LE 8/3/02
11:30AM
KAPLOW/SHAVELL

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

Louis Kaplow and Steven Shavell*

Abstract

How should moral sanctions and 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 an incentive because they are negative and positive sources of utility, when experienced. We assume that guilt and virtue are costly to inculcate and are subject to certain constraints on their use. We characterize the optimal use of guilt and virtue, including the choice between them, discuss a number of extensions of our model, and relate our results to the systems of common morality that we observe.

---

*Harvard University and National Bureau of Economic Research. We thank Gary Becker, Robert Ellickson, Oliver Hart, A. Mitchell Polinsky, Eric Posner, Eric Rasmusen, Richard Zeckhauser, and participants in workshops at the University of California at Berkeley, University of Chicago, Harvard University, and Stanford University for comments.

1. Introduction

In economic analysis of individual behavior, it generally is assumed that individuals are motivated only by the direct contribution to utility that would be produced by their actions. Yet it is obvious that moral sentiments — feelings of guilt and of virtue (along with their external correlates, disapprobation and praise) — are also springs of human action. That is, individuals are to some degree motivated by the prospect of feeling guilty or of feeling virtuous to follow moral rules and thus will sometimes act against their self-interest, conventionally interpreted. For example, an individual may refrain from telling a lie that would otherwise benefit him because telling the lie would make him feel guilty.

Whether and to what extent we experience guilt and virtue when we commit certain acts, such as telling lies or breaking promises, is not arbitrary. Rather, society's system of morality is the product of a complex process of socialization, especially in childhood, and, as we will explain, of evolution. Moreover, it seems plausible that these mechanisms have some tendency, however imperfect, to promote overall well-being.

Against this background, we ask what system of morality — in particular, what use of guilt and of virtue to induce individuals to follow moral rules — leads to the maximization of social welfare. In so doing, we build on such writers as Hume (1739, 1751) and Sidgwick (1907), who argued informally that the observed system of morality tends to advance welfare.[1]

Recent economic literature on social norms and some writing in behavioral economics, such as that exploring behavior motivated by concerns for "fairness," recognizes that individuals' behavior is not always narrowly self-interested and may reflect moral concerns.[2] (Other scholars in a variety of fields have also addressed the role of moral emotions in regulating behavior.[3]) Work by economists has tended to focus on establishing the existence of certain apparently non-self-interested motivations (such as in the ultimatum game). In contrast, we take as given certain non-self-interested motivations and examine how they may optimally be employed to advance social welfare. Another strand of literature seeks to explain cooperative behavior as rational, often as an equilibrium of a repeated game; our analysis is complementary to this work because the moral sentiments can reinforce cooperation (for example, promises support cooperation if they are credible, and the prospect of guilt for breaking promises helps to make them credible). We are not aware, however, of prior writing that seeks formally to determine optimal moral rules and their enforcement with guilt and virtue.

---

[1]More recently, Brandt (1996) has discussed the features of an optimal moral system, although not in the explicit manner that we pursue here.

[2]Analysis of the motivating force of moral sentiments begins at least as early as Smith (1790). See also Becker (1996), Becker and Murphy (2000), Ben-Ner and Putterman (1998), Binmore (1998), Fehr and Schmidt (1999), Frank (1988), Hirshleifer (1987), Kahneman, Knetsch, and Thaler (1987), Rabin (1993), and Robson (2001).

[3]See Alexander (1987), Barkow, Cosmides, and Tooby (1992), Baron (1994), Campbell (1975), Daly and Wilson (1988), Damasio (1994), Darwin (1872, 1874), Elster (1998), Gibbard (1990), Haidt (2001), Izard (1991), LeDoux (1996), Mackie (1985), Massey (2002), Pinker (1997), Trivers (1971), E.O. Wilson (1975), and J.Q. Wilson (1993). Greene et al. (2001) present brain scan evidence showing that emotions play an important role in individuals' moral judgments.

We describe our framework for analysis in section 2. Individuals decide whether or not to commit various acts, each of which directly produces utility and may also result in an externality. Individuals are subject to a process of inculcation such that they will experience guilt or 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 and/or virtue is sufficiently great. The inculcation process is assumed to be costly. We further suppose that the ability to use guilt and virtue is constrained in two ways: First, individuals are assumed to have only a limited capacity to experience these moral sentiments. (Thus we suppose that it is impossible for individuals to feel extremely guilty all of the time.) Second, guilt and virtue cannot be independently specified for every conceivable situation but rather must be inculcated over groups of acts. (For example, a specific level of guilt may be inculcated for telling lies, or various types of lies, but guilt may not be adjusted to the exact circumstances of every particular type of lie.) The social problem 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 any externalities associated with these acts — combined with moral elements — the utility associated with the experiencing of feelings of guilt or virtue and the costs of inculcation.

We analyze the optimal system of morality in section 3. Not surprisingly, we find that the moral sentiments are used only when individuals' behavior otherwise would not be first-best, due to externalities.[4] (This result involves qualifications because the experiencing of virtue is itself a source of utility.) However, the moral sentiments can only imperfectly correct externalities, an important reason arising from the need to inculcate guilt and virtue at uniform levels over groups of acts that may exhibit heterogeneity. For example, most but not all acts in a group may be undesirable; in this case, the desirable acts in the group may be deterred by the moral sentiments, and individuals may feel guilty for committing some acts even though the acts are socially desirable. An additional result 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 sanctions would be set equal to the level of the externality involved. That is, there can exist groups of acts where moral leniency or moral harshness is optimal.

We also obtain results regarding the choice between the use of guilt and of virtue. Although these moral sentiments are in many respects substitutes for the regulation of behavior, they differ in two important ways. Most obviously, guilt reduces utility and thus social welfare when it is experienced, whereas virtue raises utility and social welfare. Furthermore, when guilt is successful in controlling behavior, it is *not* experienced (because acts punished by guilt are deterred); however, when virtue is successful, it *is* experienced (because acts rewarded by virtue are committed). Since guilt and virtue are assumed to be scarce, this latter difference implies that guilt is best to use when most violations of moral rules can successfully be deterred, while virtue

[4]We include positive externalities, such as when virtue is used to encourage helping others in distress or guilt is used to penalize free-riding in the provision of public goods.

is best to use when few individuals can be induced to follow a moral rule.[5] Put another way, it is optimal for worse-than-normal moral behavior to be punished and for unusually good moral behavior to be rewarded.

In section 4, we draw on pertinent literature to discuss the social, psychological, and evolutionary basis for our assumptions, including those concerning inculcation costs, constraints on the use of moral sentiments, and the need to inculcate guilt and virtue over groups of acts, and we also consider the implications of relaxing many of these assumptions. Then we briefly examine additional extensions and issues of interpretation: how the analysis of external sanctions and rewards (disapprobation and praise) would differ from that of internal ones (guilt and virtue); how our results would change by admitting heterogeneity among actors, particularly concerning the extent to which they experience guilt and virtue; how our model might be applied to moral rules that govern prudence, that is, behavior that does not seem to involve externalities; and how our analysis illuminates certain important questions in moral philosophy.

