6

The Role of Moral Sentiments in Economic Decision Making

Timothy Ketelaar
New Mexico State University

Explanations of economic behavior often fall into one of two camps: normative and descriptive. Normative accounts typically consist of attempts to fit a wide variety of phenomena into one broad theoretical framework consisting of a single set of axioms governing rational decision making (e.g., von Neumann & Morgenstern, 1944). Descriptive approaches, in contrast, consist of attempts to accurately describe how individuals actually behave, regardless of whether these accounts are consistent with any particular normative approach to rational decision making (Gilovich, Griffin, & Kahneman, 2002; Hogarth & Reder, 1987). What makes this normative-descriptive distinction so intriguing is the observation that many descriptive accounts of economic behavior are often strikingly inconsistent with normative models. It turns out that this contradiction between how people ought to behave, according to normative models, and how they actually do behave, as evidenced by descriptive accounts, is a puzzle that has intrigued observers of human nature for centuries.

Eighteenth-century moral philosophers such as David Hume (1777) and Adam Smith (1759) attempted to bridge the gap between what people should do (normative principles) and what they actually do (descriptive accounts) by exploring the role of sentiments in decision making. Hume and Smith argued that moral sentiments, in the form of emotions and passions, exert such a powerful influence on judgment that they often compete with more rational deliberations in determining our behavior. In essence, they argued that emotions can provoke otherwise rational individuals to behave in a
manner that is inconsistent with normative principles of good judgment. The present chapter expands on this 18th-century conception of moral sentiments to explore the role of emotion in contemporary economic decision research. After discussing the historical role that emotions have played in normative models of good judgment and decision making, I provide a brief review of recent empirical work linking moral emotions to norm-obeying and norm-enforcing behaviors in experimental economics. Finally, it is argued that insights from evolutionary game theory can illuminate the possible adaptive design of moral sentiment systems and, in doing so, may help us understand why decision makers often do not behave in the strictly rational manner that some normative models suggest that they should behave.

EMOTIONS AND NORMATIVE MODELS OF GOOD JUDGMENT

Just as contemporary decision theorists often employ conflicting standards (e.g., normative vs. descriptive) to explain economic behavior, 18th-century moral philosophers once grappled with an equally conflicting set of standards for discerning an individual's moral character—namely, the standards of absolute moral perfection versus mere adherence to social convention. In his Theory of Moral Sentiments, for example, Adam Smith (1759/2000) argued that:

when we are determining the degree of blame or applause which seems due to any action, we very frequently make use of two different standards. The first is the idea of complete propriety and perfection... The second is the idea of that degree of proximity or distance from this complete perfection, which the actions of the greater part of man commonly arrive at. (pp. 29–30)

Although the two standards to which Adam Smith referred are strikingly similar to the normative–descriptive distinction employed by contemporary decision researchers, Smith and other moral philosophers of his era were more interested in the moral (rather than economic) implications of behavior. By contrast, contemporary economists and psychologists are much less interested in whether a given behavior approximates a particular standard of moral propriety and perfection. Instead economists and psychologists are more concerned with the degree to which descriptive accounts of judgment and decision making are consistent with particular normative standards of sound economic reasoning, such as those standards embodied in the normative rules of inference of the logical calculus (von Neumann & Morgenstern, 1944). Just as 18th-century moral philosophers regretted that individuals do not always behave as they should in the moral domain, so too contemporary decision theorists lament that economic decision makers routinely enact preferences that are at sharp contrast with normative standards of proper economic behavior. Regardless of whether one considers violations of normative standards to indicate moral turpitude or merely economic irrationality, it is quite obvious that the distinction between what people should do and what most people actually do looms as large today in the minds of economists and psychologists as it did in the writings of 18th-century moral philosophers.

