Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System

Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System

Louis Kaplow and Steven Shavell

Harvard University and National Bureau of Economic Research

How should moral sanctions and moral rewards—the moral sentiments involving feelings of guilt and of virtue—be employed to govern individuals' behavior if the objective is to maximize social welfare? In the model that we examine, guilt is a disincentive to act and virtue is an incentive because we assume that they are negative and positive sources of utility. We also suppose that guilt and virtue are costly to inculcate and are subject to certain constraints on their use. We show that the moral sentiments should be used chiefly to control externalities and further that guilt is best to employ when most harmful acts can successfully be deterred whereas virtue is best when only a few individuals can be induced to behave well. We also contrast the optimal use of guilt and virtue to optimal Pigouvian taxation and discuss extensions of our analysis.

I. Introduction

The influence of morality on behavior has been a long-standing theme of the analysis of human conduct. Adam Smith (1790), among others, emphasized the motivational importance of the moral sentiments: feel-

We thank the editor, Steven Levitt, the referees, Jonathan Baron, Gary Becker, Robert Ellickson, Edward Glaeser, Oliver Hart, A. Mitchell Polinsky, Eric Posner, Eric Rasmusen, Richard Zeckhauser, and participants in workshops at Berkeley, Carnegie Mellon, Chicago, Harvard, Pittsburgh, Stanford, and the NBER for comments; Anthony Fo, Felix Gilman, Clint Keller, Allon Lifshitz, Christopher Lin, Damien Matthews, Jeff Rowes, Laura Sigman, and Leo Wise for research assistance; and the John M. Olin Center for Law, Economics, and Business at Harvard Law School for financial support. A more extensive treatment of the subject of this article appears in Kaplow and Shavell (2001).

[*Journal of Political Economy*, 2007, vol. 115, no. 3]
 0022-3808/2007/11503-0005$10.00

ings of guilt and of virtue along with their external correlates, disapprobation and praise. Recent economic literature on social norms and behavioral economics, such as that examining concerns for "fairness," recognizes that individuals' actions are not always narrowly self-interested and may reflect moral concerns.[1] Likewise, significant emerging literatures in evolutionary biology and other disciplines address the role of moral emotions in regulating behavior.[2] Against this background, we assume here that individuals have the capacity to care about the morality of their actions and accordingly that their decisions may be influenced by moral factors.

The question that we examine is how a social planner would design a system of morality to maximize social welfare, that is, how a social planner would associate moral feelings with acts so as to induce behavior that fosters social welfare. This is a natural question to pose for several reasons. First, it seems that moral rules are to some degree malleable, for we see that societies devote effort to inculcate moral views and that these views vary across societies and over time. Second, many natural and social scientists believe that our moral capacities are not accidental but rather are the product of natural selection, serving the function of furnishing humans with incentives to cooperate, to punish transgressors, and to behave in additional ways that promote survival.[3] Third, the systems of morality that actually exist appear in a broad sense to be welfare enhancing, as Hume (1751), Sidgwick (1907), and others have argued.

In our model, which we present in Section II, individuals decide whether or not to commit various acts, some of which produce external effects. Individuals are subject to a process of moral inculcation such that they will experience moral sentiments—guilt and virtue—as a function of the choices they make. Accordingly, they will be led to behave other than in their narrow self-interest if the weight of guilt or virtue or both is sufficiently great. We assume that the inculcation process is socially costly and that the capacity of individuals actually to experience these moral sentiments is limited. We further assume that guilt and virtue cannot be independently specified for every conceivable situation but rather must be inculcated uniformly over certain subsets of acts. (For example, the inculcation of guilt for lying cannot be so nuanced that it results in a different level of guilt for each conceivable lie, but

[1] See Hirshleifer (1987), Kahneman, Knetsch, and Thaler (1987), Frank (1988), Rabin (1993), Ben-Ner and Putterman (1998), Binmore (1998), Fehr and Schmidt (1999), Akerlof and Kranton (2000), Becker and Murphy (2000), and Robson (2001).

[2] See Darwin (1872, 1874), Trivers (1971), Campbell (1975), E. O. Wilson (1975), Alexander (1987), Daly and Wilson (1988), Izard (1991), Barkow, Cosmides, and Tooby (1992), J. Q. Wilson (1993), Baron (1994), Damasio (1994), LeDoux (1996), Pinker (1997), Elster (1998), Haidt (2001), Massey (2002), Cohen (2005), and Hauser (2006).

[3] See the references in note 2.

rather must be the same within certain categories of lies.) The social objective is taken to be maximization of morally inclusive social welfare, which is to say conventional social welfare—the utility that individuals obtain directly from the acts that they commit and the utility consequences of any external effects of these acts—combined with morally related components of social welfare—the cost of instilling the moral sentiments and the disutility or utility associated with experiencing them.

We analyze the optimal system of morality in Section III. We find that the moral sentiments are used chiefly to control individuals' behavior when otherwise it would not be first-best because of externalities.[4] However, the moral sentiments can only imperfectly correct behavior that generates externalities, because of the cost of inculcating the moral sentiments, limits on the capacity of individuals to experience them, and also the need to instill guilt and virtue at uniform levels over groups of acts that may exhibit heterogeneity. Regarding the latter, suppose that most but not all acts in a group are undesirable; then some desirable acts in the group may be deterred by the moral sentiments, and other desirable acts will be committed but result in individuals' feeling guilty as a consequence. An additional conclusion is that, for a variety of reasons, the optimal level of the moral sanctions may be lower or higher than a Pigouvian tax benchmark, under which moral sanctions would be set equal to the level of the relevant expected externality. That is, moral leniency or moral harshness can be optimal.

A further conclusion concerns the optimal choice between the use of guilt and of virtue. Because individuals are assumed to have a limited capacity to experience these sentiments, it is best to use them in such a manner that they are actually experienced infrequently. Hence, guilt is best to employ when most violations of moral rules can be successfully deterred, whereas virtue is best to employ when few individuals can be induced to follow a moral rule. In other words, it is optimal for worse-than-normal moral behavior to be punished and for unusually good moral behavior to be rewarded.

In Section IV, we consider possible extensions of our analysis, and in Section V we comment on the consistency of our conclusions with the observed features of moral systems.[5]

[4] The externalities might be positive, such as when virtue is used to encourage individuals to help others in distress or guilt is used to induce people to contribute to the provision of public goods rather than to free ride.

[5] On the connection between our analysis and moral philosophy, see Kaplow and Shavell (2001, 2002).

II. Model

Let S denote the set of possible situations in which individuals may find themselves. In each situation, an individual chooses between committing some act and not doing so. For example, in one situation, an individual might choose whether or not to lie, in another whether or not to litter, and in another whether or not to read a book.