Finally, in section 5 we conclude by offering remarks concerning whether our results are consistent with observed features of existing moral systems and may help explain differences in moral systems across cultures and over time, and by noting some possible normative implications.

2. Framework

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.[6] 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 an individual commits the act, the individual obtains (positive or negative) utility u associated with the act per se, which we sometimes refer to as act-utility. In addition, an act causes an external harm of $h \geq 0$. (This assumption allows for positive as well as negative externalities through appropriate labeling of "acts."[7]) If the individual does not commit the act, he does not obtain any act-utility and does not cause any external harm.[8]

A situation in S may thus be identified with a pair (u,h) describing the act that an

---

[5]Compare Wittman's (1984) suggestion that one should choose between rewards and penalties based on which instrument economizes on administrative costs, determined by frequency of application.

[6]More generally, an individual who finds himself in a situation s may choose from a set of n(s) acts $a_1(s)$, $a_2(s),\ldots,a_{n(s)}(s)$, each of which is associated with a utility for the individual and an external effect. For our purposes, however, it is sufficient to assume that there are just two acts in each situation, one of which we will identify with some act and the other with not committing that act.

[7]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 in this case would not, however, affect the qualitative nature of our conclusions.)

[8]This assumption is only a normalization; the results we obtain depend only on the differences in utilities and in externalities associated with acting versus not acting.

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 guilt $g(u,h) \geq 0$ for committing an act in a particular situation (u,h). By this, we mean that a person in that situation (u,h) will experience guilt — that is, suffer disutility — of g(u,h) if and only if he commits the act. Similarly, assume that society may instill virtue $v(u,h) \geq 0$ for not committing an act in situation (u,h); that is, a person obtains utility of v(u,h) if and only if he does not commit the act.

The prospect of guilt or of virtue may lead an individual to change his behavior.[9] In the absence of guilt and virtue, an individual in a given situation will commit the act if and only if $u > 0$.[10] When guilt g(u,h) is instilled for acting and virtue v(u,h) for not acting, the person will act if and only if the overall utility from acting exceeds the utility from not acting, that is, if and only if $u - g(u,h) > v(u,h)$. It is also sometimes convenient to express this condition as $u > g(u,h) + v(u,h)$, which is to say that the utility from committing the act per se exceeds the sum of moral sanctions and rewards that favor not committing the act.

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 determined 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 an extremely nuanced manner. As we discuss in subsection 4.3, 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 results would not be affected.

Second, there is a cost of instilling guilt and of instilling 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 on each subset S_i at cost $\beta_i(v_i)$, where $\beta_i'(v_i) > 0$ and $\beta_i''(v_i) \geq 0$.

Third, there is a constraint on the actual experiencing of guilt and of virtue, namely, that the expected value of experienced guilt cannot exceed an amount $G \geq 0$ and the expected value of experienced virtue cannot exceed an amount $V \geq 0$. (As will be seen, this assumption has an important, qualitatively different effect than our assumption on inculcation costs.)

The motivation for these assumptions (as we note in the introduction and discuss in subsection 4.1) is that, as a matter of human psychology, our capacity actually to experience the

---

[9]It is not important for our analysis how individuals actually conceive of guilt and of virtue or whether moral considerations are in some psychological sense qualitatively different from ordinary sources of act-utility.

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

emotions of guilt or of virtue is bounded; there is a "crowding out" or dulling effect on further feelings of guilt or virtue as the frequency and magnitude of our experiencing these emotions increase.[11] Relaxing this assumption, to allow the increased use of guilt or virtue to decrease the marginal effectiveness of guilt or virtue, rather than using a simple capacity constraint, would complicate the exposition without materially affecting the results; see note 22.

Social welfare is taken to be the expected value of the utility that individuals experience from committing acts per se, plus any realized virtue and minus any realized guilt, minus externalities, and minus the costs of instilling guilt and virtue. As noted above, we will sometimes refer to social welfare as morally inclusive social welfare to distinguish it from conventional social welfare, which includes only act-utility and externalities.

Let us note, before proceeding, that the conventional first-best solution to the problem of social welfare maximization is for an act in a given situation to be committed if and only if $u > h$.

3. 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[12]

$$(1) \quad \sum_{i=1}^{n} W_i(g_i,v_i),$$

where

$$(2) \quad W_i(g_i,v_i) = p_i\left[\int_0^{\infty}\int_{g_i+v_i}^{\infty}(u-h-g_i)f_i(u,h)dudh + \int_0^{\infty}\int_0^{g_i+v_i} v_i f_i(u,h)dudh\right]$$

$$- \alpha_i(g_i) - \beta_i(v_i).$$

To explain, individuals commit acts in situations in subset S_i when $u - g_i > v_i$, or $u > g_i+v_i$, in which case the effect on social welfare is $u - h - g_i$ since both act-utility and guilt are experienced by the individual committing the act and since 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

[11]See, for example, Frederick and Loewenstein (1999), suggesting that there is a substantial (though not universal) regularity in the tendency of mental reactions to stimuli to fall as the stimuli are repeated.

[12]Expression (1) may naturally be interpreted as the welfare of a representative individual. Alternatively, one may interpret (1) as the average welfare of a group of possibly heterogeneous individuals, an extension that we discuss in subsection 4.5 (in which case the constraints (3) and (4) would need to be modeled differently).

the realization of guilt and virtue are

$$(3)\quad \sum_{i=1}^{n} y_i(g_i,v_i) \le G \text{ and}$$

$$(4)\quad \sum_{i=1}^{n} z_i(g_i,v_i) \le V,$$

where

$$(5)\quad y_i(g_i,v_i) = p_i \int_0^{\infty} \int_{g_i+v_i}^{\infty} g_i f_i(u,h)\,du\,dh = p_i g_i\left(1-F_i(g_i+v_i)\right) \text{ and}$$

$$(6)\quad z_i(g_i,v_i) = p_i \int_0^{\infty} \int_0^{g_i+v_i} v_i f_i(u,h)\,du\,dh = p_i v_i F_i(g_i+v_i).$$

Here, $F_i(g_i+v_i)$ is the frequency with which $u \le g_i+v_i$ on the subset S_i, that is, the fraction of acts that are deterred, and, correspondingly, $1-F_i(g_i+v_i)$ is the fraction of acts in S_i that are not deterred.

The Lagrangian for the problem of maximizing welfare (1) subject to the constraints (3) and (4) is

$$(7)\quad \sum_{i=1}^{n} W_i(g_i,v_i) - \lambda\left[\sum_{i=1}^{n} y_i(g_i,v_i) - G\right] - \mu\left[\sum_{i=1}^{n} z_i(g_i,v_i) - V\right],$$

where λ and μ are the multipliers for the constraints on the use of guilt and virtue (the shadow price or cost associated with the use of additional units of experienced guilt and experienced virtue to control acts in S_i when the constraints are binding).

