1

Forging Commitments That Sustain Cooperation

RATIONAL CHOICE MODELS typically assume that people choose among possible actions so as to maximize the extent to which they achieve their goals. Such models yield few interesting predictions, however, without first introducing specific assumptions about the nature of those goals. How can we predict what someone will do unless we first have some idea of what he or she cares about?

The most frequent move at this step is to assume that people pursue goals that are self-interested in fairly narrow terms. Yet many common actions appear to be at odds with this assumption. We leave tips at out-of-town restaurants we will never visit again. We donate bone marrow in an effort to save the lives of perfect strangers. We find wallets and return them with the cash intact. We vote in presidential elections.

People uncomfortable with the self-interest assumption often respond to such contradictions by assuming a broader range of human objectives. Tips in out of town restaurants? We leave them because we care not only about our personal wealth, but also about holding up our end of an implicit understanding with the server. Voting in presidential elections? We do it because we care about fulfilling our civic duty. And so on.

The problem with this approach, however, is that if analysts are totally unconstrained in terms of the number of goals they can attribute to people, virtually any behavior can be "explained" after the fact simply by positing a taste for it. As students of the scientific method are quick to emphasize, a theory that can explain everything ends up explaining nothing at all. To be scientifically valuable, a theory must make predictions that are at least in principle capable of being falsified.

And hence the dilemma confronting proponents of rational choice theory: versions that assume narrow self-interest are clearly not descriptive, whereas those to which goals can be added without constraint lack

real explanatory power. Yes, people seem to get a warm glow from giving to charity. But why does that give them a warm glow? Why don't they get a warm glow from *not* giving to charity, since they'll end up with more money for their own purposes that way, and since the absence of any individual's gift will make no perceptible difference?

Evolutionary psychology offers a principled way of resolving this dilemma. Instead of making essentially arbitrary assumptions about people's goals, it views our goals not as ends in themselves, but rather as means in our struggle to acquire the resources needed to survive and reproduce. At first glance it might appear that a Darwinian approach to the study of human motivation would be strongly biased toward a narrow view of self-interest. Certainly, best-selling titles such as *The Selfish Gene* and *The Moral Animal* have done little to dispel that view. Yet Darwinian analysis also suggests mechanisms that might support a considerably broader conception of human motivation. And because of its inherent bias in favor of self-interest, the Darwinian framework constitutes a conservative standard by which to judge whether a specific goal can be added to the analyst's list. Thus, according to this framework, a goal can be added only if a plausible account can be offered of why pursuit of that goal is consistent with survival under competitive conditions.

The specific idea I explore here has a distinctly paradoxical flavor. It is that people can often promote their own narrow ends more effectively by pursuing certain goals that are in clear conflict with self-interest. This idea is a special case of the broader notion that people can often improve their lot by making commitments that foreclose valuable options.

The most vivid illustration remains an early example offered by Thomas Schelling (1960), who described a kidnapper who suddenly gets cold feet. He wants to set his victim free, but is afraid the victim will go to the police. In return for his freedom, the victim gladly promises not to do so. The problem, however, is that both realize it will no longer be in the victim's interest to keep this promise once he is free. And so the kidnapper reluctantly concludes that he must kill the victim.

The kidnapper and his victim confront a commitment problem, and to solve it they need a commitment device, something that gives the victim an incentive to keep his promise. Schelling suggests the following way out (1960, 43–44): "If the victim has committed an act whose dis-

closure could lead to blackmail, he may confess it; if not, he might commit one in the presence of his captor, to create a bond that will ensure his silence." Keeping his promise will still be unpleasant for the victim once he is freed, but clearly less so than not being able to make a credible promise in the first place.

In Schelling's example, the blackmailable offense is an effective commitment device because it changes the victim's material incentives in the desired way. Is it also possible to solve commitment problems by means of less tangible changes in incentives? Can moral emotions, for example, function as commitment devices? And, if so, what evidence might persuade a skeptic that natural selection had favored such emotions at least in part for that reason? These questions are my focus in this chapter.

A first step in trying to answer them will be to adopt a common language. In what follows I will use the term "contractual commitments" to describe commitments facilitated by contracts (formal or informal) that alter material incentives. Schelling's parable is an example of contractual commitment. I will use the term "emotional commitments" to describe commitments facilitated by emotions.

Two Examples

What kinds of problems are contractual commitments and emotional commitments meant to solve? I will discuss two illustrative examples. The first is a commitment problem typically solved by legal contracts, while the second is one for which such contracts are not quite up to the task.

Consider first the problem of searching for an apartment. You have just moved to a new city, and you need a place to live. If you are in Los Angeles or some other metropolis, you cannot possibly inspect each of the thousands of vacant apartments, so you check the listings and visit a few to get a rough idea of what is available—the range of prices, amenities, locations, and other features you care about. As your search proceeds, you find a unit that seems unusually attractive on the basis of your impressions of the relevant distributions. You want to close the deal. At that point, you *know* there is a better apartment out there somewhere,

but your time is too valuable to justify looking further. You want to get on with your life.

Having made that decision, the next important step is to make a commitment with the owner of the apartment. You do not want to move in and then a month later be told to leave. After all, by then you will have bought curtains, hung your art, installed phone and cable service, and so on. If you are forced to leave, not only will those investments be for naught, but you will also have to begin searching anew for a place to live.