If in any particular situation an individual commits the act, the individual obtains (positive or negative) utility u, which we sometimes refer to as conventional utility. The commission of an act causes an external harm of $h \geq 0$. Note that this assumption allows acts with no external effects, and it also implicitly accommodates acts with beneficial external effects (through appropriate labeling of acts).[6] If the individual does not commit the act, he does not obtain any conventional utility and does not cause any external harm (these assumptions are purely a normalization in that only the difference in utility and externalities caused by acts versus inaction is relevant).

A situation in S may thus be identified with a pair (u, h) describing the act that an individual may choose to commit. The possible situations have density $f(u, h)$, which is assumed to be continuous, where u is in $(-\infty, \infty)$ and h is in $[0, \infty)$.

Assume that society may instill the predisposition to experience guilt $g(u, h) \geq 0$ for committing an act in situation (u, h). That is, a person in situation (u, h) will experience guilt and suffer disutility of $g(u, h)$ if and only if he commits the act. Similarly, assume that society may instill the predisposition to experience virtue $v(u, h) \geq 0$ for not committing an act in situation (u, h). That is, a person in situation (u, h) will obtain utility of $v(u, h)$ if and only if he does not commit the act.

The prospect of experiencing guilt or virtue may lead an individual to change his behavior.[7] In the absence of guilt and virtue, an individual in a given situation will commit the act if and only if $u > 0$.[8] When guilt $g(u, h)$ is instilled for acting and virtue $v(u, h)$ for not acting, an individual will act if and only if the overall utility from acting exceeds the

[6] For example, failing to assist others can be labeled "acting," which causes a negative externality ($h > 0$) relative to "not acting," in this case, assisting others. (Assuming that $h \geq 0$ does involve some restriction given our later assumption about the grouping of acts; relaxing the assumption that h is nonnegative would not, however, affect the qualitative nature of our conclusions.)

[7] It is not important for our analysis how individuals actually conceive of guilt and of virtue, i.e., whether moral considerations are in some psychological sense different from sources of conventional utility. It is likely, however, that moral considerations do have distinctive psychological attributes. Many of the references cited in note 2 characterize the moral sentiments as emotions, although that literature varies considerably in how these emotions are labeled, which of these are viewed as purely internal and external, and other respects.

[8] For convenience, we assume throughout that individuals do not act when they are indifferent.

utility from not acting, that is, if and only if $u - g(u, h) > v(u, h)$, or $u > g(u, h) + v(u, h)$, which is to say that conventional utility exceeds the sum of the moral sanction and reward that favor not acting.

We make three assumptions about guilt and virtue. First, when guilt and virtue are instilled, they are constrained to be the same for all situations within each of n exogenously given subsets S_i that partition the universe S of situations. Let g_i and v_i denote the uniform levels of guilt and of virtue for situations within S_i. Furthermore, let $f_i(u, h)$ denote the conditional density of (u, h) on S_i, and let p_i be the probability that a situation is in S_i.

The motivation for the assumption that guilt and virtue are constant for each S_i is that it is difficult to instill guilt and virtue in a completely tailored manner. As we discuss in Section IV.*A* below, a more elaborate model might allow the determination of the S_i to be endogenous, by permitting greater refinements of the S_i at an additional cost; but as long as perfect refinement is not optimal, our conclusions would not be affected.

Second, there is a cost of instilling guilt and virtue. Specifically, g_i may be instilled on each subset S_i at cost $\alpha_i(g_i)$, where $\alpha_i'(g_i) > 0$ and $\alpha_i''(g_i) \geq 0$. Similarly, v_i may be instilled at cost $\beta_i(v_i)$, where $\beta_i'(v_i) > 0$ and $\beta_i''(v_i) \geq 0$.[9]

Third, there is a limit to individuals' capacity to experience guilt and virtue. Specifically, we assume that the expected value of experienced guilt cannot exceed $G \geq 0$ and the expected value of experienced virtue cannot exceed $V \geq 0$.

The motivation for positing this limitation derives from human psychology. With regard to a range of feelings and stimuli, our neurological systems tend to become less sensitive or even numb to repetition of the same experience. This phenomenon suggests that there would be a "crowding-out" effect on further feelings of guilt or virtue as their frequency and magnitude increase.[10] It would be cumbersome to take the foregoing explicitly into account, and doing so would not alter the qualitative nature of the conclusions that we derive using the aforementioned simple constraints on experienced G and V. (An alternative

[9] Instead, one could assume that the total of guilt and virtue inculcated might determine the cost, and one could also incorporate a crowding-out phenomenon, in which spending more time to inculcate guilt or virtue for some types of acts leaves less inculcation time (or less effective time) for other types of acts. However, because we do not specify the level of inculcation costs and because our constraints on the experiencing of guilt and virtue (described next) have an aggregate form, introducing trade-offs among acts in the inculcation process would not change our results significantly.

[10] For example, Frederick and Loewenstein (1999) suggest that there is a substantial (although not universal) regularity in the tendency of mental reactions to stimuli to fall as the stimuli are repeated. Our approach in this regard is similar in spirit to that advanced in Rayo and Becker (2007).

version of our model under which the increased use of guilt or virtue decreases its marginal effectiveness is sketched in the margin.[11]) We also note that these constraints will be seen to have importance in regard to the choice between the use of guilt and of virtue.

Social welfare is taken to be the expected value of the conventional utility that individuals experience from committing acts, plus any realized virtue and minus any realized guilt, minus external harm, and minus the costs of instilling guilt and virtue. (Although we adopt a utilitarian framework, our conclusions would be essentially the same under a variety of other frameworks.[12]) As noted in the introduction, we refer to this measure as morally inclusive social welfare to distinguish it from the standard formulation of social welfare that incorporates only conventional utility and externalities. In the familiar first-best solution to the usual problem of social welfare maximization, an act in a situation is committed if and only if $u > h$.