The first-order condition if the optimal level of guilt on subset S_i, g_i^*, is greater than 0 is[13]

[13]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.

$$(8)\quad p_i\left[\int_0^{\infty}(h+\lambda g_i-\mu v_i)f_i(g_i+v_i,h)dh \;-\; (1+\lambda)\left(1-F_i(g_i+v_i)\right)\right] \;=\; \alpha_i'(g_i),$$

and the first-order condition if $v_i^* > 0$ is

$$(9)\quad p_i\left[\int_0^{\infty}(h+\lambda g_i-\mu v_i)f_i(g_i+v_i,h)dh \;+\; (1-\mu)F_i(g_i+v_i)\right] \;=\; \beta_i'(v_i).$$

The two terms in brackets on the left sides of (8) and (9) correspond to marginal and inframarginal effects of raising g_i and v_i. The first (integral) term reflects 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 of (8) and (9) 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.

3.1. Basic results. — We now state basic characteristics of the optimum; these are proved in the Appendix.

Proposition. For each subset S_i:

a. *If $g_i^* > 0$, guilt may sometimes be experienced; if $v_i^* > 0$, virtue may not always be experienced when situations in S_i arise.*

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

c. *If at the optimum $\beta_i'(0) > (1-\mu)p_i$, then positive guilt and/or positive virtue are instilled only if not acting is first-best in some subset of situations in S_i having positive probability.*

Part (a) states that, even when it is optimal to use guilt and/or virtue, they may not always succeed in controlling behavior. This is a consequence of the grouping of acts, that is, the need for guilt and virtue to be uniform within each subset of situations. For example, most acts in a subset may be quite harmful and produce little act-utility, making it desirable to deter them with guilt and/or virtue, but a few acts in the subset may yield very high act-utility, in which case these acts may well be committed and thus guilt would then be experienced — unless a very high level of guilt and/or virtue is employed, but that may be too costly. (Note that, if we instead had made the assumption that guilt and virtue are chosen separately for each possible situation (u,h), rather than at uniform levels for groups of acts, part (a) would not hold: For acts that cannot successfully be deterred at reasonable cost, it would not be optimal to employ any guilt or virtue.[14] An implication is that guilt would never be experienced.)

Part (b) can be explained in a straightforward manner. Undesirable acts may be committed because inculcation costs are high or because guilt and virtue are scarce. The possibility that desirable acts may be deterred is a further consequence of grouping: Although most acts in a subset may be undesirable and thus it may be optimal to use guilt or virtue to deter them, there may be some acts in the subset with atypically low h (e.g., $h = 0$) — acts that would not optimally be deterred if guilt and virtue were chosen separately for every conceivable situation (u,h).

Part (c) states that guilt and virtue will only be used for a subset S_i if there exist acts in S_i that it would be desirable to deter. This result is not entirely straightforward because the experiencing of virtue is itself a source of utility. The stated assumption is a condition sufficient to rule out the possibility that it would ever be optimal to use virtue solely because of the benefit of it being experienced. Either the marginal inculcation cost can be sufficiently high ($\beta_i'(0) > p_i$ will suffice) or the constraint on the use of virtue can be sufficiently binding ($\mu > 1$), or some combination of the two. (Regarding the constraint, observe that if virtue is scarce, the only question is *where* to use virtue rather than *how much* total virtue to use, and it will thus tend to be optimal to use virtue where the benefits of controlling behavior are the greatest.) With regard to guilt, once the possibility of using virtue solely for the sake of its being experienced is ruled out, the result is obvious: 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

[14]In the working paper version of this article, we analyze the case in which guilt and virtue are chosen separately for each possible situation (u,h). This case can be thought of as an ideal, if unrealistic, benchmark case, much as one views a setting with complete futures markets or a complete contingent contract.

it to be optimal to inculcate guilt.

3.2. Comparison of the optimal levels of guilt and virtue to the optimal level of a Pigouvian tax. — A further result is that it may well be optimal for the sums of guilt and virtue to be either above or below the Pigouvian tax benchmark, namely, the expected level of the externality.[15]

Consider a case in which it is optimal to use only guilt. Then, g_i^* may be less than the expected level of harm (associated with acts in marginal situations, those for which $u = g_i$) for three reasons. First, guilt is costly to instill. Second, for individuals who are undeterred, guilt is experienced, which in turn reduces utility. Third, guilt is scarce; if the constraint on the use of guilt is binding, a lower level of guilt raises welfare on that account. An implication of these points is that $g_i^* = 0$ is possible.

It is also possible that g_i^* exceeds the expected harm. Note initially that guilt is socially costly when experienced. Thus, it may be optimal to deter some acts that it would be first-best to have committed (i.e., for which $u > h$) because of the benefit of reducing the disutility associated with experiencing guilt. (When the deterrent effect exceeds the inframarginal effect, that is, when $g_i f_i(g_i) > 1-F_i(g_i)$, raising g_i will reduce the aggregate amount of guilt that is experienced.[16]) To illustrate, suppose that the only situations for which $u > h$ are such that u is just slightly above h. If the marginal cost of raising g_i is not too large, 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 direct social loss of $u - h$, which is assumed to be small. This social loss may be less than the benefit of avoiding the utility loss g_i that would be suffered when these acts are committed, and by enough to exceed the additional inculcation cost.[17] Furthermore, because raising g_i can reduce the total amount of guilt experienced, raising g_i may relax the constraint (3), which is valuable when the constraint is binding.

3.3. The choice between use of guilt and virtue as incentives. — Comparison of the first-order conditions (8) and (9) 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 in controlling behavior in a subset S_i. The marginal net benefits of using guilt and virtue (the first terms on the left sides of (8) and (9)) are identical, reflecting the fact that they are substitutes as deterrents. The marginal inculcation costs (the right sides of (8) and (9)) are symmetric, so this consideration favors using whichever, guilt or virtue, has the lower marginal inculcation cost. So far, there is thus no qualitative difference between the desirability of virtue and guilt as incentives.

---

[15]Our exposition involves an oversimplification in the case in which harm is unobservable and may not be 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 interpretation.

[16]We use $f_i(g_i)$ to denote the density function associated with $F_i(g_i)$.

[17]Observe that, if g_i were set equal to the expected harm, then raising g_i slightly would involve a loss in act-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 benefit from relaxing the guilt constraint.

However, consideration of the inframarginal effects of using guilt and virtue (the second terms on the left sides of (8) and (9)) suggests an important qualitative difference between them. 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.

4. Discussion

In this section, we comment on our assumptions, suggest extensions, and offer further remarks.

4.1. Inculcation costs and constraints on the experiencing of guilt and virtue. — We assume in the model that there is a cost of inculcating guilt or virtue for each subset of acts[18] because it presumably takes time and effort to inculcate guilt or virtue. The inculcation cost technology could take a number of different forms. For example, the total of guilt and virtue inculcated might determine the cost, without regard to how much of each is inculcated; or more frequently occurring acts might have lower costs of inculcation because they afford more learning opportunities.

We do not examine the possibility that inculcation involves 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 in any manner and because our constraints on the experiencing of guilt and virtue have an aggregate form, introducing tradeoffs among acts in the inculcation process would not have changed our results significantly.[19]

---

[18]Although we focus on inculcation, evolutionary explanations would tend to produce similar results because there is in a sense scarcity in natural selection: The greater the marginal benefit of a trait, the more likely (and more rapidly) it will tend to be selected; hence, when the marginal return to additional guilt or virtue is lower, we will not see as much guilt or virtue arise.