Is there any reason to believe that contemporary decision theorists have anything to learn from 18th-century moral philosophers? Although one might imagine that distinguishing the moral and logical implications of a particular behavior is a relatively straightforward task, it turns out that moral and logical implications are, in fact, quite easily confused. Nowhere is this more evident than in our evaluations of behaviors that violate normative standards of appropriate behavior. In this regard, behaviors that deviate from normative standards of sound reasoning are not only viewed as illogical, they are often met with judgments of moral disapproval that include strong emotional reactions. Even the most casual observer of human nature will note just how easy it is to confute judgments of deviant behavior with judgments of immorality. Social psychologists, for example, have shown that merely observing a person violating a culturally shared norm is enough to evoke strong moral emotions such as anger and contempt (Haidt, 2003; Ketelaar et al., 2006; Rozin, Lowery, Imada, & Haidt, 1999). Although the types of non-normative behaviors studied by social psychologists (e.g., stealing a purse from a blind person) are a bit far removed from the sorts of norm violations that interest economists (e.g., violations of the invariance axiom), empirical evidence now suggests that morally charged emotions such as anger and guilt are routinely evoked in economic decision tasks that were once conceptualized as purely logical deliberations among rational participants (see Fehr & Gächter, 2002; Ketelaar & Au, 2003; Pillutla & Murnighan, 1996; Van Kleef, De Dreu, & Manstead, 2004; Zeelenberg & Pieters, chap. 7, this volume).

Given the ubiquitous role that emotions appear to play in everyday decisions (Damasio, 1994; Elster, 2000; Nussbaum, 2004), might there be room for a more formal treatment of the role of these moral sentiments in normative accounts of economic decision making? Along these lines, the present chapter argues that a greater appreciation of the role of emotion might be informative in regards to two questions of interest to economists and psychologists. First, why are emotions so easily provoked in purely economic decisions? Second, why do emotions often compel individuals to pursue strategies that do not appear to be in their immediate material self-interest? It is argued that a large part of the puzzle concerning why irrational emotions so often intrude on our most important decisions may lie in recognizing this link between moral sentiments and judgments of normatively acceptable (and unacceptable) behavior.
6. EMOTIONS AND ECONOMIC DECISION MAKING

est in the moral reasoning literature, rationalism managed to extend its reach into the more general literature on social judgment and decision making (see Clore, 2005; Ketelaar & Clore, 1997). As one emotion researcher observed (Scherer, 1984):

Today the image of man is no longer that of an individual enslaved by his passions, but rather that of a philosopher making decisions on the basis of logical deduction and inference. In this tradition, emotion is seen as a regrettable flaw in an otherwise perfect machine. (p. 293)

The view from economics has no: been much rosier (Elster, 1995):

The standard view of the relation between rationality and emotions is, of course, that emotions interfere with rationality. They are, as it were, sand in the machinery of action. (p. 1394)

Although emotionally influenced decisions often appear to be at odds with normative principles of rational inference (von Neumann & Morgenstern, 1944), perhaps consistency with the cold rules of logic is the wrong reference point from which to view these sentiments (Clore, 2005; Ketelaar & Clore, 1997; Ketelaar & Todd, 2001). In this regard, recent research on moral judgment suggests that emotions have more in common with relatively automatic processes such as intuition and perception than they do with more calculated processes such as rational deliberation and logical inference (Haidt, 2001). To the degree that our judgments, moral or economic, are often based on moral intuitions and emotional perceptions, rather than logic and rational deliberation, one might suspect that the impact of emotion on economic decision making has been greatly underestimated. Participants in studies of moral judgment, for example, have no difficulty generating reasons for their judgments. Yet further analysis reveals that many of these reasons are actually post hoc confabulations, rather than veridical recollections of the actual thought processes that led to their judgment (Haidt, 2001). Moreover, participants in studies of moral judgment often report that they "don't really know" why a particular behavior is morally wrong, it just is! (Haidt, 2001). Although these findings do not necessarily establish that moral reasoning is enslaved by the passions, they do suggest that many decision processes that were previously thought to be the product of logical inference and rational deliberation are instead more accurately described as being the product of relatively automatic emotional intuitions and perceptions. The remainder of this chapter explores how this insight might be integrated into our understanding of certain types of economic decisions.
EMPIRICAL EVIDENCE FOR EMOTIONAL INFLUENCES ON ECONOMIC DECISION MAKING