III. Analysis

The social problem is to choose $g_i \geq 0$ and $v_i \geq 0$ on the subsets S_i to maximize social welfare, subject to the constraints on the realization of guilt and virtue. Social welfare is[13]

[11] We let g_i continue to indicate how much guilt is inculcated for situations in S_i, but we introduce a separate term, $\gamma_i(g_i, G)$, to indicate effective guilt—the level of disutility—which in turn influences behavior. (Thus, in a model with guilt only, an individual commits an act if and only if $u > \gamma_i$.) In this formulation, G is not a constraint on guilt that may be experienced but rather refers to the total amount of guilt that will be experienced. Finally, assume that γ_i is increasing in g_i and decreasing in G. For example, one might have $\gamma_i(g_i, G) = g_i/(1 + G)$. Then, if more guilt is used on subset S_i, G will increase, which will decrease the effectiveness of guilt in controlling all behavior. The first-order conditions for this model are similar to (5) and (6) below. The primary difference is that the shadow prices would be replaced by terms reflecting an increasing marginal cost of experienced guilt or virtue (corresponding to the diminished marginal effectiveness of guilt or virtue that is already employed to control other acts). Additionally, in such a model—in which guilt and virtue are not literally fixed in supply—the fact that experiencing guilt and virtue affects welfare would have a more clearly identifiable effect on the optimum: Ceteris paribus, this consideration favors using less guilt and more virtue than otherwise.

[12] Our measure of external harm h, conceived here as the aggregate reduction in others' utilities, could also be interpreted as a nonwelfarist measure of social harm from an act, such as from rights violations. Likewise, our results do not depend on the particular choice of an additive welfare function; it would be sufficient for social welfare to depend positively on different individuals' u's, negatively on h, and so forth.

[13] Expression (1) may be interpreted as the welfare of a representative individual. Alternatively, expression (1) may be interpreted as the average welfare of a group of possibly heterogeneous individuals, an extension that we discuss in Section IV.*C*. In this case, however, the constraints (2) and (3) would need to be formulated somewhat differently.

$$\sum_{i=1}^{n} W_i(g_i,\ v_i) = \sum_{i=1}^{n} \left\{ p_i \left[\int_0^{\infty} \int_{g_i+v_i}^{\infty} (u - h - g_i) f_i(u,\ h) du dh + \int_0^{\infty} \int_{-\infty}^{g_i+v_i} v_i f_i(u,\ h) du dh \right] - \alpha_i(g_i) - \beta_i(v_i) \right\}. \quad (1)$$

To explain, individuals commit acts in situations in subset S_i when $u > g_i + v_i$, in which case the effect on social welfare is $u - h - g_i$ because both conventional utility and guilt are experienced by the individual committing an act and the externality occurs; when individuals do not commit acts, they obtain utility of v_i; and the costs of instilling g_i and v_i are subtracted. The constraints on the realization of guilt and virtue are

$$\sum_{i=1}^{n} y_i(g_i,\ v_i) = \sum_{i=1}^{n} \left[p_i \int_0^{\infty} \int_{g_i+v_i}^{\infty} g_i f_i(u,\ h) du dh \right] = \sum_{i=1}^{n} p_i g_i [1 - F_i(g_i + v_i)] \le G \quad (2)$$

and

$$\sum_{i=1}^{n} z_i(g_i,\ v_i) = \sum_{i=1}^{n} \left[p_i \int_0^{\infty} \int_{-\infty}^{g_i+v_i} v_i f_i(u,\ h) du dh \right] = \sum_{i=1}^{n} p_i v_i F_i(g_i + v_i) \le V. \quad (3)$$

Here, $y_i(g_i, v_i)$ and $z_i(g_i, v_i)$ are, respectively, expected experienced guilt and virtue from situations in S_i; $F_i(g_i + v_i)$ is the frequency with which $u \le g_i + v_i$ on the subset S_i, that is, the fraction of acts in S_i that are deterred; and, correspondingly, $1 - F_i(g_i + v_i)$ is the fraction of acts in S_i that are committed.

The Lagrangian for the problem of maximizing welfare (1) subject to the constraints (2) and (3) is

$$\sum_{i=1}^{n} W_i(g_i,\ v_i) - \lambda \left[\sum_{i=1}^{n} y_i(g_i,\ v_i) - G \right] - \mu \left[\sum_{i=1}^{n} z_i(g_i,\ v_i) - V \right], \quad (4)$$

where λ and μ are the multipliers for the constraints on the use of guilt

and virtue. The first-order condition if the optimal level of guilt g_i^* on subset S_i is positive is[14]

$$p_i\left[\int_0^{\infty} (h + \lambda g_i - \mu v_i) f_i(g_i + v_i, h)dh - (1 + \lambda)[1 - F_i(g_i + v_i)]\right] = \alpha_i'(g_i), \tag{5}$$

and the first-order condition if the optimal level of v_i^* is positive is

$$p_i\left[\int_0^{\infty} (h + \lambda g_i - \mu v_i) f_i(g_i + v_i,\ h)dh + (1 - \mu)F_i(g_i + v_i)\right] = \beta_i'(v_i). \tag{6}$$

The two terms in brackets on the left sides of (5) and (6) correspond to marginal and inframarginal effects of raising g_i and v_i. The first (integral) terms reflect the marginal net benefit of deterring additional acts. When g_i or v_i is raised slightly, the marginal acts that are deterred are those for which $u = g_i + v_i$; hence, with regard to the utility experienced by an individual with an act just at the margin, deterrence has no effect on social welfare. However, when an act is deterred, the external harm h is also avoided. Moreover, when an act is deterred, the fact that the individual no longer experiences g_i relaxes the constraint on the use of guilt by that amount, which has an implicit value per unit of λ. But the individual now experiences v_i, which tightens the constraint on the use of virtue by that amount, which has an implicit cost per unit of μ. Each of these marginal benefits and costs is weighted by $f_i(g_i + v_i, h)$, the density of acts deterred at the margin.

The second terms in brackets in (5) and (6) are the inframarginal effects on welfare of raising g_i and v_i, respectively. For those acts that are not deterred, whose relative proportion in the subset S_i is $1 - F_i(g_i + v_i)$, there are two costs of raising g_i: Individuals suffer an additional unit of guilt, and an additional unit of the constrained pool of guilt is used. Likewise, for those acts that are deterred, whose relative proportion in the subset S_i is $F_i(g_i + v_i)$, raising v_i has two effects: Individuals experience an additional unit of virtue and an additional unit of the constrained pool of virtue is used.

These two types of effects, the marginal (or deterrence) effects and the inframarginal effects, are equated with the direct marginal cost of instilling a higher level of guilt or of virtue, as the case may be.

[14] As will be apparent from the discussion to follow, $g_i^* = 0$ and $v_i^* = 0$ are each possible. In addition, the first-order conditions are not sufficient conditions for a global optimum.

A. *Basic Results*

We now state characteristics of the optimum; these are proved in the Appendix.

