ISSN 1045-6333

MORAL RULES AND THE MORAL SENTIMENTS: TOWARD A THEORY OF AN OPTIMAL MORAL SYSTEM

Louis Kaplow
Steven Shavell

Discussion Paper No. 342

11/2001

Harvard Law School
Cambridge, MA 02138

The Center for Law, Economics, and Business is supported by
a grant from the John M. Olin Foundation.

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

Louis Kaplow and Steven Shavell*

Abstract

We examine how moral sanctions and rewards, notably the moral sentiments involving feelings of guilt and virtue, would be employed to govern individuals' behavior if the objective were to maximize social welfare. In our model, we analyze how the optimal use of guilt and virtue is influenced by the nature of the behavior under consideration, the costs of inculcating moral rules, constraints on the capacity to experience guilt and virtue, the fact that guilt and virtue often must be applied to groups of acts rather than be tailored to every conceivable type of act, and the direct effect of feelings of guilt and virtue on individuals' utility. We also consider a number of ways that the model could be extended, discuss the extent to which our analysis is consistent with the observed use of guilt and virtue, and relate our conclusions to longstanding philosophical debates about morality.

JEL Classes D11, D62, H00, K00

---

*Harvard Law School and National Bureau of Economic Research. We thank David Cope, Robert Ellickson, Oliver Hart, Christopher Kutz, A. Mitchell Polinsky, Eric Posner, Eric Rasmusen, Richard Zeckhauser and participants in workshops at the University of California at Berkeley, Harvard University, and Stanford University 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.

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

Louis Kaplow and Steven Shavell

1. Introduction

In economic analysis of individual behavior, it typically is assumed that individuals are motivated by the direct contribution to utility that would be produced by the actions that they might choose. Yet it is obvious that our moral sentiments — feelings of guilt and virtue, along with their external correlates, disapprobation and praise — are also springs of human action. That is, individuals may be motivated by the prospect of feeling guilty or of feeling virtuous (or, essentially equivalently, by the desire to do what is right) to follow moral rules when doing so would otherwise be contrary to their self-interest, conventionally interpreted.

Whether and to what extent we experience guilt and virtue when we commit certain acts is not arbitrary. It is evident that society's system of morality is the product of a complex process of socialization, especially in childhood, and of evolution. Moreover, it seems plausible that these mechanisms have some tendency, however imperfect, toward maximization.

Against this background, we ask what system of moral rules — and, notably, what use of guilt and virtue to induce individuals to follow these rules — leads to the maximization of social welfare. One motivation for our inquiry is to compare the answer to this theoretical question with what we observe to be society's common morality, particularly to the actual pattern of use of guilt and virtue.[1] In this respect, we build on the work of such writers as Hume (1739, 1751) and Sidgwick (1907), who argued informally that the observed system of morality tends to advance welfare.[2] Another motivation for our analysis is that it can in principle be useful in considering what system of morality should be inculcated.

Recent economic literature on social norms and some writing in behavioral economics (such as that exploring behavior motivated by concerns for "fairness"), along with scholarship in other disciplines, recognize that individuals' behavior is not always narrowly self-interested and may reflect

---

[1]Of course, we would not expect the observed moral system to maximize social welfare, both because evolution (biological and social) is imperfect and because the objectives of some of those who socialize and of evolution differ from social welfare. See our discussion in subsection 5.2.

[2]We also note Brandt (1996), who inquires into the features of an optimal moral system, although not in as explicit manner as we pursue here.

moral concerns.[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 (though we give explanations for the existence of) a particular and broad set of non-self-interested motivations, and our object is to examine how these motivations might optimally be employed to advance welfare. Another strand of literature seeks to explain cooperative behavior as rational, often as a possible equilibrium of a repeated game; in this regard, our article is complementary in that one use to which the moral sentiments could be put is to reinforce cooperation (for example, promises support cooperation if they are credible, and the prospect of feeling guilty 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.

Our conclusions about what moral system is optimal are as follows. Initially, we find, consistent with intuition, that guilt and virtue will tend to be used only to induce individuals to change behavior that involves externalities.[4] However, this simple result is potentially misleading because, in realistic settings, guilt and virtue will typically be assigned wholesale to groups of similar acts (such as all lies). As a consequence, the match between the assignment of guilt or virtue and the ideal provision of incentives so as to promote desirable behavior will be imperfect. Some socially undesirable acts will not be deterred, and, perhaps more interestingly, some socially desirable acts will be deterred (some lies that happen to be socially desirable). In addition, some socially desirable acts that are committed will nevertheless result in the actors bearing moral sanctions; that is, the acts will be treated as wrongs even though the acts promote welfare.