The landlord also has an interest in seeing you stay for an extended period, since he, too, went to a lot of trouble and expense to rent the apartment. He advertised it and showed it to dozens of other prospective tenants, none of whom seemed quite as stable and trustworthy as you seemed to be.

The upshot is that even though you know there is a better apartment out there, and even though your landlord knows that a better tenant will eventually come along, you both have a strong interest in committing yourselves to ignore such opportunities. The standard solution is to sign a formal lease—a contractual commitment that prevents each of you from accepting other offers that might later prove attractive. If you move out, you must still pay your rent for the duration of the lease. If your landlord asks you to leave, the lease empowers you to refuse.

The ability to commit by signing a lease raises the amount a tenant would be willing to pay for any given apartment, and reduces the amount that its owner would be willing to accept. Without the security provided by this contractual commitment, many valuable exchanges would not occur. Leases foreclose valuable options, to be sure. But that is exactly what the signatories want them to do.

The person searching for a mate confronts an essentially similar commitment problem. You want a mate, but not just any old mate. In the hope of meeting that special someone, you accept additional social invitations and make other efforts to expand your circle of friends. After dating for a while, you feel you know a fair amount about what kinds of people are out there—what sorts of dispositions they have, their ethical values, their attitudes toward children, their cultural and recreational interests, their social and professional skills, and so on. Among the peo-

ple you meet, you are drawn to one in particular. Your luck holds, and that person feels the same way about you. You both want to move forward and start investing in your relationship. You want to get married, buy a house, have children. Few of these investments make sense, however, unless you both expect your relationship to continue for an extended period.

But what if something goes wrong? No matter what your mate's vision of the ideal partner may be, you know there's someone out there who comes closer to that ideal than you. What if that someone suddenly shows up? Or what if one of you falls seriously ill? Just as landlords and tenants can gain by committing themselves, partners in marriage have a similar interest in foreclosing future options.

The marriage contract is one way of attempting to achieve the desired commitment. On reflection, however, we see that a legal contract is not particularly well-suited for creating the kind of commitment both parties want in this situation. Even fiercely draconian legal sanctions can at most force people to remain with spouses they would prefer to leave. But marriage on those terms hardly serves the goals each partner had originally hoped to achieve.

A far more secure commitment results if the legal contract is reinforced by emotional bonds of affection. The plain fact is that many relationships are not threatened when a new potential partner who is kinder, wealthier, more charming, and better looking comes along. Someone who has become deeply emotionally attached to his or her spouse often does not *want* to pursue new opportunities, even ones that, in purely objective terms, might seem more promising.

That is not to say that emotional commitments are fail-safe. Who among us would not experience at least mild concern upon hearing that his wife was having dinner with Ralph Fiennes this evening, or that her husband was having a drink with Gwyneth Paltrow? Yet even imperfect emotional commitments free most couples from such concerns most of the time.

Again, the important point is that even though emotional commitments foreclose potentially valuable opportunities, they also confer important benefits. An emotional commitment to one's spouse is valuable in the coldly rational Darwinian cost-benefit calculus because it pro-

motes investments that enhance reproductive fitness. But note the ironic twist. These commitments work best when they deflect people from thinking explicitly about their spousal relationships in cost-benefit terms. People who consciously approach those relationships with scorecards in hand are much less satisfied with their marriages than others; and when therapists try to get people to think in cost-benefit terms about their relationships, it often seems to backfire (Murstein, Cerreto, and MacDonald 1977). That may just not be the way we're *meant* to think about close personal relationships.

Sustainable Cooperation

Solving commitment problems is important not only for successful pair bonding, but also for achieving a variety of other forms of cooperation. Indeed, the prisoner's dilemma—the ubiquitous metaphor for the difficulty of achieving cooperation among rational, self-interested individuals—is in essence a simple commitment problem. Both players in a prisoner's dilemma get higher payoffs when both cooperate than when both defect, yet no matter which choice one player makes, the other can get a still higher payoff by defecting. When each player defects, however, each receives a lower payoff than if both had cooperated, and hence the dilemma. Each player would be happy to join a mutual commitment to cooperate if he could. But when such commitments cannot be made, the dominant strategy is to defect.

The metaphor is a powerful one. It helps to explain the generally pessimistic tone of evolutionary biologists who have written on the subject of altruism and cooperation. Consider a population consisting of two distinct types of people, cooperators and defectors, who earn their living by interacting with one another in a game whose payoffs take the form of a prisoner's dilemma. When interacting with others, cooperators always cooperate and defectors always defect. Both types do best when they interact with cooperators. But if the two types looked exactly the same, they would interact with other individuals at random. And because defection is the dominant strategy in the prisoner's dilemma, defectors would always receive a higher expected payoff than cooperators in these random interactions. By virtue of their higher payoffs, defectors

would eventually drive cooperators to extinction—and hence the standard result in behavioral biology that genuine cooperation or altruism cannot survive in competitive environments.

If the same individuals face a prisoner's dilemma repeatedly, cooperation can often be sustained, because individuals will have future opportunities to retaliate against those who defect in the current round (Rapoport and Chammah 1965; Trivers 1971; Axelrod and Hamilton 1981; and Axelrod 1984). But although cooperation motivated by threat of punishment is surely better than none at all, such behavior does not really capture what we mean by genuine cooperation. When cooperative play is favored by ordinary material incentives, as when interactions are repeated, it is more aptly called prudence than cooperation.