PROPOSITION. For each subset S_i, the following conditions hold:

a. Neither positive guilt g_i^* nor positive virtue v_i^* is instilled unless there exists a subset of situations in S_i having positive probability for which individuals otherwise would act $(u > 0)$ but for which acting is not first-best $(u < h)$—provided that, at the optimum, $\beta_i'(0) > (1 - \mu)p_i$.

b. Both possible types of deviations from first-best behavior may occur: the commission of undesirable acts and the deterrence of desirable acts.

c. If $g_i^* > 0$, guilt may sometimes be experienced; if $v_i^* > 0$, virtue may not always be experienced.

Part *a* states the chief point motivating this article, that the moral sentiments are instilled in order to control socially undesirable behavior. That is, guilt or virtue will be used for a subset S_i only if there exist acts in S_i that it would be desirable to deter. If we consider guilt alone, the reason for this conclusion is clear: Because guilt is costly to inculcate and involves costs if it is ever experienced, there must be some behavioral benefit—some acts that it is desirable to deter—for guilt to be socially desirable to inculcate. If we consider virtue as well, however, the argument for the conclusion must take into account that the experiencing of virtue is itself a source of utility. This point raises the possibility that it would be optimal to instill virtue solely because of the benefit of its being experienced. Our condition that $\beta_i'(0) > (1 - \mu)p_i$ is sufficient to rule out this possibility. Either the marginal inculcation cost can be sufficiently high ($\beta_i'(0) > p_i$ will suffice) or the constraint (3) on the use of virtue can be sufficiently binding $(\mu > 1)$, or some less demanding combination of the two. (Regarding the constraint, observe that if virtue is sufficiently scarce, the only question is *where* to use virtue rather than *how much* total virtue to use, and it will tend to be optimal to use virtue where the benefits of controlling behavior are the greatest relative to how much virtue must be used; see subsection *C* below.)

With regard to part *b*, undesirable acts may be committed because g_i^* and v_i^* are not sufficient to deter them, which may arise because inculcation costs are high or because guilt and virtue are scarce. The possibility that desirable acts may be deterred is a consequence of the grouping of acts, that is, the need for guilt and virtue to be uniform within each subset S_i of situations. Although most acts in a subset S_i may be undesirable, making it optimal to use guilt or virtue to deter them, there may be some acts in the subset with atypically low h (for instance, $h = 0$).

Part *c* states that, even when it is optimal to use guilt or virtue or both, they may not always succeed in controlling behavior. This conclusion is a further consequence of the grouping of acts into subsets S_i. For example, most acts in a subset may be quite harmful and produce little conventional utility, making it desirable to deter them with some combination of guilt and virtue. Nevertheless, a few acts in the subset may yield very high conventional utility, in which case these acts may well be committed and thus guilt rather than virtue would then be experienced—unless a very high combined level of guilt and virtue is employed, but that may be too costly.

B. *Comparison of the Optimal Levels of Guilt and Virtue to the Optimal Level of a Pigouvian Tax*

It is of interest to compare the optimal use of guilt and virtue to control externalities, as described above, with the optimal Pigouvian tax. It may be best for the moral incentive—the sum of guilt and virtue, $g_i^* + v_i^*$—to be either below or above the Pigouvian tax benchmark, namely, $E(h \mid S_i)$, the expected level of the externality on S_i.[15] The case in which the optimal moral incentive is below expected harm may arise because of the cost of instilling guilt and virtue and because of the constraints on their use (and, regarding guilt, also because of the utility cost of its being experienced by individuals). An implication is that $g_i^* = 0$ or $v_i^* = 0$ (or both) is possible even when there exist acts in S_i that ideally should be deterred. The case in which the optimal moral incentive exceeds expected harm may occur because raising deterrence may reduce the application of moral sanctions, which as noted is costly.[16]

[15] Our exposition involves an oversimplification in the case in which harm is not independent of the utility of the externality-causing activity; the reader may interpret our remarks for the case of independence or add the appropriate adjustments to our exposition.

[16] More precisely, consider a subset in which only guilt is used: Raising g_i will reduce the aggregate amount of guilt that is experienced when the deterrent effect exceeds the inframarginal effect, i.e., when $g_i f_i(g_i) > 1 - F_i(g_i)$. To illustrate the claim in the text, suppose further that the only situations for which $u > h$ are those in which u is just slightly above h. Then it may be optimal to deter all such acts by setting g_i equal to the highest level of u. The deterred acts involve a direct social loss of $u - h$, which is assumed to be small; suppose further that the marginal cost of raising g_i is not large. These social costs may be less than the benefits arising from individuals not experiencing g_i. (Observe that if g_i were set equal to the expected harm, then raising g_i slightly would involve a loss in conventional utility equal to the expected harm—just as in the case of a Pigouvian tax—but a savings of the expected harm plus g_i, which at that point itself equals the expected harm, plus a possible further benefit from relaxing the guilt constraint.)

C. *Choice between the Use of Guilt and Virtue as Incentives*

Comparison of the first-order conditions (5) and (6) for the optimal use of guilt and of virtue sheds light on whether it is optimal to rely primarily (or exclusively) on guilt or primarily on virtue to control behavior in a subset S_i. The marginal net benefits of using guilt and virtue—the first terms on the left sides of (5) and (6)—are identical, reflecting the fact that they are substitutes as deterrents. The marginal inculcation costs—the right sides of (5) and (6)—are symmetric, so this consideration favors using whichever sentiment, guilt or virtue, has the lower marginal inculcation cost. So far, therefore, there is no qualitative difference between the desirability of guilt and of virtue as incentives.

However, consideration of the inframarginal effects of using guilt and virtue—the second terms on the left sides of (5) and (6)—suggests an important distinction. In what we take as our benchmark, the case in which $\mu > 1$ at the optimum, both second terms are negative, indicating that greater experiencing of both guilt and virtue is costly. But in general the amounts of guilt and virtue experienced are not the same: For guilt, the fraction experienced is $1 - F_i(g_i + v_i)$, and for virtue, the fraction experienced is $F_i(g_i + v_i)$. Thus, when most individuals will be deterred from committing acts in S_i, so that F_i is large, very little guilt will actually be experienced, whereas a significant amount of virtue will be experienced (each per unit inculcated). Accordingly, when most acts in S_i will be deterred, it will tend to be optimal to use guilt and not virtue. Likewise, when few acts in S_i will be deterred, so that F_i is small, it will tend to be optimal to use virtue and not guilt. Moreover, because the effect of raising g_i or v_i on inframarginal costs may be large even when, initially, $g_i = 0$ or $v_i = 0$, it may well be optimal to rely exclusively on guilt in the former case and exclusively on virtue in the latter case.