[19]The main effect of using a technology under which increasing the use of guilt or virtue for one subset of acts raises the marginal cost of using guilt or virtue for other subsets of acts would have been to make it optimal to use less guilt and virtue for any particular subset of acts. But since we do not specify how high is the marginal cost of instilling guilt and virtue for each subset of acts, and since one could interpret each of the separate guilt and virtue cost functions as incorporating the average extent to which other costs are raised as one increases the use of guilt and virtue for a particular act or subset of acts, the analysis would be much the same.

We also assume that there are constraints on the total amounts of guilt and of virtue that can be experienced. The motivation is that emotions — including moral emotions — tend to be relative and, like many other feelings and stimuli, our neurological system is most sensitive to changes, often becoming numb to repetition of the same experience. (Thus, one may become numb to pain or to positive experiences, such as incremental consumption of sugar not tasting as sweet as one ingests larger quantities on a single occasion.) In the present context, the import is that one cannot feel tremendously guilty or virtuous all of the time.[20] All of this seems plausible to us, as a matter of introspection and observation and is supported by psychological evidence, see Frederick and Loewenstein (1999); but we are not yet aware of what research may exist that would allow a more precise statement of the phenomenon in the present context.[21]

Rather than employ a simple constraint, we could assume that, as the total amount of experienced guilt or virtue increases, its marginal effectiveness diminishes.[22] The shadow prices in the first-order conditions (8) and (9) would then 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 deployed to control other acts). Similar qualitative conclusions to ours could be obtained (although the exposition would be more cumbersome). Moreover, in such a model — in which guilt and virtue are not literally fixed in supply — the fact that the experiencing of guilt and virtue affects welfare would have a more clearly identifiable effect on the optimum: Ceteris paribus, this consideration favors using somewhat less guilt and more virtue than otherwise.

Finally, we note that there may also be limits on the ability to feel guilt or virtue for a particular act, simply because, at any given moment, there are limits to how much guilt or virtue one can experience. Had we included such a limitation in our model, the results would not be

---

[20]One simplification we made is that we did not take a certain sort of "credibility" issue into account. Notably, we assumed that, for example, the prospect of guilt could deter even though, if one were not deterred and actually experienced the guilt, the constraint might be violated. This seems to be a modest consideration (one may simply need to stop short of the guilt constraint by enough to leave room to deter the marginal act), but if a very high level of guilt is to be used for some types of acts, or if there is a per act constraint of the sort suggested in the text to follow, this problem could be more important.

[21]In our consideration of external sanctions and rewards, see subsection 4.4, it seems that there are also reasons to assume that there is an aggregate constraint (or at least diminishing marginal effectiveness or increasing marginal cost in using moral sanctions and rewards). These reasons include the costs to individuals who mete out the sanctions and rewards, in terms of time and effort, and the crowding out of moral messages in the public domain, as well as corresponding limits on the targets of disapprobation and approbation to react to external sanctions and rewards. Another factor could be that social esteem is to some degree a relative phenomenon, making social sanctions and rewards, to an extent, zero-sum.

[22]Consider, for example, the following model. The term g_i continues 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 now refers not to the constraint on guilt that may be experienced but rather to the total amount of guilt that will be experienced. Finally, γ_i would be assumed to be increasing in g_i and decreasing (or at least not increasing) 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 (8) and (9); the main difference is that described in the text to follow.

greatly affected because we already assume that there are increasing marginal inculcation costs for each subset of situations.[23]

4.2. Evolution and inculcation. — In modeling common morality and in applying the results to analyze actual moral systems, it is important to understand the roles of evolution and inculcation (nature and nurture).[24] Initially, we observe that the general capacity to feel 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, for example, Darwin (1874), E.O. Wilson (1975), and Izard (1991).[25] Likewise, the manner by which guilt and virtue may be inculcated, and associated limitations or costs, must have biological foundations in the way that our brains process information and in the mechanisms by which various emotions are triggered.

Our analysis is also motivated by the idea that society is able to influence which types of acts are subject to the moral sentiments and to what extent. 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.[26] That human nature is indeed programmable in this sense is further implied by a wide range of practices, notably, substantial efforts to inculcate 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. This is particularly apparent in extreme cases, in which feelings of patriotism or fidelity to a religious belief are able to motivate individuals or groups to engage in suicidal behavior. 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.[27] It is not surprising that morality has such flexibility because it would seem to confer an evolutionary advantage by allowing adaptation to changed circumstances. (Nevertheless, some of our particular feelings of guilt and virtue may have more direct evolutionary foundations. Thus, although society may instill guilt for stealing, it is also true that a hard-wired reluctance to steal may have arisen to help overcome acquisitive urges that, if freely acted upon, would be met with retaliation, which can prove very costly to the initial

---

[23]In all, we have one restriction (inculcation costs) on the *inculcation* of guilt and virtue and another (the pool constraint) on the *use* of guilt and virtue. In addition, we have one restriction (inculcation costs) that is *per subset* and another (the pool constraint) that is *across subsets*. Hence, our particular assumptions capture aspects of a number of plausible features that could have been included separately.

[24]Many of the ideas discussed in this subsection are developed in the literature cited in note 3.

[25]See also Darwin (1874) and de Waal (1996), who suggest that certain other species exhibit aspects of morality and conscience, and see Darwin (1872), who argues at length that the facial expressions corresponding to different emotions are universal in humans and evident in some other species and hence must have an evolutionary origin (which implies that the emotions being expressed must too have an evolutionary origin).

[26]See many of the references cited in note 3. 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, and this occurs 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.

[27]See, for example, Izard (1991), Nisbett and Cohen (1996), Tangney and Fischer (1995).

aggressor.)

The relative roles of evolution and inculcation are also important in assessing whether one would expect actual moral systems to maximize social welfare. First, consider whether it is indeed welfare that would be maximized. Evolution tends to maximize survival (more precisely, replication of the pertinent genes) whereas inculcation, particularly in a society not on the brink of subsistence, may reflect a concern with maximizing welfare.[28] If controlling aggression was (in the relevant evolutionary period) far more important to survival than helping others pursue their ambitions, and if the pattern of moral emotions is determined primarily by evolution, one would predict a heavy use of guilt to control aggression but little use of guilt or virtue to induce individuals to assist others' attempts to maximize their utility. Nevertheless, some acts of helping may have been important to survival, such as sharing food among members of one's tribal group (as long as they did not shirk), as a form of insurance. If so, guilt or virtue might be used heavily to encourage cooperative, sharing behavior. As a result, if inculcation can affect the situations in which cooperation can be induced, this human capacity can be usefully employed to serve a wider range of purposes (including the promotion of welfare) and thus be more adaptive to modern circumstances.