Emotion researchers have found it useful to conceptualize passions and sentiments in terms of their intrapersonal and interpersonal functions (Bowley, 1969; Van Kleef et al., 2004). Intrapersonal functions refer to the impact that emotions have on individual decision making, such as when postdecision regret motivates you to pursue an economic opportunity that you had previously rejected (Zeelenberg, 1999). Interpersonal functions, in contrast, focus on the impact that emotions have on social decision making, such as when anger motivates you to punish a selfish contributor in a public goods game (Fehr & Gächter, 2002). In relation to these intrapersonal and interpersonal functions, emotion theorists have identified at least two important clusters of moral sentiments: (a) self-focused sentiments such as shame, embarrassment, regret, and guilt; and (b) other-focused sentiments such as contempt, anger, disgust, and schadenfreude1 (Haidt, 2003; Lewis, 1993; Ortony, Clore, & Collins, 1988). Self-focused emotions emphasize an evaluation of the appropriateness of one's own behavior, whereas other-focused emotions tend to direct one's attention to the appropriateness of others' behavior. Although there exists a great variety of ways in which the relation between emotions and economic decisions can be conceptualized (see Zeelenberg & Pieters, chap. 7, this volume), the remainder of this chapter focuses on two areas in which recent theoretical and empirical developments have been quite illuminating: (a) self-focused emotions and norm-obeying behaviors, and (b) other-focused emotions and norm-enforcing behaviors.

Self-Focused Emotions and Norm-Obeying Behavior

Scottish philosopher Adam Smith (1759/2000) was among the first to develop a theory of how self-focused emotions can motivate an individual to obey (or disobey) certain normative standards of good judgment. Smith argued that when we experience certain moral sentiments, such as guilt, they intuitively compel us to do what is morally appropriate, despite the fact that rational economic calculation might suggest otherwise. The sentiment of gratitude, for example, can compel an individual to repay an act of kindness even when the cost of repayment exceeds the benefit initially bestowed on the actor (Hirshleifer, 1987). So strong are these emotional influences on behavior that certain sentiments, such as greed or lust, can even motivate an individual to disobey these same moral standards:

The man who acts according to the rules of perfect prudence, of strict justice, and of proper benevolence, may be said to be perfectly virtuous. But the most perfect knowledge of those rules will not alone enable him to act in this manner; his own passions are very apt to mislead him — sometimes to drive him, and sometimes to seduce him, to violate all the rules which he himself, in all his sober and cool hours, approves of. (Smith, 1759/2000, p. 349; italics added)

In essence, Smith (1759/2000) argued that our capacity to obey normative standards of appropriate behavior was more often determined by our experience of certain moral sentiments than our employment of rational deduction and logical inference. Depending on the nature of the particular emotion that one experienced (e.g., greed vs. guilt), Smith argued that moral sentiments could compel an individual to pursue or avoid a particular course of action even when this choice ran contrary to one's immediate economic self-interest.

Although contemporary economists and psychologists are more concerned with earthly rewards than heavenly utilities, Smith's insights regarding moral sentiments can be applied to questions that interest contemporary decision theorists. For example, why do emotions often compel individuals to pursue economic strategies that do not appear to be in their immediate material self-interest? Along these lines, there is an ever-expanding body of theoretical and empirical research on self-focused emotions and economic decision making that is consistent with Smith's insights regarding the influence of moral sentiments on decision making (Ketelaar, 2004; Ketelaar & Au, 2003; Van Kleef et al., 2004). For example, economist Robert Frank's (1988, 2004) theory of emotions as commitment devices essentially argues that the experience of self-focused emotions such as guilt can explain why some individuals forgo the immediate rewards associated with cheating in favor of more cooperative strategies that pay their benefits, not immediately, but rather in the long run. According to this view, moral sentiments can:

1Schadenfreude is a German term that refers to the experience of positive emotions while observing another's misfortune.
how to behave (Ketelaar & Clore, 1997). One source of information (from deliberate economic calculations) informs the individual about the immediate material consequences of a given strategy, and the second source (from the intuitive moral emotions) informs the individual about the future (in this case, negative) consequences of that particular strategy choice. In this manner, the mental pain of guilt can compel an individual to not cheat even when it is in their immediate material interest to do so (Frank, 1988).