Writers in the standard tradition seem to agree that universal defection is the expected outcome in prisoner's dilemmas that are not repeated. Yet examples abound in which people cooperate in one-shot prisoner's dilemmas. Waiters usually provide good service in restaurants located on interstate highways, and diners in those restaurants usually leave the expected tip at meal's end, even though both realize they are unlikely ever to see each other again. People return wallets they find on street corners, often anonymously, and usually with the cash intact. Millions of people brave long lines and unpleasant November weather to vote in presidential elections, even though they know their individual votes will not be decisive, even in a contest as close as the one for Florida's twenty-five electoral votes in 2000.

The pessimistic conclusion that genuine cooperation is impossible would be reversed completely if cooperators and defectors could somehow be distinguished from one another at a glance. Suppose, for example, that cooperators had a birthmark on their foreheads in the form of a red C, and that defectors had a birthmark in the shape of a red D (or no birthmark at all). Then those with a C on their foreheads could pair off together and reap the higher payoff from mutual cooperation. The defectors would be left to interact with one another. In this situation, the defectors are the ones who would be driven to extinction.

If experience is any guide, however, this optimistic conclusion is also flawed. Although millions of public radio listeners make generous contributions to support the programming they enjoy, substantially larger num-

bers take a free ride. And at least some people keep the cash in the wallets they find on city sidewalks.

To describe the mixture of motives and behavior we actually observe, we need an intermediate model, one in which cooperators and defectors are observably different in some way, but not transparently so. We may have some idea of whether a specific individual is likely to cooperate in the prisoner's dilemma, but we cannot be sure.

As Adam Smith and David Hume realized, the emotion of sympathy is a good candidate for the moral sentiment that motivates cooperation in social dilemmas. Your goal as an individual is to interact with someone who feels sympathy for your interests, in the hope that such a person will be internally motivated to cooperate, even though he could earn more by defecting.

But how do you know whether someone feels sympathy for your interests? Darwin (1965 [1872]) wrote of the hard-wired link between emotional states in the brain and various details of involuntary facial expression and body language. Consider the crude drawing shown in figure 1.1. This drawing shows only a few details, yet people in every culture recognize even this simple abstraction as an expression of sadness, distress, sympathy, or some other closely related emotion. Most people cannot produce this expression on command (Ekman 1985). (Sit in front of a mirror and try it!) Yet the muscles of the human face create the expression automatically when the relevant emotion is experienced (Darwin 1965 [1872]). Suppose you stub your toe painfully, leading an acquaintance who witnesses your injury to manifest that expression immediately. Such a person is more likely to be a trustworthy trading partner than someone who reacted to the same incident without expression.

Simple facial expressions, of course, are not the only clues on which we rely, or even the most important ones. In ways I will describe presently, we construct character judgments over extended periods on the basis of a host of other subtle signals, many of which enter only subconscious awareness. On the basis of these impressions, among potential trading partners we choose those we feel are most likely to weigh not just their own interests when deciding what to do, but our own interests as well.

Defectors have an obvious incentive to mimic whatever signs we use

Figure 1.1

for identifying reliable trading partners. Selection pressure should strongly favor capacities for effective deception, and examples of such capacities clearly abound in human interaction. If signals of emotional commitment could be mimicked perfectly and without cost, these signals would eventually cease to be useful. Over time, natural selection would mold false signals into perfect replicas of real ones, driving the capacity for signaling genuine commitment to extinction.

Whether that capacity has been able to stay a step ahead of attempts to mimic it is an issue that is difficult to settle on a priori grounds. Granted, natural selection ought to be good at building a copy of a useful signal. But it also ought to be good at modifying an existing signal to evade mimicry. Which of these opposing tendencies wins out in the end is an empirical question, one to which I devoted considerable attention in my 1988 book, *Passions Within Reason*. There, I also argued that even if we grant the existence of reliable signals of emotional commitment, the resulting equilibrium must entail a mixed population of cooperators and defectors. In any population consisting only of cooperators, no one would be vigilant, and opportunities would thus abound for defectors. In a mixed population, cooperators can survive only by being sufficiently vigilant and skilled in their efforts to avoid good mimics.

Can we, in fact, identify people who are emotionally predisposed to cooperate? My Cornell colleagues Tom Gilovich and Dennis Regan and I (1993b) present evidence from an experimental study that appears to support this possibility. In this study, which I describe in chapter 2, we

found that subjects of only brief acquaintance were able to identify defectors with better than twice chance accuracy in a one-shot prisoner's dilemma game. And as the following thought experiment suggests, substantially higher accuracy rates may be possible among people who know one another well:

> Imagine yourself just having returned from a crowded sporting event to discover that you have lost an envelope containing $5000 in cash from your coat pocket. (You had just cashed a check for that amount to pay for a used car you planned to pick up the next morning.) Your name and address were written on the front of the envelope. Can you think of anyone, not related to you by blood or marriage, who you feel certain would return your cash if he or she found it?