A separate, but important point is that even if social inculcation is operative and welfare rather than survival is the concern, most processes of inculcation would not reliably maximize *social* welfare. Parents may be concerned primarily with their family's well-being, religious or other organizations with the welfare of their members, and governments with their constituents to an extent (depending on the form of government) and also with promoting government officials' self-interest. There may be overlap between these objectives and broader social welfare. For example, parents may find it easiest to teach simple rules (do not lie or steal, do not free ride) to children to regulate behavior within the family, but the same rules may tend to spill over to their children's behavior in interacting with others. Also, parents might teach broad rules for prudential reasons, believing that their children are more likely to be successful in society at large if they internalize norms of honesty and cooperation, because individuals thus inculcated will find better opportunities available to them.

Second, the tendency for moral systems to be optimal — with reference to whatever is being maximized — 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, reciprocity, and so forth). Also, with evolution, there must be a feasible, step-by-step path for a desirable trait to arise. With inculcation, there is the problem that inculcators do not bear all the costs and benefits of their actions. For example, parents may fail to inculcate guilt concerning a type of behavior that does not harm other family members or contribute to the ability to establish a reputation.

---

[28]This explains why humans do not have a vastly greater capacity for virtue, the experiencing of which itself contributes to welfare. As a matter of survival, guilt and virtue, being substitutes, would seem equally useful. But, as we have noted, by using guilt to control behaviors that usually are successfully controlled and virtue for behaviors where few can be induced to behave optimally, one can economize on the use of these emotions. Thus, it may have been optimal for evolution to have produced some capacity for each rather than, say, the same total capacity, but all for guilt or all for virtue.

Likewise, 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.

4.3. Grouping of situations. — We assume that certain situations are naturally (and exogenously) grouped into distinct subsets, so that if, say, guilt is to be inculcated for a particular act, it is inculcated (at the same level) for all acts in the subset. In this subsection, we set forth some of the motivations for our assumption and consider certain modifications.

Motivations related to human psychology. — Many of the reasons it is appropriate to think of acts as falling into groups rather than being considered individually have to do with the organization of our brain.[29] Initially, an individual must perceive the relevant characteristics of a situation even to know what options are available and for it to be possible for emotions, such as guilt and virtue, to be triggered. Yet perception does not simply involve the brain's instant and perfect absorption of surrounding stimuli (which themselves may not constitute a complete depiction of all that may be relevant). Rather, our minds make use of various rules of interpretation and other techniques of pattern recognition in order to construct and categorize mental images. This process involves groupings of sorts, many of which are beyond our conscious control. With regard to perception, emotions, and other brain functions, no doubt, there are important scale economies: It is easier to apply a single response to a range of activity than to have systems and responses customized for each task. There is scarcity in the evolutionary process that limits how such systems develop. Moreover, generality is directly valuable. For example, if the mechanisms supporting some cooperation among individuals were highly specialized, applicable only to the precise instances that had previously and repeatedly been confronted by a species, then even slight changes in the environment would render prior systems and rules useless.

Grouping is also favored by 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. Perhaps more importantly, the proper functioning of the moral emotions requires that their application be largely automatic. If whether one ultimately feels guilty depends upon a 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 their act. This phenomenon would undermine the function of guilt in regulating behavior that harms others. When one adds that moral rules are inculcated to a significant degree during childhood, these points assume greater significance. Thus, it seems plausible that there are important limits on how refined the categories of acts can be, consistent with guilt and virtue being effective motivators of socially desirable behavior.

Motivations relating to inculcation costs; endogenous groupings. — Despite the foregoing, groupings are not entirely fixed; hence, a natural extension of our model would allow

---

[29]See, for example, Kosslyn and Koenig (1992) and Pinker (1997).

choice as to the breadth of categories over which to inculcate guilt or virtue, or might allow for the inculcation of exceptions to general rules. Regarding the former, broader grouping would still tend to be favored because there probably exist significant scale economies. That is, it may be less costly, presumably far less costly, to teach the lesson that one should not lie than to teach the same lesson separately for each and every possible lie one might ever be in a position to tell. Even if some benefits from precise tailoring of levels of guilt to characteristics of acts are lost through grouping, the cost savings would often justify the practice. Relatedly, the optimality of inculcating guilt or virtue at all depends on the frequency with which situations will arise (because the inculcation costs are fixed, borne ex ante, whereas the benefits are ex post and depend on whether and how often situations arise), so it will tend to be optimal to engage in wholesale inculcation for acts that, taken alone, are infrequent, but combined in a sufficiently large group, are frequent.

When the resulting groupings are broad, and thus in some instances overinclusive, it might make sense to expend additional resources to inculcate an exception. For example, certain "white lies," such as those needed to facilitate a surprise party, may not be subject to guilt. An alternative, of course, would be to inculcate guilt over a narrower subset in the first place; which approach is cheaper will depend on the technology, which is to say, on whether individuals (typically, children) can more readily learn in one or another fashion.[30]

Overlapping groups. — We assume that the subsets over which guilt and virtue are inculcated are distinct, but in reality they sometimes overlap. For example, physically interfering with others may be a group of acts subject to guilt, and aiding others in need may be subject to virtue; but it is possible that a person would push someone away in order to help someone else who is in distress. A prospective rescuer may help the person in distress and thereby feel virtuous, but still feel guilty for having pushed someone out of the way. Or, the prospect of that guilt, when combined with the rescuer's own direct costs of aiding another, may exceed the virtue he would feel, thus deterring the act of assistance. Society might choose to inculcate an exception to the physical interference rule for cases of rescue, but as noted above this would be costly.

An interesting case of overlap involves moral rules that do not apply only to a particular type of act (such as lying or stealing) but rather quite broadly. Notably, the Golden Rule enjoins individuals always to take into account the effects of their behavior on others. One can understand such a rule as associating guilt with all undesirable acts and/or virtue with all desirable acts, perhaps with the levels of guilt or virtue rising with the extent of negative or positive

---

[30]It should be apparent that there is an important relationship between the sort of grouping that is assumed and the form of the inculcation cost functions. An alternative modeling approach could posit a single cost function that depends on the level of guilt and virtue inculcated for each act, allowing for interdependencies, which would thereby make it possible to capture the possible natural groupings of acts. We did not adopt this formulation because, at the level of the analysis we have undertaken, the exposition would have been needlessly complex and would have made less transparent our basic points about the grouping of acts. (Consider that the implication of two acts being in the same group is not merely that the marginal cost of inculcating, say, guilt for one act falls — in our case, to zero — when one inculcates guilt for the other act, but also that one *must* have the same level of guilt for the other act — so that, in our model, there is implicitly an infinite marginal cost of reducing guilt for an act below the level of guilt for any other act in the same subset.)

externalities. This raises the question of why society does not simply inculcate the Golden Rule or some variant, eschewing all other rules, and thereby in effect enjoin all individuals always to act in a socially optimal manner.

The foregoing discussion suggests many reasons. It would be difficult to inculcate the command to engage in complex calculations concerning all behavior to young children, even as adults the application of such a rule would be costly and difficult, and there would arise the problem of rationalization (that individuals would miscalculate in their own self-interest to avoid the restraining force of guilt).[31] Moreover, our analysis suggests that, even if successful, such a broad rule 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. With the Golden Rule in full force, many individuals would still commit undesirable acts, which would quickly consume the scarce pool of guilt, making it difficult to deter other acts that may be more important to control; likewise, if virtue were instilled for all good acts, virtue would rapidly be consumed on routine good behavior, leaving little to encourage certain types of behavior that may be particularly valuable. Thus, although broad rules like the Golden Rule have the benefit of being all-inclusive, it seems that such rules would optimally be associated with only modest levels of guilt and virtue, and they would be supplemented by the more focused kind of moral rules that we have emphasized.