IV. Discussion

A. *Grouping of Situations*

We assumed that the universe of situations S is exogenously divided into distinct subsets S_i. An important reason to view acts as falling into groups rather than being considered individually is that the basic nature of perception and cognition is thought to be categorical (see, e.g., Kosslyn and Koenig 1992; Pinker 1997). (It is therefore not surprising that economists have begun to study categorical thinking in various settings; see, e.g., Mullainathan 2002; Fryer and Jackson 2007.)

Another reason favoring the assumption that the moral sentiments are inculcated over groups rather than act by act is that the former is advantageous in light of a number of concerns about the application of moral rules. More act-specific rules require more information to

apply, the information may not always be available, and even when present, it is costly to process. Further, the proper functioning of the moral emotions requires that their application be largely automatic. If, instead, whether one ultimately feels guilty depends on a conscious, complex assessment of highly context-specific information, the ability to rationalize in one's self-interest would often lead individuals not to feel guilty when they should, that is, when it would be socially desirable for them to refrain from actions that advance narrow self-interest. This phenomenon would undermine the function of guilt in regulating behavior that harms others.

In addition to these benefits of grouping with regard to the application of moral rules, the inculcation process itself may be more efficient and effective when acts are grouped into categories. There may be scale economies and related savings of time and effort in teaching broader, if somewhat imprecise, lessons than in attempting to treat every conceivable future situation as distinctive. Moreover, the fact that moral rules are inculcated to a significant degree during childhood favors the use of categories.

It does not follow, however, from the foregoing that the groups of situations S_i that are treated as morally equivalent are entirely fixed. Hence, a natural extension of our model would allow for choice as to the breadth of categories over which to inculcate guilt and virtue or for the inculcation of exceptions to general moral rules. Obviously, the greater the heterogeneity of acts within a group S_i, the greater would be the benefits of a finer partition of situations. Nevertheless, given the many factors that favor grouping, it seems clear that allowing the groupings to be endogenous, optimizing on all relevant margins in light of residual human limitations, would not alter our basic conclusions. Specifically, some grouping would be optimal, and within the optimal categories, all our results would hold.

We also assumed that the subsets over which guilt and virtue are inculcated are distinct, but in reality they may overlap. An interesting case of overlap occurs because moral rules that have a particular focus, such as those directed against lying or stealing, are accompanied by a more general moral rule, notably the Golden Rule, which broadly enjoins individuals to take into account the effects of their behavior on others. One can understand the Golden Rule as associating guilt with all undesirable acts and virtue with all desirable acts, perhaps with the levels of guilt and virtue rising with the extent of negative and positive externalities. This consideration raises the question why society does not simply inculcate the Golden Rule, eschewing all other rules, and thereby attempt to induce all individuals always to act in a socially optimal manner. The answer may lie in the reasons given for grouping in the previous paragraphs, including problems of complex calculations,

susceptibility to self-interested rationalization, and the difficulty of instructing young children. Additionally, recent research, such as that summarized by Cosmides and Tooby (1994), suggests that the human mind may perform better in specialized rather than in general problem solving. Moreover, our analysis suggests that, even if successful, such an approach would be problematic if the associated levels of guilt or virtue were high because of the constraints on the ability to experience the moral emotions. For example, many individuals would still commit undesirable acts, which would quickly consume the scarce pool of guilt, making it difficult to deter those acts that are particularly important to control.

B. Internal versus External Sanctions and Rewards

Corresponding to the internal mechanisms of guilt and virtue, there are external sanctions and rewards, namely, disapprobation or blame and approbation or praise.[17] These external analogues to the moral sentiments complement guilt and virtue in regulating individuals' behavior. Despite the similarities between these internal and external sanctions (taken in what follows to include rewards), a more complete analysis would also account for their differences.

External sanctions involve the actions of third parties, sometimes victims (or, in the case of helpful acts, beneficiaries) but often unrelated individuals. This requirement implies that there are three prerequisites for external sanctions to be effective: The individuals imposing the sanctions need information about the actor's behavior, they must be motivated to mete out the sanctions, and the actor must care about others' expressions of blame and praise. The third element is closely related to the internal sanctions and rewards of guilt and virtue: It seems that those who would feel guilty committing an act would usually feel badly if others express disapproval, and conversely. The second element, individuals' motivation to impose sanctions on actors, cannot be taken for granted. One explanation for individuals' motivation in this regard is that the very process by which, for example, guilt may be inculcated for committing a particular type of act would lead an individual to express disapproval of others' commission of the same type of act.[18] The first element, third parties' information, will sometimes arise automat-

[17] Prior work by economists on social sanctions for failure to adhere to social norms includes Akerlof (1980) and Bernheim (1994). Smith (1790) devoted significant attention to the similarities and differences between internal and external moral sanctions and rewards.

[18] For example, Fehr and Gächter (2002) present evidence that negative emotions triggered by antisocial behavior motivate punishment of third parties; this reaction is anticipated, thereby inducing cooperation. Cohen (2005) discusses evidence that emotions explain the rejection of low offers in the ultimatum game.

ically but often will involve associated activity, such as gossip, which itself requires motivation. One could model disapprobation and approbation explicitly using our framework by defining such behavior as involving additional subsets of acts, which themselves might have moral sentiments associated with them.

C. *Heterogeneity of Actors*

Our model presumed that individuals were identical, but it could easily be modified to accommodate a heterogeneous population of individuals. If individuals' utilities from acts or the external effects of their acts differ, the acts of different individuals can be labeled as distinct acts. In this case, different distributions of the likelihood of situations would be associated with different individuals. These distributions could then be aggregated across the population, and our social welfare maximization problem would refer to the average expected utility of individuals rather than to the expected utility of a single, representative individual. A complication is that the constraints on the experiencing of guilt and virtue would then apply separately to each individual.

Another important source of heterogeneity is that different individuals may be differentially susceptible to feelings of guilt and virtue. This could be due to differences in their constitution or differences in their upbringing. Izard (1991) indicates genetic differences in individuals' susceptibility to emotions. With regard to inculcation, since much of it is done by parents or local institutions, the potential for variation is substantial. Relatedly, as Akerlof and Kranton (2000) emphasize, individuals to an extent choose among different personal identities; these different identities, in turn, will tend to be associated with different norms and different internal and external sanctions. To model this heterogeneity, one could allow for a distribution of types with regard to individuals' personal sensitivity to guilt and virtue or to the degree to which inculcation succeeds. This, too, would not greatly alter the nature of our conclusions. The primary effect of actor heterogeneity on our analysis would be to augment the impact of the grouping of situations that themselves are heterogeneous.