For over one decade, this view of moral sentiments (Frank, 1988) has been so compelling that it has been accepted at face value without being subject to empirical test. Recently, research at the intersection of social psychology and experimental economics has begun exploring the role of moral sentiments in economic decision making (Ketelaar & Au, 2003; Van Kleef et al., 2004). Consistent with this emotions-as-commitment-devices view (Frank, 1988, 2004; Hirsheifer, 1987), several studies have shown that individual differences in the capacity to experience feelings of guilt can translate into rather large differences in the magnitude of monetary offers made in the final round of a repeated Ultimatum Game (Haley & Fessler, 2004; Ketelaar & Au, 2003). In one study, over 91% of individuals who felt guilty after proposing an unfair monetary offer in the first round of play were later observed to make a generous offer in the second round (Ketelaar & Au, 2003). By contrast, less than 22% of individuals who reported no feelings of guilt (after proposing a similarly unfair offer) were observed to make a generous offer in the second round. Similar findings have been obtained in a repeated Prisoner's Dilemma game, where it was observed that individuals who felt guilty after behaving noncooperatively subsequently displayed 25% more cooperation compared with those who felt no guilt (see Ketelaar & Au, 2003). Finally, in studies of postdecision regret in the Ultimatum Game, it has been shown that feelings of regret are a good predictor of subsequent offers when the game is repeated with another partner (Zeelenberg & Beattie, 1997).

In one study, participants reported their level of regret after receiving feedback informing them that their offer was either (a) only slightly higher than the responder's minimal acceptable offer, or (b) much higher than the responder's minimal acceptable offer. Not surprising, participants who received feedback that their offer was much higher than the responder's minimal acceptable offer reported significantly more regret than participants who were informed that their offer was only slightly higher than the responder's minimal acceptable offer. Interestingly, when analyzed together, only feelings of regret (and not feedback about minimal acceptable offers) were able to predict the magnitude of proposers' offers in the subsequent Ultimatum Game.

Taken together, research on guilt and regret suggests that individuals differ in their capacity to experience particular moral sentiments, and these individual differences are predictive of subsequent decision behavior (Schwartz, Ward, Monterosso, Lyubomirsky, White, & Lehman, 2002; Tangney, 1991). If self-focused moral emotions (e.g., regret, guilt) are a universal solution to the so-called commitment problem (Frank, 1988), one might wonder why we observe these individual differences in the capacity to experience guilt and regret? Nowhere is this puzzle regarding individual differences in moral sentiments more evident than in research examining other-focused emotions and norm-enforcing behavior.

Other-Focused Emotions and Norm-Enforcing Behavior

The existence of individuals willing to enforce norms is an intriguing paradox for any model of decision making that treats individuals as purely self-interested agents. This is the case because many norm-enforcing behaviors, ranging from specific acts of punishment to credible threats directed at potential norm violators, appear to be altruistic acts (Fehr & Gächter, 2002). The logic behind this claim is straightforward: Individuals who display threatening signals and/or perform acts of punishment must directly incur the costs of carrying out these actions, yet the benefits accrued by these norm-enforcing behaviors are often distributed across the entire social group. In this manner, one might argue that norm-enforcing behaviors constitute a form of second-order public good (Fehr & Fischbacher, 2004). As such, one expects that most self-interested agents will refuse to contribute to this public good (fail to threaten or punish violators) and instead will free ride on the actions of a few punisher types who enforce social norms for the entire group. Interestingly, experimental economists have demonstrated that a significant portion of individuals are often willing to incur a cost to inflict punishment on others who violate social norms such as undercontributing in a public goods game (Fehr & Fischbacher, 2004; Fehr & Gächter, 2002). This section explores the claim that emotions operate as the proximate psychological mechanisms underlying these norm-enforcing behaviors (Fehr & Gächter, 2002).

Just as certain moral sentiments seem to play a central role in motivating individuals to obey norms, other sentiments appear to play a central role in motivating individuals to enforce these norms. In particular, emotion researchers have identified three other-focused emotions—contempt, anger, and disgust—that are activated by detecting violations of normative standards (Rozin, Lowery, Imada, & Haidt, 1999). In addition to being associated with distinct facial signals, these three moral sentiments appear to be triggered by distinct types of norm violations (Haidt, 2003). Contempt, for example, has been shown to be associated with violations of normative rules regarding community standards and customs, as when one observes someone failing to carry out his or her duties in the community or social hierarchy.
Anger, on the other hand, is associated with violations of normative standards governing one's personal rights, as when one directly experiences another person infringing on his or her personal liberties. Finally, disgust appears to be triggered by violations of normative standards governing purity and divinity, as when one observes another person disrespecting culturally shared sacred beliefs or religious traditions. Moreover, these three moral sentiments appear to be excellent candidates for relatively universal, culturally shared reactions to these particular types of norm violations (Rozin et al., 1999).