4.4. 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.[32] These external analogues to the moral sentiments complement guilt and virtue in regulating individuals' behavior.

Despite the obvious similarities between these internal and external sanctions (hereinafter to include rewards), a more complete analysis would also take into account their differences. External sanctions require the actions of third parties, sometimes one's victim (or, in the case of helpful acts, beneficiary) but often unrelated individuals. 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 seems quite closely related to the internal sanctions and rewards of guilt and virtue: It would appear that those who would feel guilty committing an act would usually feel badly if others express disapproval, and vice versa. The second element, individuals' motivation to impose sanctions on actors, cannot be taken for

---

[31]See, for example, Smith (1790), Brandt (1996), Hare (1981), and Mackie (1985). In addition, Cosmides and Tooby (1994) suggest that the human mind is better at specialized than general problem solving, implying that we are more capable of properly applying rules targeted to particular contexts than a broad command like the Golden Rule.

[32]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.

granted.[33] 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. The first element, third parties' information about the actor's behavior, is an independent factor; in some contexts, certain third parties will automatically learn about behavior; in others, they may learn about it indirectly, such as through gossip (which itself requires information and motivation).

To model disapprobation and approbation explicitly using our framework, one would first define such behavior as involving additional subsets of acts, which themselves might have moral sentiments associated with them. For example, one may be motivated to express disapproval of someone who behaved badly — perhaps by shunning him rather than continuing to greet him cheerfully — because one would feel disgusted associating with him.[34] The externality associated with the act would be the act's effect on welfare through enforcing or undermining, as the case may be, the moral rules that directly govern primary behavior — under the assumption that those subject to blame or praise care about this and accordingly will be induced to comply with moral rules by the prospect of external sanctions. Likewise, there may be guilt and virtue associated with conveying information about others' behavior.

As in our analysis, the moral sentiments would not only affect behavior but would also sometimes be experienced by third parties, which itself would affect social welfare. Furthermore, disapprobation and approbation would sometimes be experienced by primary actors, which would affect their utility and may also involve third parties incurring costs of expression. Considering our other assumptions, there would also be costs associated with inculcating guilt and virtue with regard to external sanctioning behavior, although there may be synergies, as suggested previously: If guilt is to be inculcated for committing a particular type of act, it may not add much cost, if any, simultaneously to inculcate a sense of disgust at others' commission of that type of act, which in turn would lead one to express disapprobation. Moreover, there would be indirect costs associated with constraints on the use of moral emotions. For example, there are undoubtedly limits on the extent to which individuals can be perpetually upset at third parties' behavior and on the ability of individuals to express their disapproval in a manner that influences others.

4.5. Heterogeneity of actors. — Our model can be interpreted as applying to a representative individual in a society. The differences in utilities and external harms or benefits are thus understood as referring to different acts or situations, not to different people. However, no two individuals are entirely alike.

---

[33]Some external sanctions are motivated by ordinary self-interest, such as when one chooses not to deal with a third party known to be unreliable. We view this as distinct from the expression of disapprobation for its own sake, which may include refusal to deal with an unreliable party even when it would be in one's interest to do so in spite of his unreliability. Of course, reputational sanctions motivated by self-interest, narrowly and conventionally understood, sometimes reinforce moral sanctions. Interestingly, even when reputational sanctions operate, morality may be at work, for the third party's misbehavior is, one supposes, taken as a signal of his underlying type — here, perhaps, the extent to which he feels guilty when he behaves opportunistically. See our discussion of heterogeneity, in subsection 4.5.

[34]Moreover, society might use a further level of external sanctions to enforce third parties' enforcement against primary behavior, and so forth. See, for example, Axelrod (1986).

Some heterogeneity could be incorporated with little modification of our model. In particular, if individuals' utilities of acts or the external effects of their acts differ, they can simply be labeled as different 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 naturally apply separately for 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. Thus, to the extent that one can speak of a social decision — or an evolved tendency — for guilt or virtue of a specific magnitude to be associated with a class of acts, one will be speaking about averages, not about the moral emotions of each and every individual. To model this, 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 heterogeneity on our analysis would be to augment the impact of the grouping of situations that themselves are heterogeneous. For example, when we described the possibility that a given level of guilt might deter most but not all acts in a given natural cluster, one could think of an additional reason being that some individuals, when committing acts in that cluster, would fail to be deterred not because their particular situation involves an unusually high level of utility from committing the act, but rather because they experience atypically low levels of guilt. (Psychopaths may be viewed as extreme cases.) Individual heterogeneity also helps to explain why even modest levels of inculcated virtue will induce some individuals (such as Mother Theresa) to do desirable acts that most individuals could not be induced to commit even by the prospect of great rewards.[35]

4.6. Prudence. — Many acts involve no externalities of a conventional sort. Accordingly, there would seem to be no role for the use of guilt and virtue to regulate them because, in the absence of moral sanctions, individuals would commit such acts if and only if their own benefit from doing so was positive, and this behavior would be socially optimal. Nevertheless, discussions of virtue and vice over the ages have often included categories of acts that seem to involve only self-regarding behavior. And psychologists indicate that individuals experience guilt when they act in ways that harm 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

---

[35]Additionally, as suggested in note 33, heterogeneity helps to explain certain responses to others' past behavior, such as refusing to deal with someone who is of an untrustworthy type, which might be translated as the person having little capacity to experience guilt or as the person not having been well inculcated with respect to certain moral rules. Another issue made more important in a model with heterogeneity — when some individuals are deterred and others are not — is that it may be more difficult to inculcate or maintain the effectiveness of guilt (as well as a social practice of expressing disapprobation) for committing an act if too many individuals are committing the act.

folly, and individuals who fail to do so may feel guilty.

One explanation is that externalities are in fact associated with apparently self-regarding behavior. Others may feel badly when individuals act in ways that harm themselves; moreover, such others might be motivated to expend resources to aid those who have fallen victim to their own imprudence. Another explanation is that individuals may lack self-control. (This explanation is particularly important because it constitutes a reason that imprudent behavior might arise in the first place.) In particular, many instances in which guilt and virtue seem to be associated with self-regarding behavior involve problems of myopia. These problems can be thought of as involving two selves — in the case of myopia, a present self whose decisions negatively affect a future self — as Schelling (1984), Thaler and Shefrin (1981), and others have suggested. Under such a formulation, the behavior of the present self does create an externality, affecting a different self, and hence our analysis suggesting the potential benefits of employing guilt and virtue is applicable.

4.7. Relationship to literature on moral philosophy. — Most twentieth-century moral philosophers do not view moral rules as a system that is supposed to maximize social welfare. Instead, they conceive of moral rules as indicating which acts are intrinsically right or wrong. Furthermore, such philosophers frequently present situations in which our moral instincts and intuitions indicate to be wrong acts that consequentialist (often utilitarian) accounts of morality would endorse. Familiar examples include cases in which an actor would have to kill a person to save many, or where a sheriff, by framing an innocent person, could avoid a riot.