Most people say yes. Typically, the persons they name are friends of long duration, choices that seem natural for two reasons. First, the more time one spends with someone else, the more opportunities there are to observe clues to that person's emotional makeup. And second, the more time people spend with a friend, the deeper their emotional bonds are likely to be. Sympathy, affection, and other emotions that motivate trustworthy behavior are likely to be more strongly summoned by interactions with close friends than with perfect strangers.

Notice that, although the people named are usually ones with whom we engage in repeated interactions, the particular episode involving the cash is not a repeated game: keeping the cash would not lead to retaliation in the future because there would be no way of knowing that your friend had found and kept the cash. You are also unlikely to have direct evidence regarding your friend's behavior in similar situations in the past.

When pressed, most people respond that they named the people they did because they felt they knew them well enough to be able to say that they would *want* to return the cash. The prospect of keeping a friend's cash would make them feel so bad that it just wouldn't be worth it.

If you named the people you did for roughly similar reasons, then you accept the central premise of my signaling argument, which is that we can identify (possibly strongly context-dependent) behavioral tendencies such as trustworthiness in at least some other people. That doesn't prove

that this premise is correct. But it constitutes a hurdle for those who would persuade us that it is false.

Toward a More Realistic Model

Darwinian models like the ones illustrated in the preceding section are clearly stick-figure caricatures. Although they may capture important features of the reality they attempt to represent, they cannot hope to embody the complex range of human behavior and emotion triggered when individuals have conflicting interests. Yet even such simple models often afford powerful insights.

For precisely this reason, however, we are often prone to interpret them too literally. For example, my stick-figure model encourages us to view people as being either cooperative or not. Under the influence of this model, I once equated the task of solving prisoner's dilemmas to that of finding a trustworthy trading partner. But social psychologists have long been skeptical about the existence of stable individual differences of this sort. They believe that differences in behavior are far more likely to be explained by the details of the situation than by stable differences in individual traits. This insight seems clearly applicable to situations that test our willingness to cooperate. Although individual differences in the overall tendency to cooperate surely do exist, most of us cannot be easily assigned to a single category from the cooperator/defector pair. All but the most extreme sociopaths have within them the capacity to experience sympathy for others and to weigh others' interests when deciding what to do. And although almost all of us have cooperated in situations in which it would have paid to defect, most of us have also let others down on occasion.

The Emergence of Sympathy

I now believe that the search for a reliable trading partner is not a quest to identify an indiscriminately trustworthy individual, but rather a process of creating conditions that make us more likely to elicit cooperative tendencies in one another. In a remarkably insightful essay, David Sally

(2000) has summarized a large literature that bears precisely on this process.

Beginning with the writings of David Hume and Adam Smith, Sally traces the intellectual history of the concept of sympathy and reports on some extremely fascinating results on the mechanics of how it develops in human interactions. I use the term "mechanics" advisedly, for an important thread in the studies he reviews is that we are often remarkably mechanical in the ways we respond to stimuli.

Many of these studies remind me of a behavior that I myself have puzzled over for a long time, which is that I usually set my watch five minutes ahead. Not everyone does this, of course, but I know many who do. We do it because it seems to help us get to appointments on time. But *why*? When someone asks me what time it is, I always report the correct time by simply subtracting five minutes from whatever my watch says. So I am not really fooling myself by setting it ahead. Yet if I have an appointment across campus at 11:00, the mere act of seeing a dial saying 10:55 apparently triggers an emotional reaction in my brain, which in turn gets me going a little more quickly than if I had relied only on my knowledge that the correct time was 10:50. Whatever the details of *how* the actual mechanism works, it clearly does work, as I know from experiments in which I have set my watch to the correct time for extended periods.

Other studies confirm the importance of seemingly mindless physical motions. For example, if you are pulling a lever toward you when an experimenter shows you a Chinese ideograph, you are much more likely than control subjects to give the image a positive evaluation when you are queried about it later. But if you are pushing a lever away from you when you are shown the ideograph, you are much more likely to give it a negative evaluation later (Cacioppo, Priester, and Berntson 1993). If you put a pen between a person's teeth—forcing him to smile, as it were—and then show him a cartoon, he is much more likely to find it funny than if he does not have a pen between his teeth (Strack, Martin, and Stepper 1988).

Similar mechanical stimulus-response patterns are also strongly implicated in the processes by which sympathetic bonds form between people. An important factor in these processes is the concept of valence—

an evaluation that is either positive or negative. Psychologists have identified a universal human tendency to assign an initial valence in response to virtually every category of stimulus—even words that may seem neutral, or photographs, or visual scenes of any kind (Lewin 1935; Bargh 1997).

So, too, with persons. When you meet someone, you make an initial up/down categorization very quickly, probably before you are even consciously aware of it (if indeed you ever become consciously aware of it). Likeness seems to play a role in these judgments (Lazarsfeld and Merton 1954). You are more apt to assign positive valence to someone who is like you in some way—say, in dress, speech patterns, or ethnic background. Reputation matters, as does the character of your initial exchange. Attractiveness is important. Physically attractive persons are far more likely than others to receive a positive initial evaluation (Eagly et al. 1991; Sally, 2000).