This linkage between specific moral sentiments and specific types of norm violations suggests that the same event can evoke any one of these three moral sentiments depending, of course, on how the event is perceived. Consistent with this view, a sample of registered Democrats were observed in one study to spontaneously display a range of moral emotions (contempt, anger, and disgust) when asked how they felt about the very same event: the recent presidential election in which the Republican candidate had won (Ketelaar, 2005). In this context, one supposes that those individuals who reacted with contempt displays (62.5% of the sample) were focusing on the impact that the election results had on their social group or community. Similarly individuals who reacted with anger (12.5%) may have been focusing on the impact that the election results had on their personal rights, whereas individuals who reacted with disgust (25%) were presumably focusing on the impact that the election results had on fundamental beliefs that they held sacred. Such an interpretation would be consistent with the view that contempt, anger, and disgust involve subjective perceptions that give rise to systematic experiences of specific moral emotions in response to violations of specific types of normative standards of appropriate behavior. The clear empirical evidence for several distinct other-focused moral sentiments (contempt, anger, disgust; see Haidt, 2003) leads one to ask why an individual would possess powerful emotions to condemn behaviors in others that (the observer) are rarely or never motivated to perform? Before discussing the intriguing question of whether it is in one's material self-interest to experience these other-condemning emotions, I briefly review empirical research that suggests that these other-directed moral sentiments can indeed exert a strong influence on economic decision making.

Several studies have demonstrated a clear link between anger and punitive behavior in economic bargaining games. In one study of investment behavior in a public goods game, punishment took the form of the ability to deduct monetary earnings from other members of the collective after each round of play (Fehr & Gächter, 2002). Interestingly these acts of punishment could only be made at a cost to the punisher and resulted in no monetary gain for the punishing party. Consistent with the view that anger acts as a moral sentiment that compels an individual to punish individuals who violate normative standards governing one's personal rights, these punishing acts were significantly correlated with self-reports of anger and were typically executed by above-average contributors and inflicted on below-average contributors (Fehr & Gächter, 2002).

Anger has also been shown to motivate individuals to punish norm violators by spitefully rejecting monetary offers that appear to violate norms of fairness. In one study of unfair ultimatum offers, it was observed that self-reported feelings of anger were a better predictor of rejections of unfair offers than actual perceptions of unfairness (Pillutla & Murmann, 1996). One reason that angry reactions may not be synonymous with perceptions of unfairness may be due to the fact that some unfair offers are not the result of intentional actions, and thus fail to elicit the appropriate attributions of responsibility and blameworthiness that some theorists would claim are necessarily but not sufficient conditions for producing anger (Lerner, Goldberg, & Tetlock, 1998; Ortony, Clore, & Collins, 1988; but see Kuppens, Van Mechelen, Smits, & de Bock, 2003). Along these lines, other studies have shown that feelings of anger are strongly correlated with attributions of responsibility and blameworthiness directed toward one's negotiation partner (Allred, Mallozi, Matsui, & Rai, 1997). Consistent with the claim that anger is a norm-enforcing emotion, other studies have manipulated participants' perceptions of the emotional state of their interaction partner (angry vs. happy) and found that participants make greater concessions when they believe their negotiation partner is angry than when they believe their partner is happy (van Kleef et al., 2004). In short, there appears to be a clear link between the experience of anger and the motivation to punish individuals who have violated norms of appropriate behavior, even when these acts of punishment entail a cost for the angry enforcer. Because anger is just one of several other-directed moral sentiments, future research might explore the possible norm-enforcing role of additional other-focused emotions such as contempt, disgust, and schadenfreude.

Game-Theoretic Insights Into Emotion-Based Strategy Types

From the perspective of normative models of rational decision making, it is clear that emotions are often associated with a number of peculiar economic behaviors. Guilt-prone individuals often forgo monetary rewards by obeying norms that prescribe cooperation rather than selfishness (Ketelaar & Au, 2003), and angry individuals often enforce norms by punishing norm violators even when doing so provides a benefit to the group at a significant cost to the punisher (Fehr & Gächter, 2002). As paradoxical as these norm-obeying and norm-enforcing acts may appear to be, it must be emphasized that not all individuals engage in such behavior. Thus, one might argue that a complete understanding of the influence of emotion on norm-obeying and
norm-enforcing behavior must also explain why only a certain portion of the population behaves in this manner. In this final section, I review several evolutionary game-theoretic insights regarding how individual differences in the capacity to experience certain moral sentiments could be functionally linked to individual differences in strategic economic behavior (Ketelaar, 2004).