In contrast, some earlier philosophers, notably Hume (1739, 1751), Mill (1861), and Sidgwick (1907), argued that when one examines the conventional categories of virtue and vice, one discovers that nearly all rules of common morality serve to promote welfare. These scholars, along with some modern writers, advance what is now described as a two-level view of morality.[36] At the first (higher) level is the ultimate criterion of judgment, social welfare. At the second (lower) level are the imperfect moral rules that are supposed to guide behavior. (These correspond to the subsets of acts in our model, and the corresponding uniform levels of guilt and virtue applying to acts in each subset.)

Implicit in their analysis is what economists would recognize as a standard problem of constrained maximization. Because the choices regarding the design of moral rules are limited — due to given facts of human nature — this problem is of a second-best character. Accordingly, one does not expect the moral system to generate ideal behavior (by the first-level standard) in all cases. In particular, some acts will violate moral rules (i.e, be subject to guilt in our model) — and thus be deemed "wrong" — even though the acts, if committed, would raise social welfare;

[36]This concept is often associated with rule utilitarianism, in contrast to act utilitarianism, but discussions of the subject often fail to illuminate because there is so much confusion about the meaning of each version of utilitarianism and whether, at a deep level, they can be distinguished at all. Twentieth-century two-level accounts that seek to address these issues include Brandt (1996), Hare (1981), Harrod (1936), and Rawls (1955).

likewise, some "right" acts may be welfare-reducing.[37]

This two-level view of morality helps to reconcile our moral instincts and intuitions, which sometimes conflict with the maximization of welfare, with the view that our moral system tends to advance welfare. Because moral rules must apply to categories of situations, it is inevitable that they will sometimes classify as wrong a particular act that in fact would increase welfare. (Thus, the act of killing an innocent person when many would be saved may be in the general category of killing innocent people, there being no exception for unusual circumstances that would be unlikely to arise, especially when a circumstance-dependent exception may lead individuals to rationalize undesirable behavior more often than it would encourage beneficial acts of killing.) Accordingly, the occasional failure of our intuitive sense of right and wrong to reflect first-best choices of acts can be seen simply as an ordinary feature of a (second-best) optimal moral system rather than as a deep problem with consequentialist moral theories.

Our article can be regarded as formalizing and extending the two-level theories of the philosophers who have advanced them. Our main contribution concerns the question of how best to employ moral sanctions and rewards, which they do not consider. In particular, they do not focus on whether guilt or virtue, or some combination of the two, is best to use. In addition, most of them do not take into account that the experience of guilt and virtue is part of utility and, accordingly, will influence the optimal use of the moral sentiments.[38] Nor do they make explicit many of their assumptions about human nature and systematically trace the implications.

5. Conclusion

Our analysis offers a theory of how moral sanctions and rewards — feelings of guilt and virtue — would be optimally employed if the purpose was to maximize social welfare. In this respect, our paper complements economists' extensive attention to other means of regulating externalities, namely, government action and Coasean bargaining.

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 expect a close fit because, as we discuss in subsection 4.2, the processes that produce common morality hardly guarantee optimality and they may not involve the maximization of social welfare in particular. Furthermore, our model considers morality in isolation, whereas the optimal use of morality will depend on the availability of other instruments to control behavior, notably, the legal system, regulation, taxes and subsidies, and so forth. Nevertheless, since our model and results are basic, one might expect there to be some explanatory power.

Most obviously, it is indeed the case that behavior condemned as immoral tends to be

---

[37]Relatedly, psychologists have suggested that moral rules function as decisionmaking heuristics that are subject to error in application due to overgeneralization. See, for example, Baron (1994), Spranca, Minsk, and Baron (1991).

[38]Both Mill (1861) and Sidgwick (1907) recognized that the moral sentiments were a component of welfare, but did not pursue how this should affect the design of a system of morality.

socially undesirable, on account of negative externalities, and that behavior deemed morally praiseworthy tends to be socially desirable, due to positive externalities. In addition, many moral rules (such as those concerning lying, promises, and aggression) apply to groups of related acts; even when there are exceptions or other refinements, categories are typically used rather than having a separate rule finely tailored to the particulars of each possible situation. Moreover, it has long been recognized that certain acts (such as lies in specific situations) might be condemned 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. Furthermore, regarding our result that optimal moral sanctions may be below or above the Pigouvian tax benchmark, it does seem that moral sanctions seem excessively lenient for some types of acts in some moral systems and rather harsh for other types.

Finally, the choice whether to rely primarily on guilt or on virtue seems 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 actually to be 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.[39] Because of differences in both 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 still may be possible to illuminate recognized variations by using a framework like ours that relates moral systems to underlying social conditions.

Also outside our current analysis are normative implications. Others, such as Harsanyi (1953-1954) and Weisbrod (1977), have explored the idea that social welfare may be raised by changing individuals' utility functions, and more recently there has been growing attention to how certain government policies (such as laws having symbolic effects, like civil rights legislation) may be used to reinforce or modify common morality. Our analysis suggests the importance of such inquiries and but raises some cautions. Notably, raising 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 it is not fully successful: Experiencing guilt involves a direct cost and, given the scarcity of the moral sentiments, may reduce their effectiveness in controlling other behavior. Furthermore, when families, religious and educational institutions, and the

[39]See, for example, Miller (2001).

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

It is obviously premature to offer confident statements about the extent to which common morality in our society or others in fact is well designed to maximize welfare or to make pronouncements about how our moral system might be reformed to promote individuals' well-being to a greater extent. Our hope is that the present article serves to illustrate the potential usefulness of explicit economic modeling of what seems to be an important incentive device: the use of guilt and virtue and related external moral sanctions and rewards to regulate behavior.

Appendix

Proof of parts (a) and (b): To prove these parts, it suffices to construct an example for each claim. For all of the claims except the latter claim of part (a), 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 (3) is not binding. To ensure this, suppose that u never exceeds 1, so that g_i^* cannot exceed 1. (As $g = 1$ is sufficient to deter any act, no higher g can be optimal on any subset S_i because the only effect of raising g above 1 would be to increase inculcation costs.) Now, assume that $G > 1$, so that (3) 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 that 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 (8) for $g_i^* > 0$, and moving p_i to the denominator on the right side, we have $1(1-g_i) - (1+0)(1-(\frac{1}{2}+g_i-\frac{1}{2}g_i^2)) = 0.0375/0.1$. Solving this, $g_i^* = 0.5$.[40] Part (a), 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 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 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, that $F(0.1) = 0.99$, $F(1) < 1$, $\beta(0.1) < 0.2$, $\beta(1) > 1$, and $V > 1$.[41] 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 act-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 $v^* < 1$. This is because 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.

Proof of part (c): We first observe that, from expression (7), 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

---

[40]This must be a maximum because 0.5 is the only solution to (8) and the derivative of (7) with respect to g_i is positive at $g_i = 0$ (it is $0.1(1-\frac{1}{2})-0.0375 = 0.0125$).