D. *Prudence*

Although our analysis suggests that moral sanctions should be employed only when externalities are present, discussions of virtue and vice over the ages have often included categories of acts that seem to involve only self-regarding behavior. Moreover, psychologists have found that individuals experience guilt when they act in ways that harm only themselves (see Izard 1991). For example, individuals are urged to save for a rainy

day, not to overeat, and otherwise to protect themselves from their own folly, and individuals who fail to do these things may feel guilty. The primary explanation for such use of moral sanctions appears to be individuals' potential lack of self-control. In particular, many instances in which guilt and virtue seem to be associated with self-regarding behavior involve problems of myopia (hyperbolic discounting). These problems can be conceived as involving two selves, a present self whose decisions negatively affect a future self—as Thaler and Shefrin (1981), Schelling (1984), and many others have suggested. Under such a formulation, the behavior of the present self creates an externality, affecting a different self, and hence our analysis suggesting the potential benefits of employing guilt and virtue is relevant. An interesting application that involves self-control problems, the use of external sanctions, and Akerlof and Kranton's (2000) view that individuals to an extent choose their own identities, is the decision of some individuals to join groups such as Alcoholics Anonymous in order to raise the moral sanctions on their own self-destructive behavior.

E. Evolution and Inculcation

The general capacity to experience guilt and virtue—as distinct from how that capacity may be employed in a given society—must have an evolutionary origin, just as does any other capacity we might have (see, e.g., Darwin 1874; E. O. Wilson 1975; Izard 1991). The ability of individuals to learn associations between actions and emotions and subsequently to have their behavior influenced by their emotions is well supported by recent work in the social and natural sciences.[19] That human nature is indeed programmable in this sense is further implied by the substantial efforts devoted to inculcating guilt and virtue to enforce various moral rules in the rearing of children, in organized religion, in educational institutions, and in some acts of government. The possibility of inculcation, moreover, is important in attempting to explain cross-cultural variation in moral rules as well as their rate of change over time, which seems greatly to exceed the rate of biological evolution (see, e.g., Izard 1991; Nisbett and Cohen 1996).

It does not follow, of course, that the moral sentiments will in fact be employed in a manner that maximizes social welfare. First, biological

[19] See many of the references cited in note 2. For example, Massey (2002) surveys literature in evolutionary psychology and neurology in explaining how implicit memories are created by the pairing of external stimuli with hard-wired emotions so that, when the relevant stimuli are subsequently experienced, emotions are triggered; these responses occur even before the mind is able to begin rational analysis of the situation. Moreover, these links between stimuli and emotions influence cognition and are highly durable and thus difficult to eliminate.

evolution tends to maximize survival (more precisely, replication of the pertinent genes) rather than welfare, although it may well be the case that in societies not on the brink of subsistence, inculcation will reflect an interest in maximizing welfare. Second, it may not be society's overall welfare that is maximized by morality. Parents may be concerned primarily with their family's well-being, religious or other organizations with the welfare of their members, and so forth. Nevertheless, there often is overlap: Teaching simple rules (do not lie or steal, do not free ride) to children to regulate behavior within the family may spill over to future interactions with others, and parents may teach more altruistic rules because individuals known to be honest and cooperative may face better opportunities in society at large. Third, maximization is not assured. With evolution, there is the familiar point that selection is fundamentally at the level of individual genes, so traits that would benefit a group as a whole may not emerge (although they may arise to some extent through kin selection and reciprocity). With inculcation, there is the problem that inculcators do not bear all the costs and benefits of their actions. For example, when there are multiple inculcators (family, religious institutions, and government), each may impose externalities on others through excessive use of the scarce capacity to experience guilt and virtue, a sort of common-pool problem.

V. Conclusion

Our analysis offers a theory of how moral sanctions and rewards—feelings of guilt and virtue—would best be employed if the purpose were to maximize social welfare. Although beyond the scope of the present inquiry, it is natural to ask whether common morality is roughly consistent with our results. One would not predict a close fit because, as we discuss in Section IV.*E*, the processes that produce common morality hardly guarantee optimality and may not involve the maximization of social welfare in particular. Furthermore, our model considers morality in isolation from other social methods of controlling behavior, whereas the optimal use of morality will depend on the availability of other instruments, notably, the legal system, regulation, and taxes and subsidies (see Shavell 2002). Nevertheless, since our model and results are basic, one might expect them to possess a measure of explanatory power.

Most obviously, it is indeed the case that behavior censured as immoral (lying, stealing) tends to be socially undesirable because of negative externalities, and behavior deemed morally praiseworthy (rescue) tends to be socially desirable because of positive externalities. In addition, many moral rules (such as those concerning lies, promises, and aggression) apply to groups of related acts; even when there are exceptions or other refinements, moral rules apply to categories rather than there

existing separate rules finely tailored to the particulars of each possible situation. Moreover, it has long been recognized that certain acts (such as lies in specific circumstances) might be considered immoral despite their being socially desirable. Hence, individuals may sometimes be deterred by the prospect of moral sanctions from committing desirable acts, and others may sometimes feel guilty for committing acts that are in fact socially desirable. Additionally, regarding our result that optimal moral sanctions may be below or above the Pigouvian tax benchmark, it does seem that moral sanctions seem lenient for some types of acts in some moral systems and excessive for other types.

Furthermore, the observed choice in moral systems whether to rely primarily on guilt or on virtue seems to be in accord with our model, which indicates that limits on individuals' capacities to experience guilt and virtue make it optimal to employ whichever one would least need to be actually experienced. Thus, individuals who fail to comply with rules that are usually followed (such as norms against cutting in line or unprovoked aggression) tend to feel guilty, rather than the majority who routinely comply continually feeling virtuous. Likewise, individuals who engage in unusual acts of sacrifice to help others do seem to feel virtuous and are subject to praise, rather than most people who, say, fail to devote the majority of their wealth to help those less fortunate always feeling guilty and being subject to pervasive disapprobation.

We also note that, in principle, our model could be employed to help understand the frequently noted differences in moral systems across cultures and over time. Because of differences both in the relative importance of various external harms and benefits and in the role and effectiveness of systems of inculcating morality (religious institutions versus government versus families), one would not expect systems of common morality to be the same. Although measuring the relevant parameters would be quite difficult, it may still be possible to illuminate recognized variations by using a framework like ours that relates moral systems to underlying social conditions.

One could also attempt to employ a model like ours to analyze how certain government policies (e.g., laws having symbolic effects, such as civil rights legislation) may be used to reinforce or modify common morality. Our analysis suggests the importance of such inquiries but also raises some cautions. Notably, increasing the level of guilt and social disapprobation associated with socially undesirable behavior may be beneficial by deterring it, but it also may be costly to the extent that it is not fully successful: Experiencing guilt involves a direct cost and, given the scarcity of individuals' capacities to experience the moral sentiments, may reduce their effectiveness in controlling other behavior. Furthermore, when families, religious and educational institutions, and the

government all compete in attempting to use this scarce common resource for their own ends, the results are unlikely to be optimal.