One of the most compelling explanations for the apparently irrational influence of emotion on economic behavior centers on the claim that certain emotions function as commitment devices that compel individuals to maximize long-term payoffs at the expense of sometimes forgoing immediate rewards (Frank, 1988, 2004; Hirshleifer, 1987). However, this commitment device view of emotion does not explain why certain individuals (e.g., guilt-free individuals and emotionless sociopaths) lack these important emotional commitments. In other words, if one wishes to explain the disruptive influence of emotion on economic decision making by saying that certain emotions, such as guilt or gratitude, serve important functions, one can ask why all individuals do not experience these putatively functional sentiments. In one study of the Ultimatum Game, for example, only 57% of individuals reported experiencing feelings of guilt after proposing an unfair split of the money (Ketelaar & Au, 2003). However, these individual differences in emotion resulted in quite striking differences in ultimatum offers in the second round of play. When the Ultimatum Game was repeated 1 week later, over 91% of individuals who felt guilty (the previous week) now made generous offers, whereas only 22% of individuals who reported no feelings of guilt made similarly generous offers.

Such strategically relevant individual differences are not limited to studies of emotional influences on norm-obeying behavior; they also appear to play a prominent role in studies of norm-enforcing behavior. In one study of altruistic punishment, for example, approximately 16% of the sample never punished other group members even when they contributed significantly less than the group average in a public goods game. Moreover, only one third of the sample consistently (across several rounds of play) punished group members who made deviant contributions, and these costly acts of punishment were shown to be strongly linked to individual differences in the experience of anger directed toward low contributors (Fehr &.Gächter, 2002). In summary, not all individuals are compelled by feelings of guilt to resist the temptation to defect (Ketelaar & Au, 2003), and not everyone becomes angry enough to incur the cost of punishing norm violators (Fehr & Gächter, 2002). Although the emotions-as-commitment-device approach is able to explain how emotions can compel individuals to obey norms or punish norm violators, such models leave open the question of why not everyone behaves as if they are emotionally committed to these norm-obeying and norm-enforcing actions.

6. EMOTIONS AND ECONOMIC DECISION MAKING

Strategies as Decision Rules

Although Adam Smith (1759/2000) explained individual differences in the capacity to experience moral sentiments in terms of a failure to exercise self-command over these emotions, one can also utilize evolutionary game theory to explore another, more intriguing explanation for these individual differences. Specifically, one can examine the possibility that individual differences in the capacity to experience moral sentiments may represent a relatively stable (frequency-dependent)\(^2\) distribution of emotion-based strategy types (Ketelaar, 2004; Mealey, 1995). One way to think about individual differences in emotion-based strategy types is to think of strategies as analogous to computer-programmed decision rules (Binmore, 1998). According to this way of thinking, there are at least two ways that multiple strategies (i.e., individual differences in strategic behavior) can manifest themselves in a population of agents. First, a population may contain multiple behavioral strategies because the population consists of several distinct strategy types, each corresponding to a distinct decision rule governing each agent’s behavior (e.g., compare “IF you cooperate, THEN I will cooperate” to “IF you cooperate, THEN I will defect”). A second reason that one might observe different behavioral strategies in the same population concerns the presence of certain types of reactive strategies (see Tooby & Cosmides, 1990a, 1990b). In other words, a population may consist of just one strategy type, but this strategy type may correspond to a complex decision rule that generates different behavioral phenotypes contingent on the phenotypic characteristics of the other agents that it encounters. For example, imagine a decision rule of the form “IF I am weaker than you AND IF you cooperate, THEN I will cooperate; ELSE IF I am stronger than you AND IF you cooperate, THEN I will defect.” A reactive strategy of this form appears cooperative when interacting with strong opponents and noncooperative when encountering weak opponents. The presence of such a reactive strategy might lead one to the erroneous conclusion that their exists two strategy types in the population (cooperators and noncooperators), when, in reality, the population is comprised of just one strategy type that reacts differentially to particular phenotypic characteristics (e.g., relative strength) that vary across agents.