[41]We observe that these assumptions are consistent with the assumption in part (c) 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 $\beta(0.1) < 0.2$.

expression is positive for all $g_i > 0$ and $v_i > 0$.[42]

$$(A.1)\ \left(W_i(0,0) - \lambda y_i(0,0) - \mu z_i(0,0)\right) - \left(W_i(g_i,v_i) - \lambda y_i(g_i,v_i) - \mu z_i(g_i,v_i)\right)$$

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

$$+ p_i(1+\lambda)g_i\left(1 - F_i(g_i+v_i)\right) - p_i(1-\mu)v_i F_i(g_i+v_i).$$

If $u \geq h$ for all (u,h) in S_i (except possibly on a set of measure zero), then (A.1) will be shown to be positive for any $g_i > 0$ and/or $v_i > 0$, meaning that $g_i^* = 0$ and $v_i^* = 0$ if acting is first-best for all acts in S_i. Now, each of the five terms are obviously strictly positive or equal to zero except possibly for the last one. But we now show that it, combined with the $\beta_i(v_i)-\beta_i(0)$ term, is positive. Given that $\beta_i'(0) > (1-\mu)p_i$, 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, $\beta_i(v_i)-\beta_i(0) > \beta_i'(0)v_i$ 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.

[42]When writing (A.1) we find it convenient, with respect to using expression (2) 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.

References

Akerlof, George A. 1980. A Theory of Social Custom, of Which Unemployment May Be One Consequence. *Quarterly Journal of Economics* 94: 749-75.

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

Axelrod, Robert. 1986. An Evolutionary Approach to Norms. *American Political Science Review* 80: 1095-1111.

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

Baron, Jonathan. 1994. *Thinking and Deciding*. Second edition. Cambridge: Cambridge University Press.

Becker, Gary S. 1996. *Accounting for Tastes*. Cambridge: Harvard University Press.

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

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

Bernheim, B. Douglas. 1994. A Theory of Conformity. *Journal of Political Economy* 102: 841-77.

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

Brandt, Richard B. 1996. *Facts, Values, and Morality*. Cambridge: Cambridge University Press.

Campbell, Donald T. 1975. On the Conflicts Between Biological and Social Evolution and Between Psychology and Moral Tradition. *American Psychologist* 30: 1103-26.

Cosmides, Leda, and John Tooby. 1994. Better than Rational: Evolutionary Psychology and the Invisible Hand. *American Economic Association Papers and Proceedings* 84: 327-32.

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

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

Darwin, Charles. 1872. *The Expression of the Emotions in Man and Animals*. Paul Ekman, ed., third edition. Oxford: Oxford University Press (1998).

Darwin, Charles. 1874. *The Descent of Man; and Selection in Relation to Sex*. Second edition. Amherst, NY: Prometheus Books (1998).

de Waal, Frans. 1996. *The Origins of Right and Wrong in Humans and Other Animals*. Cambridge: Harvard University Press.

Elster, Jon. 1998. Emotions and Economic Theory. *Journal of Economic Literature* 36: 47-74.

Fehr, Ernst, and Klaus M. Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 114: 817-68.

Frank, Robert H. 1988. *Passions within Reason*. New York: W.W. Norton & Co.

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

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings; A Theory of Normative Judgment*. Cambridge: Harvard University Press.

Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D.

Cohen. 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science* 293: 2105-08.
Haidt, Jonathan. 2001. The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review* 108: 814-34.
Hare, R.M. 1981. *Moral Thinking: Its Level, Method, and Point.* Oxford: Oxford University Press.
Harrod, R.F. 1936. Utilitarianism Revised. *Mind* 45: 137-56.
Harsanyi, John C. 1953-1954. Welfare Economics of Variable Tastes. *Review of Economic Studies* 21: 204-13.
Hirshleifer, Jack. 1987. On the Emotions as Guarantors of Threats and Promises. In John Dupré, ed., *The Latest and The Best.* Cambridge: MIT Press.
Hume, David. 1739. *Treatise of Human Nature*. Buffalo: Prometheus Books (1992).
Hume, David. 1751. *An Enquiry Concerning the Principles of Morals*. Tom L. Beauchamp, ed. Oxford: Oxford University Press (1998).
Izard, Carroll E. 1991. *The Psychology of Emotions.* New York: Plenum Press.
Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1987. Fairness and the Assumptions of Economics. In Robin M. Hogarth and Melvin W. Reder, eds., *Rational Choice: The Contrast between Economics and Psychology.* Chicago: University of Chicago 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.
Mackie, J.L. 1985. *Persons and Values: Selected Papers, Volume II.* Joan Mackie and Penelope Mackie, eds. Oxford: Oxford University Press.
Massey, Douglas S. 2002. Emotion and the History of Human Society: The Origin and Role of Emotion in Social Life. *American Sociological Review* 67: 1-29.
Mill, John Stuart. 1861. *Utilitarianism*. Edited by Roger Crisp, New York: Oxford University Press (1998).
Miller, Joan G. 2001. Culture and Moral Development. In David Matsumoto, ed., *The Handbook of Culture and Psychology*. New York: Oxford University Press.
Nisbett, Richard E., and Dov Cohen. 1996. *Culture of Honor: The Psychology of Violence in the South.* Boulder: Westview Press.
Pinker, Steven. 1997. *How the Mind Works.* New York: W.W. Norton & Co.
Rabin, Matthew. 1993. Incorporating Fairness into Game Theory and Economics. *American Economic Review* 83: 1281-1302.
Rawls, John. 1955. Two Concepts of Rules. *Philosophical Review* 64: 3-32.
Robson, Arthur J. 2001. The Biological Basis of Economic Behavior. *Journal of Economic Literature* 39: 11-33.
Schelling, Thomas C. 1984. *Choice and Consequence*. Cambridge: Harvard University Press.
Sidgwick, Henry. 1907. *The Methods of Ethics.* Seventh edition. Indianapolis: Hackett Publishing Company (1981).
Smith, Adam. 1790. *The Theory of the Moral Sentiments.* Sixth edition. Oxford: Oxford University Press (1976).
Spranca, Mark, Elisa Minsk, and Jonathan Baron. 1991. Omission and Commission in Judgment

and Choice. *Journal of Experimental Social Psychology* 27: 76-105.
Tangney, June Price, and Kurt W. Fischer, eds. 1995. *Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. New York: Guilford Press.
Thaler, Richard. H., and H.M. Shefrin. 1981. An Economic Theory of Self-Control. *Journal of Political Economy* 89: 392-406.
Trivers, Robert L. 1971. The Evolution of Reciprocal Altruism. *Quarterly Review of Biology* 46: 35-57.
Weisbrod, Burton A. 1977. Comparing Utility Functions in Efficiency Terms or, What Kind of Utility Functions Do We Want? *American Economic Review* 67: 991-95.
Wilson, Edward O. 1975. *Sociobiology*. Cambridge: Harvard University Press.
Wilson, James Q. 1993. *The Moral Sense*. New York: Simon & Schuster.
Wittman, Donald. 1984. Liability for Harm or Restitution for Benefit? *Journal of Legal Studies* 13: 57-80.