Appendix

Proof of the Proposition

Part a.—We first observe that, from expression (4), we know that g_i^* and v_i^* must maximize $W_i(g_i, v_i) - \lambda y_i(g_i, v_i) - \mu z_i(g_i, v_i)$. Thus neither $g_i > 0$ nor $v_i > 0$ can be optimal if the following expression is positive for all postulated optimal nonzero pairs $(g_i, v_i) \geq 0$:[20]

$$[W_i(0, 0) - \lambda y_i(0, 0) - \mu z_i(0, 0)] - [W_i(g_i, v_i) - \lambda y_i(g_i, v_i) - \mu z_i(g_i, v_i)]$$

$$= p_i \int_0^{\infty} \int_0^{g_i + v_i} (u - h) f_i(u, h) du dh + [\alpha_i(g_i) - \alpha_i(0)] + [\beta_i(v_i) - \beta_i(0)]$$

$$+ p_i(1 + \lambda) g_i [1 - F_i(g_i + v_i)] - p_i(1 - \mu) v_i F_i(g_i + v_i). \quad \text{(A1)}$$

Now suppose, contrary to the assumption of part *a* of the proposition, that there does not exist a subset of situations in S_i having positive probability for which individuals otherwise would act ($u > 0$) but acting is not first-best ($u < h$). That is, suppose that $u \leq 0$ or $u \geq h$ for all (u, h) in S_i (except possibly on a set of measure zero). In this case, it can be shown that (A1) is positive for any postulated optimal nonzero pair (g_i, v_i). The first four terms are nonnegative. The final term has an ambiguous sign. If $v_i = 0$, this term equals zero; moreover, in that event it must be that $g_i > 0$ because we are considering only nonzero pairs (g_i, v_i), and this in turn implies that the second term is positive, so (A1) must be positive. Suppose instead that $v_i > 0$. In this case, the ambiguous final term combined with the fourth term, $\beta_i(v_i) - \beta_i(0)$, can together be shown to be positive. Given that $\beta_i'(0) > (1 - \mu)p_i$ at any optimum, we have $\beta_i'(0)v_i > p_i(1 - \mu)v_i \geq p_i(1 - \mu)v_i F_i(g_i + v_i)$. Moreover, $\beta_i(v_i) - \beta_i(0) \geq \beta_i'(0)v_i$ because $\beta_i'(0) > 0$ and $\beta_i''(0) \geq 0$. Therefore, $\beta_i(v_i) - \beta_i(0) > p_i(1 - \mu)v_i F_i(g_i + v_i)$, so the two terms combined are positive, and (A1) must be positive in this case as well.

Parts b and c.—To prove these parts, it suffices to construct an example for each claim. For all the claims except the latter claim of part *c*, that virtue may not always be experienced, we consider an example in which $V = 0$, so that virtue cannot be used. Furthermore, we choose an example in which the constraint on guilt (2) is not binding. To ensure this, suppose that u never exceeds 1, so that g_i^* cannot exceed 1. (As $g_i = 1$ is sufficient to deter any act, no higher g_i can be optimal on any subset S_i because the only effect of raising g_i above 1 would be to increase inculcation costs.) Now, assume that $G > 1$, so that constraint (2) cannot be binding and thus $\lambda = 0$. For the remainder of the example, we confine attention to a particular subset S_i. Assume that the distributions of u and of h on S_i are independent, so that $f_i(g_i + v_i, h) = f_{i1}(g_i + v_i) f_{i2}(h)$. For u, assume a triangular distribution on $[-1, 1]$ such that $f_{i1}(-1) = f_{i1}(1) = 0$ and $f_{i1}(0) = 1$. For h, assume a distribution that is positive on $(0, 2)$ and has a mean of 1. Let $p_i = 0.1$, and let α' be constant and equal to 0.0375. Now, using the first-order condition (5) for $g_i^* > 0$ and moving p_i to the denominator on the

[20] When writing (A1) we find it convenient, with respect to using expression (1) for W_i, to state g_i and v_i separately, taking advantage of the fact that g_i and v_i are constants when integrating with respect to u and h.

right side, we have $1(1 - g_i) - (1 + 0)[1 - (0.5 + g_i - 0.5g_i^2)] = 0.0375/0.1$. The solution to this is $g_i^* = 0.5$.[21] Part *c*, that guilt may sometimes be experienced, is true because, whenever $u > 0.5$, the act is committed and guilt is therefore experienced. Part *b* is also true. That undesirable acts may be committed follows because, as just noted, all acts for which $u > 0.5$ are committed, but for any such u, some situations will be such that $h > u$ because the distribution of h is positive (and independent of u) on (0, 2). That desirable acts may be deterred follows because all acts for which $u \leq 0.5$ are deterred, but for all acts such that $u > 0$, there will be situations in which $h < u$.

Finally, to show that it is possible that $v_i^* > 0$ but virtue may not be experienced when situations in S_i arise, we can construct a different type of example. Suppose that there is only one subset (with probability 1). Suppose further that $G = 0$, so that guilt cannot be used. In addition, assume that h is distributed independently of u and has a mean of 1; the cumulative distribution of u on this single subset $F(0.1) = 0.99$, $F(1) < 1$; $\beta(0.1) < 0.2$, $\beta(1) > 1$; and $V > 1$.[22] First, observe that $v^* > 0$. This is necessarily true because welfare at $v = 0.1$ exceeds welfare at $v = 0$ (and $v = 0.1$ is feasible since $V > 1$): Raising v from 0 to 0.1 involves an inculcation cost less than 0.2, deters 0.99 of the acts and thus causes a total loss in conventional utility of less than 0.1 (since each deterred act is such that $u \leq 0.1$) and avoids harm of 0.99 (since the mean of h is 1). Second, observe that $v^* < 1$. The reason is that the inculcation cost at $v = 1$ exceeds 1, which in turn exceeds the maximum possible benefit from avoiding harm, which equals 1, so total welfare at $v = 1$ is less than that at $v = 0$. Finally, this implies that, even though $v^* > 0$, virtue will not always be experienced, for there are situations in which $u > 1$, where the act is committed (because $v < 1$ and $g = 0$), and thus virtue is not experienced.

References

Akerlof, George A. 1980. "A Theory of Social Custom, of Which Unemployment May Be One Consequence." *Q.J.E.* 94 (June): 749–75.

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *Q.J.E.* 115 (August): 715–53.