Finally, there exists a third means of obtaining a diversity of strategy types in a single population of agents, and I argue that this explanation may be key to understanding the influence of emotions on economic decision making. This third explanation for multiple strategies is a hybrid of the two types of

\(^2\)Frequency-dependent selection occurs when the evolutionary success of a strategy type is dependent on the relative frequency of certain other strategy types in the population (Maynard Smith, 1982).
explanation discussed earlier. According to this hybrid view, all agents possess essentially the same reactive decision rule for determining strategic behavior (e.g., IF X THEN Y, but only IF B > C), yet each agent differs in terms of the actual values they assign to certain parameters in this decision rule. To appreciate this idea, imagine an evolved decision rule comprised of a number of parameters (X, Y, B, C, etc.), some of which are invariant across individuals and others that vary systematically across individuals. Consider, for example, a decision rule that essentially translates into the following form: "IF you fail to contribute to the public good by some factor X, THEN I will punish you by some factor Y, but only IF the benefit B of inflicting this punishment is greater than the cost C associated with punishing you." Further imagine that the amount of punishment Y that is inflicted on a norm violator is a function of the degree of norm violation entailed by X and, moreover, that this parameter is some fixed function \( Y = f(X) \) shared by all agents in the population. Thus, all agents who encounter a given instance of norm violation will necessarily punish the same amount as determined by the function \( Y = f(X) \).

However, now imagine that the two cost–benefit parameters (C and B in this hypothetical decision rule) are determined by natural selection according to their evolutionary success. Because agents are constantly interacting with each other across time and space, one might imagine that natural selection could eventually converge on a single algorithm that determines how all agents will calculate these cost–benefit parameters (C and B). Because the algorithm that calculates these cost–benefit parameters would essentially correspond to a sort of threshold for determining whether an agent will punish a particular instance of norm violation, if all agents possess the same algorithm for calculating these parameter values, all agents will essentially possess the same behavioral phenotype in regard to punishing norm violators. By contrast, if different agents possessed somewhat different algorithms for calculating these parameters (costs and benefits of punishing), this would necessarily generate several distinct behavioral phenotypes (i.e., punishers and nonpunishers). Therein lies the rub, or at least a possible evolutionary source for individual differences in punishment behavior or any other strategic behavior for that matter. The only difference between this explanation for multiple strategies and the reactive strategy explanation (discussed earlier) is the subtle point that several distinct strategy types can emerge merely as a result of subtle differences in how certain parameter values are calculated in the agent's decision rule. How might this work?

If agents are analogous to preprogrammed decision rules and evolutionary selection pressures do not converge on just one method for calculating parameter values for all agents, the population, by definition, will consist of several different strategy types. For example, if evolutionary dynamics allowed for different agents to possess different values for the cost–benefit parameters contained in the hypothetical punishment decision rule "IF X THEN Y, but only IF B > C," this could result in a diversity of strategy types in regards to their propensity to punish norm violators. Depending on how these cost–benefit parameters varied, some versions of this rule would dictate that certain agents never punished (punishment would always be perceived as too costly), whereas other versions of this rule, with somewhat different cost–benefit parameters, would dictate that certain agents would invariably punish every instance of norm violation that they encountered (it would always be perceived as beneficial to punish).

Again one could imagine that evolutionary forces might invariably converge on just one method of calculating these cost–benefit parameters for all agents, whereby the population of agents (decision rules) that emerges is at some sort of game-theoretic equilibrium. Yet it turns out that the ecological conditions under which such decision rules often operate (i.e., indefinitely repeated interactions) lend themselves to a vast number of possible equilibria and, thus, an equally vast number of successful decision rules. The implication of this insight—known as the folk theorem—is that evolutionary selection pressures are capable of generating a vast number of successful distributions of strategy types (each consisting of a distinct decision rule) and, thus, one must consider the possibility that human populations may be comprised of a multitude of decision rules (not just one) governing the implementation of norm-obeying and norm-enforcing behaviors.