Once the initial valence has been assigned, a biased cognitive filter becomes activated. You still evaluate further aspects of your experience with a new acquaintance, but with a slant. If the initial evaluation was positive, you are much more likely to treat ambiguous signals in a positive light. But if your initial impression was negative, you are more likely to assign negative interpretations to those same signals. Such feedback effects often make first impressions far more important than we might like them to be on ethical grounds.

A colleague of mine once described a vivid example of how an initial negative assessment had distorted several subsequent judgments made by his three-year-old son. He had taken his son to visit Will Rogers's ancestral home, a dark, forbidding gothic structure. The boy did not want to go in but finally yielded to his father's urgings. As they toured the house, a tape of Will Rogers reading from one of his works was playing in the background. To passages in Rogers's narrative that had an ambiguous sound or meaning, the boy seemed to assign the darkest possible interpretations. For example, when Rogers said at one point, "Well, I tried," the boy asked his father, "Why'd he die?" Time and again, the boy's interpretations were slanted to the negative.

Given this biased filter, the development of successful personal relationships hinges powerfully on getting off to a good start. If your first

experience in a relationship is positive, you engage further. But if you begin with a negative experience, things are likely to get worse.

Psychologists report that an important component of normal sympathetic responses in relationships is a subconscious impulse to mimic what your conversation partner is doing. If she smiles, you smile. If she yawns, you yawn. If she leans to one side, you lean the same way (Bavelas et al. 1986; Hatfield, Cacioppo, and Rapson 1994).

Although such mimicry turns out to be critically important, most people are not consciously aware of it. In one study, for example, psychologists had separate conversations with two groups of subjects—a control group in which the psychologists interacted without special inhibition, and a treatment group in which the psychologists consciously did not mimic the postures and other movements and expressions of their conversation partners (Chartrand and Bargh 1998). Subjects in the treatment group reported generally negative feelings toward the psychologists, while those in the control group found the same psychologists generally likeable. Apart from the suppression of physical acts of mimicry in the treatment group, no other observable details of the interactions differed between groups. This finding is consistent with the view that people may subconsciously interpret failure to mimic as signifying a deficit of sympathy.

Studies of how the appearance of married couples evolves over time also suggest that physical mimicry is an important aspect of social interaction. In one study, subjects were shown individual wedding-year photographs of a large sample of men and women, and then asked to guess which men had married which women. The accuracy of their guesses was no better than chance. But when other subjects were given the same matching task on the basis of individual photos taken after twenty-five years of marriage, the accuracy of their guesses was far better than chance (Zajonc et al. 1987). Over the course of a quarter-century of married life, apparently, the furrow of the brow, the cast of the lip, and other subtle details of facial geography seem to converge perceptibly. I have two friends, a married couple, who have been professional storytellers for several decades. As is common among storytellers, they employ exaggerated facial expressions to highlight the emotional ebbs and flows of their tales. I don't know how much they resembled one another in

their youth. But people often remark on how strikingly similar they look today.

The process of bonding with another person influences, and is influenced by, physical proximity and orientation. Being too close invites a negative response, but so does being too far away, where "too close" and "too far" depend partly on cultural norms (Hall 1982). The gaze is also important (Sally 2000). Frequency and intensity of eye contact correlates strongly with the duration and intimacy of personal relationships (Patterson 1973). Among recent acquaintances, both extremely high levels of eye contact and extremely low levels often prove aversive. If experimenters seat subjects too close together, they will look at one another less frequently than if they are seated at a more comfortable distance (Argyle and Dean 1965).

The intensity of the initial interaction—even if purely the result of chance—has important consequences for long-term bonding. For example, combat troops who were under heavy shelling in the same unit corresponded with one another for many more years and much more frequently than combat troops who were not shelled heavily in the same engagement (Elder and Clipp 1988). The heavyweight fighters John Tunney and Jack Dempsey wrote to one another for years after their legendary title bouts, and did many favors for one another. They were not friends. They never even particularly liked one another, but they were thrown together in very intense circumstances that seemed to forge a bond (Heimer 1969).

Mere exposure also matters. As Robert Zajonc and his colleagues have shown, the simple fact that we have been repeatedly exposed to an initially neutral stimulus—such as a Chinese ideograph or the shape of a polygon—is enough to make us like it (Zajonc et al. 1987). Repeated exposure to persons has essentially the same effect. Relative to people we have never seen, we strongly prefer to interact with those we have seen repeatedly—in the same elevator or on the same train platform—even though we have never acknowledged one another's presence before (Brockner and Swap, 1976). As David Hume wrote, "When we have contracted a habitude and intimacy with any person; tho' in frequenting his company we have not been able to discover any very valuable quality,

of which he is possess'd; yet we cannot forbear preferring him to strangers, of whose superior merit we are fully convinc'd" (Hume 1978 [1740], 352, as quoted by Sally [2000]).

Laughter also seems to be important in the development of relationships. Why do we have such a pronounced capacity to experience mirth in our interactions with one another? While other animal species may have something analogous to this capacity, even our closest relatives among primates do not have it to anything like the same degree. One possibility is that laughter not only promotes the development of social bonds, it may also be an unusually effective test of shared sympathy and understanding. People who find the same things funny often find they have many other attitudes and perceptions in common.

In short, the emergence of sympathetic bonds among people is a very complex physical, cognitive, and emotional dance. People feel one another out, respond to one another, choose to develop closer bonds with some, and abandon further contact with others.