Another of our main conclusions concerns an important but previously neglected question: whether guilt or virtue (or both or neither) should be used to govern particular types of behavior. Although past writers, like Hume and Sidgwick, discussed at length how acts deemed to be vices, and thus subject to guilt and disapprobation, are socially undesirable (telling lies undermines our ability to engage in cooperative, wealth-enhancing activity), whereas acts deemed to be virtuous, and thus resulting in feelings of virtue and in social approval, are socially desirable, they did not attempt to explain why a given act was deemed to be a vice rather than abstention from that act being deemed to

---

[3]Among economists, analysis of the motivating force of moral sentiments begins at least as early as Smith (1790). See also Becker (1996), Ben-Ner and Putterman (1998), Binmore (1998), Fehr and Schmidt (1999), Frank (1988), Geanakoplos, Pearce, and Stacchetti (1989), Hirshleifer (1987), Kahneman, Knetsch, and Thaler (1987), Ostrom (2000), Rabin (1993), Robson (2001), and Romer (1996). Other scholars (in addition to those mentioned in the preceding text), including psychologists, sociologists, anthropologists, neurobiologists, sociobiologists, and philosophers have addressed the role of moral emotions in regulating behavior. See, for example, Alexander (1987), Barkow, Cosmides, and Tooby (1992), Baron (1994), Campbell (1975), Daly and Wilson (1988), Damasio (1994), Darwin (1872, 1874), Elster (1998, 1999), Gibbard (1990), Hechter and Opp (2001), Izard (1991), Kagan (1984), LeDoux (1996), Mackie (1985), Pinker (1997), Trivers (1971), E.O. Wilson (1975), and J.Q. Wilson (1993).

[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. We also consider briefly the use of the moral sentiments to control what may appear to be entirely self-regarding behavior (for example, to enforce self-discipline) in subsection 5.7.

be a virtue. (Why, for instance, do we feel guilty when shoving someone out of our way rather than virtuous when we refrain from such action?) Many disputes in contemporary moral philosophy wrestle with such questions.

Our approach, by contrast, answers the question whether guilt or virtue should be employed, and in a manner that we suggest is roughly consistent with what we observe to be true about our system of moral rules. Notably, our analysis suggests that it will not be optimal to employ guilt when doing so would be insufficient to induce most individuals to act in a socially optimal manner, because most individuals would then simply suffer disutility from feeling guilty. This would be costly directly (the experiencing of guilt in itself reduces welfare) and indirectly (as we discuss, the capacity to experience guilt is limited, so substantial depletion of what is, in a sense, a scarce resource for little behavioral benefit is undesirable). In fact, it seems that virtue rather than guilt is primarily used in this type of case, such as in inducing individuals to rescue others at substantial risk to themselves.

Our article is organized as follows. We begin in section 2, where we set out our framework of analysis. We consider a set of possible situations, in each of which individuals must decide whether or not to commit an act. Acts directly produce utility (positive or negative) for individuals, and they may also result in externalities. Individuals are subject to a process of inculcation such that they will experience guilt or virtue as a function of the choices they make, and, accordingly they will be led to behave other than in their narrow self-interest if the weight of guilt and virtue exceeds their personal benefit from an act.[5] This inculcation process involves a cost. We further suppose that the ability to use guilt and virtue is constrained, because individuals have a limited capacity to experience these moral emotions. (Thus we suppose, for example, that it is impossible for individuals to feel extremely guilty all of the time.) The social problem is taken to be maximization of morally inclusive social welfare, which is to say conventional components of 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.

In section 3, we consider the optimal system of morality when enforcement of moral rules is assumed to be accomplished through guilt alone. We first consider the admittedly unrealistic case in which moral rules can be perfectly specific in character, that is, when the level of guilt can be made to depend on the particulars of each possible situation. (This case is analogous to that of the complete contingent contract in contract theory.) In this case, guilt is optimally used only when it is sufficient to deter an undesirable act that otherwise would be committed, but it is not always so used due to inculcation costs. Moreover, because guilt is used only when deterrence will succeed, guilt is never actually experienced.

---

[5]Certain philosophers, and others, may question our assumption that it is the prospect of feelings of guilt and virtue, as components of individuals' utility, that motivates moral behavior. As we discuss further in note 9 and in subsection 5.8, however, we believe that a range of different interpretations of why individuals behave morally (to the extent that they do) are consistent with our assumption because individuals would nevertheless behave as if they were motivated by the prospect of guilt and virtue.