Evolutionary game theory gives us several hints at how dynamic forces (such as those encapsulated in the concept of frequency-dependent selection and the folk theorem) could give rise to more than one equilibria in a single population of agents (Binmore, 1998; Boyd, 1989; Boyd & Richardson, 1992; Maynard Smith, 1982). For example, if one considers strategy types as analogous to finite decision rules (finite automata) that encounter one another in indefinitely repeated games, an interesting phenomenon emerges. Specifically, one observes that:

any two finite automata playing each other for long enough in a repeated game will eventually end up cycling through the same sequence of plays. A player's per-game payoff can therefore be calculated simply by taking the average payoff in a cycle. (Binmore, 1998, p. 301)

For example, if the finite strategy of tit for tat interacts with itself (or any other finite strategy that is never the first to defect), these two strategies will
end up in a continuous cycle of mutual cooperations and obtain whatever payoffs are associated with this particular cycle. By contrast, tit for tat interacting with a finite strategy of always defect will end up in a continuous cycle of mutual defections and obtain a necessarily different set of payoffs associated with this very different cycle. The folk theorem emerges from the simple fact that, because finite strategies invariably end up infinitely cycling through a particular sequence of outcomes (mutual cooperation, mutual defection, etc.), this sets up an incentive for either agent (i.e., each finite decision rule) to locate and shift to an alternative pairing of strategies that can generate a different cycle with a greater payoff than that obtained in their current cycle (Binmore, 1998). It follows that indefinitely repeated games that occur in a diverse population of strategy types can allow for a great variety of equilibria simply because this context (indefinitely repeated interactions, different strategy types) allows for a vast number of more profitable alternative cycles on which various strategy types can coordinate (see Binmore, 1998; Lomborg, 1996, for a fuller treatment of this so-called folk theorem). This argument depends, of course, on agents being able to maintain credible commitments to sticking with a particular alternative cycle of payoffs (Lomborg, 1996). The claim here is that emotions are a plausible source of these credible commitments (Frank, 1988, 2004; Hirsheifer, 1987, 2002; Ketelaar, 2004). Because individuals who experience different emotions (e.g., guilt vs. no guilt, anger vs. no anger) often end up making quite different choices in economic bargaining games, it may be useful for future research to explore the possibility that individual differences in moral sentiments constitute one source of these individual differences in strategy types.

CONCLUSIONS

Moral sentiments—such as feelings of anger—are easily labeled as irrational when they are observed to motivate individuals to engage in behaviors (e.g., altruistic punishment) that benefit others, but are costly to the individual who performs them. To the degree that the psychological mechanisms underlying our emotions have been shaped by natural selection, one wonders how adaptively designed moral sentiment systems could routinely generate preferences that compel individuals to pursue outcomes with lower (rather than higher) payoffs. This chapter argues that an intriguing answer to this question may lie in the realization that not all individuals confront identical payoffs even when they find themselves in the same situation (see Kelley, 1984; Kelley & Thibaut, 1978). For example, an agent who embodies the angry punisher type may experience a quite different set of payoffs associated with inflicting punishment as compared with a relatively angerless, nonpunishing type. In this manner, if one is forced to select just one set of payoffs as the rational way of viewing the utilities associated with punishing and not punishing, by definition, one of these two agents (strategy types) will be viewed as possessing irrational preferences. Yet if the payoffs associated with punishment are determined not by one theoretical gold standard, but rather by virtue of the evolutionary (including frequency-dependent) success of the specific parameter settings contained in an agent's decision rule, one might wonder whether the specific parameter values, and the corresponding "payoff estimates" that they generate, must necessarily be identical for all agents.

Just as the predicted payoff for approaching elephant dung ineluctably depends on whether you are a human or a dung beetle, it may turn out that the ultimate future payoff associated with punishing norm violators depends on which strategy type you are, as well as the particular distribution of other strategy types in the same environment. In a population comprised largely of defectors, for example, it is more costly to be a punisher than in an environment comprised largely of norm-obeying agents. Along these lines, it was argued that individual differences in the capacity to experience certain moral sentiments might correspond to frequency-dependent individual differences in the settings (values) of certain decision rule parameters that influence norm-obeying and norm-enforcing behavior. The obvious implication is that, by virtue of their ability to generate individual differences in how payoffs are assigned to outcomes, emotion-based strategies might play a significant role in explaining why some decision makers often do not act in the strictly rational manner that some normative models suggest that they should behave. Along these lines, the present chapter argued that a large part of the puzzle concerning why irrational emotions so often intrude on our most important decisions lies in recognizing this important link between moral sentiments and judgments of normatively acceptable and unacceptable behavior.

REFERENCES


6. EMOTIONS AND DECISION MAKING