This brief account describes only a small sample of the literature surveyed in David Sally's paper. Suffice to say, however, that this literature suggests a far more complex phenomenon than the one I sketched in *Passions Within Reason*. My simple stick-figure model gave the impression that some people feel sympathy toward others and some people do not, suggesting that the challenge is to interact selectively with those in the first group. David Sally's insight is that it would be far more descriptive to say that most people have the capacity to experience sympathy for the interests of others *under the right circumstances*. The challenge is to forge relationships in which mutual sympathy will develop sufficiently to support cooperation.

Does Sympathy Predict Cooperation?

Substantial evidence suggests that the same factors that promote the development of sympathetic bonds between individuals also predict an increased likelihood of cooperation. A large literature, for example, documents the importance of physical proximity and communication as predictors of the likelihood of cooperation in prisoner's dilemmas (Sally [1995] offers a review). If you are sitting next to your partner and there's

a screen between you so you can't see one another, you are more likely to cooperate than if you are sitting across the table with a screen between you. You are closer, physically, in the first condition, even though you can't see one another in either case. But take the screens away and the people sitting side by side are less likely to cooperate than the people who are sitting opposite one another (Gardin et al. 1973). Apparently, the side-by-side pairs are sitting too close together to feel comfortable with extended eye contact, while those seated opposite one another do not suffer from this inhibition.

Many experiments have found that friends are much more likely to cooperate in social dilemmas with one another than are others of lesser acquaintance (Sally [2000] reviews several studies that confirm this finding). All else being constant, the longer you have known a person, the stronger your mutual bond, and the greater your assurance of cooperation. Written exchanges among participants stimulate cooperation in prisoner's dilemma experiments, but not by nearly as much as face-to-face exchanges, even if the content of the exchanges is essentially the same (Sally 1995; Valley, Moag, and Bazerman 1998).

Considered as a whole, the evidence is consistent with an affirmative answer to our question of whether moral emotions such as sympathy facilitate commitment. This evidence does not rule out alternative interpretations conclusively. But in my view, it places a substantial burden of proof on those who argue that moral emotions do not facilitate commitment.

Was the Capacity for Sympathy Forged by Natural Selection?

What about the second question I posed at the outset? Does available evidence provide any reason to believe that natural selection favored the evolution of sympathy at least in part *because* of its ability to solve commitment problems? A moral emotion won't be favored by natural selection merely because it motivates cooperation. It must motivate cooperation in such a way that cooperation *pays*. Does sympathy meet that test? Here, too, existing studies suggest an affirmative answer.

Consider again the two conditions that must be satisfied for a moral

emotion like sympathy to facilitate mutual cooperation in one-shot prisoner's dilemmas. (By "one-shot" prisoner's dilemmas I do not mean only games that are played once between perfect strangers. Such dilemmas also include interactions among friends of long standing, as in situations in which partners are unable to discover who is responsible for the bad outcome they experience.) First, it must motivate players to cooperate, even though they would receive higher payoffs by defecting. And second, players must have statistically reliable means of predicting which potential trading partners will be trustworthy. Available evidence provides support for both conditions. In the first instance, conditions that have been shown experimentally to foster the development of sympathy have also been shown to promote cooperation in one-shot prisoner's dilemmas (see Sally [2000] for a detailed summary).

As for whether people can predict whether their partners will cooperate, some studies also show that people are aware—sometimes unconsciously, but in ways that influence observable behavior—of the degree of sympathetic bonding that exists between themselves and others. One such study, described in detail in the next chapter, suggests that experimental subjects are able to predict their partners' choices in social dilemmas on the basis of only brief periods of interaction. And as suggested by the thought experiment involving the lost envelope full of cash, substantially higher accuracy rates may be possible among people who know one another well.

How can someone tell that a potential trading partner is genuinely sympathetic to his interests? As noted earlier, psychologists have confirmed Darwin's claim that certain facial expressions are characteristic of specific emotions. Psychologists have also found that posture and other elements of body language, the pitch and timbre of the voice, the rate of respiration, and even the cadence of speech are systematically linked to underlying motivational states. Because the relevant linkages are beyond conscious control in most people, it is difficult to conceal the experience of certain emotions, and equally difficult to feign the characteristic expressions of these emotions on occasions when they are not actually experienced. For this reason, such clues provide reliable information about the emotions we trigger in others. In addition to facial expressions and other physical symptoms, we rely on reputation and a variety of

other clues to predict how potential partners will treat us in specific situations (for a discussion of the role of reputation and other factors, see chapter 4 of my *Passions Within Reason*).

If we possess the capacity to discern whether others will treat our interests with respect, it is an undeniably useful one. Equally clear is that this capacity is extremely complex. As we have seen, the development of sympathetic bonds is a process involving multiple perceptual, cognitive, and emotional capacities. No one in the scientific community questions that these capacities exist in most humans. Nor does anyone question that these capacities involve specialized components of the inborn neural circuitry of humans. Nor, to my knowledge, has anyone offered a plausible theory other than natural selection that could account for the presence of such components.