Alexander, Richard D. 1987. *The Biology of Moral Systems*. New York: Aldine de Gruyter.

Barkow, Jerome H., Leda Cosmides, and John Tooby, eds. 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford Univ. Press.

Baron, Jonathan. 1994. *Thinking and Deciding*. 2nd ed. Cambridge: Cambridge Univ. Press.

Becker, Gary S., and Kevin M. Murphy. 2000. *Social Economics: Market Behavior in a Social Environment*. Cambridge, MA: Harvard Univ. Press.

Ben-Ner, Avner, and Louis Putterman, eds. 1998. *Economics, Values, and Organization*. Cambridge: Cambridge Univ. Press.

[21] This must be a maximum because 0.5 is the only solution to (5) and the derivative of (4) with respect to g_i is positive at $g_i = 0$: it is $0.1(1 - 0.5) - 0.0375 = 0.0125$.

[22] We observe that these assumptions are consistent with the assumption in part *a* that $\beta_i'(0) > (1 - \mu)p_i$: As will be seen, $v^* < 1$, so the constraint is not binding, which implies that $\mu = 0$. Thus, the right side equals p_i, which here equals 1. Finally, the assumption that $\beta'(0) > 1$ is consistent with the assumption in this example that $\beta(0.1) < 0.2$.

Bernheim, B. Douglas. 1994. "A Theory of Conformity." *J.P.E.* 102 (October): 841–77.

Binmore, Ken. 1998. *Game Theory and the Social Contract.* Vol. 2. *Just Playing.* Cambridge, MA: MIT Press.

Campbell, Donald T. 1975. "On the Conflicts between Biological and Social Evolution and between Psychology and Moral Tradition." *American Psychologist* 30 (December): 1103–26.

Cohen, Jonathan D. 2005. "The Vulcanization of the Human Brain: A Neural Perspective on Interactions between Cognition and Emotion." *J. Econ. Perspectives* 19 (Fall): 3–24.

Cosmides, Leda, and John Tooby. 1994. "Better than Rational: Evolutionary Psychology and the Invisible Hand." *A.E.R. Papers and Proc.* 84 (May): 327–32.

Daly, Martin, and Margo Wilson. 1988. *Homicide.* New York: Aldine de Gruyter.

Damasio, Antonio R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Putnam.

Darwin, Charles. [1872] 1998. *The Expression of the Emotions in Man and Animals.* Edited by Paul Ekman. 3rd ed. Oxford: Oxford Univ. Press.

———. [1874] 1998. *The Descent of Man; and Selection in Relation to Sex.* 2nd ed. Amherst, NY: Prometheus Books.

Elster, Jon. 1998. "Emotions and Economic Theory." *J. Econ. Literature* 36 (March): 47–74.

Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415 (January 10): 137–40.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Q.J.E.* 114 (August): 817–68.

Frank, Robert H. 1988. *Passions within Reason: The Strategic Role of the Emotions.* New York: Norton.

Frederick, Shane, and George Loewenstein. 1999. "Hedonic Adaptation." In *Well-Being: The Foundations of Hedonic Psychology*, edited by Daniel Kahneman, Ed Diener, and Norbert Schwarz. New York: Sage Found.

Fryer, Roland G., Jr., and Matthew O. Jackson. 2007. "A Categorical Model of Cognition and Biased Decision-Making." *Contributions in Theoretical Econ.*, forthcoming.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Rev.* 108 (October): 814–34.

Hauser, Marc D. 2006. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong.* New York: HarperCollins.

Hirshleifer, Jack. 1987. "On the Emotions as Guarantors of Threats and Promises." In *The Latest on the Best: Essays on Evolution and Optimality*, edited by John Dupré. Cambridge, MA: MIT Press.

Hume, David. [1751] 1998. *An Enquiry Concerning the Principles of Morals.* Edited by Tom L. Beauchamp. Oxford: Oxford Univ. Press.

Izard, Carroll E. 1991. *The Psychology of Emotions.* New York: Plenum.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1987. "Fairness and the Assumptions of Economics." In *Rational Choice: The Contrast between Economics and Psychology*, edited by Robin M. Hogarth and Melvin W. Reder. Chicago: Univ. Chicago Press.

Kaplow, Louis, and Steven Shavell. 2001. "Moral Rules and the Moral Sentiments: Toward a Theory of an Optimal Moral System." Working Paper no. 8688 (December), NBER, Cambridge, MA.

———. 2002. *Fairness versus Welfare*. Cambridge, MA: Harvard Univ. Press.
Kosslyn, Stephen M., and Olivier Koenig. 1992. *Wet Mind: The New Cognitive Neuroscience*. New York: Free Press.
LeDoux, Joseph E. 1996. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon & Schuster.
Massey, Douglas S. 2002. "A Brief History of Human Society: The Origin and Role of Emotion in Social Life." *American Sociological Rev.* 67 (February): 1–29.
Mullainathan, Sendhil. 2002. "Thinking through Categories." Preliminary draft, Massachusetts Inst. Tech.
Nisbett, Richard E., and Dov Cohen. 1996. *Culture of Honor: The Psychology of Violence in the South*. Boulder, CO: Westview.
Pinker, Steven. 1997. *How the Mind Works*. New York: Norton.
Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *A.E.R.* 83 (December): 1281–1302.
Rayo, Luis, and Gary S. Becker. 2007. "Evolutionary Efficiency and Happiness." *J.P.E.* 115 (April): 302–37.
Robson, Arthur J. 2001. "The Biological Basis of Economic Behavior." *J. Econ. Literature* 39 (March): 11–33.
Schelling, Thomas C. 1984. *Choice and Consequence*. Cambridge, MA: Harvard Univ. Press.
Shavell, Steven. 2002. "Law versus Morality as Regulators of Conduct." *American Law and Econ. Rev.* 4 (Fall): 227–57.
Sidgwick, Henry. [1907] 1981. *The Methods of Ethics*. 7th ed. Indianapolis: Hackett.
Smith, Adam. [1790] 1976. *The Theory of Moral Sentiments*. 6th ed. Oxford: Oxford Univ. Press.
Thaler, Richard H., and H. M. Shefrin. 1981. "An Economic Theory of Self-Control." *J.P.E.* 89 (April): 392–406.
Trivers, Robert L. 1971. "The Evolution of Reciprocal Altruism." *Q. Rev. Biology* 46 (March): 35–57.
Wilson, Edward O. 1975. *Sociobiology: The New Synthesis*. Cambridge, MA: Harvard Univ. Press.
Wilson, James Q. 1993. *The Moral Sense*. New York: Simon & Schuster.