Of course, the claim that moral emotions help solve commitment problems could be valid even if the relevant capacities through which those emotions act were selected for altogether different purposes—just as, for example, the human capacity to produce and enjoy music could have emerged as an accidental by-product of intellectual and emotional capabilities selected for other purposes. Indeed, theoretical considerations from the animal signaling literature suggest that moral sentiments such as sympathy almost certainly could not have *originated* purely because of their capacity to solve one-shot dilemmas.

The basic problem is that natural selection cannot be forward-looking. It cannot recognize, for example, that a series of mutations might eventually produce an individual with the capacity to solve one-shot prisoner's dilemmas, and then favor the first costly step to that end, even though it yields no immediate benefit. As I will explain presently, it is this first step that presents the difficulty, because the initial appearance of a signal would have no meaning to external observers. It would thus entail costs, but no benefits. And the Darwinian rule is that a mutation must offer an *immediate* surplus of benefits over costs, or else be consigned to the evolutionary scrap heap.

How do signals ever originate, then? Essentially by accident, according to the derivation principle developed by Niko Tinbergen (1952). The constraint imposed by this principle is clearly illustrated by the example of the dung beetle. The insect gets its name from the fact that it

escapes from predators by virtue of its resemblance to a fragment of dung. Biologists argue, however, that this advantage cannot explain how this beetle came to resemble a fragment of dung in the first place. The problem is that if we start with a species whose individuals bear not the slightest resemblance to a dung fragment, a minor mutation in the direction of a dung-like appearance would not have been of any use, since, as Stephen Jay Gould asks, "can there be any edge in looking 5 percent like a turd?" (1977, 104). A mutation toward dung-like appearance will enhance fitness only if the individual's appearance *already* happened to be similar enough to a dung fragment for the mutation to have fooled the most myopic potential predator. Thus the initial path toward near-resemblance must have been essentially a matter of chance—the result of mutations that were favored for other reasons and just happened to produce a dung-like appearance in the process. Once the resemblance crosses the recognition threshold by chance, however, natural selection can be expected to fine-tune the resemblance, in the same ruthlessly effective way it fine-tunes other useful traits.

Essentially the same logic should apply to the emergence of an observable signal of a moral emotion such as sympathy. If the *only* behavioral effect of having sympathy were to motivate cooperation in one-shot prisoner's dilemmas, the first mutants with a small measure of this emotion would have enjoyed no advantage, even if their mutation happened to be accompanied by an observable signal. By virtue of its novelty, no one would have known what the signal meant, so it could not have facilitated selective interaction among sympathetic individuals. And since an undiscriminating tendency to cooperate entails costs, natural selection should have worked against sympathy, for the reasons just described.

If sympathy and other moral emotions were favored by natural selection in their earliest stages, they must therefore have conferred some other benefit. For example, perhaps a mutant with the capacity for sympathy was a more effective parent, a fitness enhancement that might have compensated for the initial costs of an indiscriminately sympathetic posture toward unrelated individuals.

A second possibility, one I explore in more depth here, is that the moral sentiments may function as self-control devices. In a world populated by utility maximizers of the sort usually assumed in economics,

self-control problems would not exist. Such individuals would discount future costs and benefits at a constant exponential rate, which means that any choice that would seem best right now would also seem best in hindsight. Extensive evidence summarized by George Ainslie, however, suggests that all creatures, animal and human, tend to discount future rewards not exponentially but hyperbolically (1992). As Ainslie explains, hyperbolic discounting implies a temporary preference for "the poorer but earlier of two goals, when the earlier goal is close at hand." Seated before a bowl of salted cashews, for example, people often eat too many, and then later express sincere regret at having spoiled their dinners.

A similar time-inconsistency problem confronts people who interact in a sequence of repeated prisoner's dilemmas. In such situations, Rapoport and Chammah (1965), Axelrod (1984), and others have demonstrated the remarkable effectiveness of the tit-for-tat strategy—in which you cooperate in the first interaction, then in each successive interaction mimic whatever your partner did in the immediately preceding one. Note, however, that implementation of tit-for-tat entails an inherent self-control problem. By cooperating in the current round, the tit-for-tat player must incur a small present cost in order to receive a potentially much larger benefit in the future. In contrast, a player who defects in the current round receives a benefit immediately, whereas the costs of that action are both delayed and uncertain. Thus someone might realize he would come out ahead in the long run if he cooperated in the current interaction, yet find himself unable to resist the temptation to reap the immediate gains from defecting.

A person who is sympathetic toward potential trading partners is, by virtue of that concern, less likely than others to yield to temptation in the current interaction. Such a person would still find the gains from defecting attractive, but their allure would be mitigated by the prospect of the immediate aversive psychological reaction that would be triggered by defecting. For this reason, persons with sympathy for their trading partners would find it easier than others to implement the tit-for-tat strategy in repeated prisoner's dilemmas. To the extent that the ability to execute tit-for-tat enhances fitness, people who experienced sympathy would have fared better than those who did not, even if no observable signal of sympathy were generally recognized.

Similar reasoning applies in the case of commitment problems that entail deterrence. It will often be prudent to exact revenge against an aggressor, even at considerable personal cost, when doing so would help create a reputation that will deter future aggression. Self-interested rational persons with perfect self-control would always seek revenge whenever the future reputational gains outweighed the current costs of taking action. As before, however, the gains from a tough reputation come only in the future while the costs of vengeance-seeking come now. A person may know full well that it pays to be tough, yet still be tempted to avoid the current costs of a tough response. Thus an angry person may be more likely to behave prudently than a merely prudent person who feels no anger.

The empirical literature I described earlier documents the existence of reliable markers of sympathy and other moral emotions that influence human interaction. The animal-signaling literature provides compelling theoretical reasons for believing that both the emotions themselves and their observable signals are unlikely to have originated because of their capacity to resolve one-shot dilemmas. But given that these emotions and their markers exist, for whatever reasons, there is every reason to expect natural selection to have refined them for that purpose. We know, for example, that individual differences in emotional responsiveness are at least weakly heritable (Bruell 1970). If selective trustworthiness is advantageous and observable, natural selection should favor individual variants who are both more trustworthy and better able to communicate that fact to others.

Strategic Issues

Robert Solomon has stressed the importance of viewing emotions not as purely exogenous events but rather as something over which we often have considerable control (2003). The literature that describes how sympathetic bonds develop among people is strongly supportive of this view. In predictable ways, people react to the things we do and say to them, and we react to the things they say and do. Over time these reactions change both them and us. And because the outcomes of our choices are

to a large extent predictable, decisions about the details of interpersonal interaction have a potentially strategic dimension.

Consider the decision to associate with someone. In the absence of unexpected negative feedback, and sometimes even in the presence of it, deciding to spend time with someone is tantamount to a decision to like him and to develop sympathy for his interests. If his values were initially different from yours, a decision to spend time together is likely to diminish those initial differences. Our choice of associates, therefore, is at least in part a strategic choice about the kind of values we want to hold.

The knowledge of how sympathetic bonds emerge between people also has strategic implications for the design of organizations and institutions. Most university administrations are keenly aware, for example, of how sympathy (and antipathy) can effect people's decisions in promotion cases. Accordingly, few universities delegate ultimate decision power in such cases to a faculty member's departmental colleagues. Most rely heavily on ad hoc committees composed of faculty outside the department.

Widespread prohibitions on gift-giving in institutional settings can be understood in a similar way. The fear is not, for example, that by giving a gift to a professor, the student will bribe the professor to overrule his honest judgment about the true grade she feels the student deserves. Rather, it is that the gift may foster a strong sympathetic bond between the two, which in turn may distort even the professor's most determinedly objective assessment of the student's performance.

Knowledge of the processes that forge sympathetic bonds among people also sheds light on the common practice of cronyism among corporate executives and political leaders. When such people ascend to positions of power, their first step is often to hire assistants from the ranks of their long-standing friends and subordinates. This practice invariably exposes them to the criticism that they value loyalty above competence. Yet for leaders to achieve the goals they were chosen to implement, they require subordinates who are not only competent, but also trustworthy. And because the sympathetic bonds that support trust are strongest when nurtured over a period of many years, the preference for long-term associates is, on its face, neither surprising nor blameworthy. What *would* be

cause for alarm would be the observation that an executive's long-term friends and subordinates were mostly incompetent hacks. Those are choices for which we have every reason to hold a potential executive accountable.

Concluding Remarks

Traditional rational choice theories confront a painful dilemma. Without making specific assumptions about people's goals, they cannot generate testable implications for observable behavior. Most rational choice models thus assume that people pursue narrowly selfish goals. Yet people make anonymous gifts to charity, leave tips at restaurants on interstate highways, vote in presidential elections, and take a variety of other costly actions with little prospect of personal gain. Some analysts respond by introducing tastes for such behavior. But if the list of goals is not circumscribed in some way, virtually any behavior can be rationalized by simply positing a taste for it. When a man dies shortly after drinking the used crankcase oil from his car, we do not really explain anything by asserting that he must have had a powerful taste for crankcase oil.

Darwinian analysis offers a principled way of resolving this dilemma, one that is not vulnerable to the crankcase-oil objection. Evolutionary models view our goals not as ends in themselves, but as means to acquire the material resources needed for survival and reproduction. In this framework, we are free to offer a "taste for cooperation" to explain why people cooperate in one-shot prisoner's dilemmas, but only if we first can explain how having such a taste might help a person acquire the resources needed to survive and reproduce.

I have argued that cooperation in one-shot prisoner's dilemmas is sustained by bonds of sympathy among trading partners. The models I employed in my earlier work on this subject encouraged the view that some people have genuinely cooperative tendencies while others do not. I now believe it is far more descriptive to say that most people have the capacity to develop bonds of sympathy for specific trading partners under the right set of circumstances. The preference for cooperation is not an unconditional one, but rather one that depends strongly on the history of personal interaction between potential trading partners. But this amend-

ment, in the end, is a detail. Even traditional preferences depend on context in essentially similar ways. We don't desire food at every moment, for example, but only after a suitable delay since the ingestion of our last meal.

Narrow versions of the rational choice approach leave the moral emotions completely out of the picture. Naked self-interest is not an unimportant motive, of course, and these models can help us understand much of the observed human behavioral repertoire. But there is also much that is simply beyond the reach of these models. And there is some evidence that the models themselves may do social harm by encouraging us to expect the worst from others (a point to be developed in some detail in chapter 9). By giving us a principled framework for broadening our assumptions about human motives, the Darwinian approach points the way to long-overdue enrichments of the narrow rational choice approach.