Using these results as a standard for comparison, we then analyze the case in which the level of guilt cannot be perfectly tailored to each act. In particular, we suppose that there exist natural groups of acts, and if guilt is to be inculcated, it must be inculcated at the same level for all acts in a given set of acts. Although this assumption is obviously an oversimplification (as we discuss later), it seems to capture an important aspect of our environment. When guilt must be uniformly applied to heterogenous acts (for example, lies that have different consequences), a number of the conclusions just mentioned change. First, guilt may sometimes be suffered because the level of guilt that is optimal to employ may not be sufficient to deter all acts in the group. Second, some of the acts that are deterred will be socially desirable ones. Third, when guilt is experienced, it may be due to the commission of an undesirable act or a desirable one. Fourth, the optimal level of guilt may be lower or higher level than would be optimal considering only its effects on behavior and inculcation costs, because raising the level of guilt from that point may increase or reduce the extent to which guilt is actually experienced.

In section 4, we analyze the same two cases — perfectly specific and general moral rules — but this time allowing for the use of both virtue and guilt. Although some of the results obtained are similar to those just described, this is not always so. The reason is that there are two fundamental differences between the use of guilt and of virtue: When optimal behavior is induced through guilt, no guilt is experienced, but when optimal behavior is induced through virtue, virtue is experienced. In addition, when individuals actually experience guilt, social welfare falls on that account, but when they actually experience virtue, social welfare rises as a result. These differences help to explain, as previously noted, why our analysis yields conclusions about the circumstances in which virtue rather than guilt should be employed. We also note that, because virtue is itself a source of utility, the magnitude of inculcation costs and the constraint on the capacity to experience virtue play a more central role than is the case in the analysis of guilt alone.

In section 5, we interpret our results, describe a range of possible extensions, and offer further discussion. In particular, we consider whether the actual use of guilt and virtue is consistent with our results; the relative roles of inculcation and evolution in determining the form of moral rules and the implications thereof for our analysis; our assumptions about inculcation costs and constraints on the use of guilt and virtue; the differences between internal sanctions and rewards (guilt and virtue) and external ones (disapprobation and praise); how acts are grouped and the effect of different assumptions concerning grouping on our results; the relevance of heterogeneity among actors, particularly concerning the extent to which they experience guilt and virtue; the apparent fact that some moral rules govern prudence, that is, behavior that does not seem to involve externalities; the relationship between our analysis and certain strands of the literature on moral philosophy; and the choice between using morality and the law to control behavior. In section 6, we conclude.

Before proceeding, we note that this article is preliminary and speculative in important respects. Our particular assumptions about the costs of inculcation and about limits on the experiencing of guilt and virtue are not grounded in firm knowledge of the evolution and functioning of moral emotions — although they seem plausible and much of what we say does not depend upon the particular formulation that we adopt. In any event, we believe that the main contribution of this article is to illuminate the

structure of moral rules and the moral sentiments by viewing them as an incentive scheme and thus susceptible to analysis using conventional economic tools. More specifically, we examine how the solution to this problem is influenced by various limits on the use of moral sanctions as well as by the fact that moral sanctions and rewards themselves enter into well-being and thus into social welfare.

2. Framework of Analysis

Let S denote the set of possible situations in which individuals may find themselves. In each situation, an individual chooses between committing some act or not doing so.[6] For example, in one situation, an individual might choose whether or not to lie, in another he might choose whether or not to litter, and in another whether or not to read a book. If he chooses the act, the individual obtains utility associated with the act per se (which we sometimes refer to as act-utility) of u, which can be positive or negative. In addition, an act causes an external harm of h $\geq$ 0. If he does not commit the act, he does not obtain any act-utility and does not cause any external harm. The act that an individual may choose in a particular situation is thus identified with a pair (u,h). The possible situations have density f(u,h), which is assumed to be continuous, where u is in $(-4,4)$ and h is in $[0,4)$.

In interpreting the foregoing, a number of observations should be borne in mind. First, the assumption that not acting results in no utility for the person and in no external effect is a convenience; the results that we obtain depend only on the difference between the utilities obtained from acting and not acting, and on the difference between the externalities associated with acting and not acting. Second, and related, our analysis should be understood to apply to acts that cause positive externalities, even though in the model acting can only cause a negative externality: we can label an act that causes a positive externality (such as rescuing a person) as "not acting," so that "acting" (failing to rescue) relative to it causes a negative externality. Third, in stating that the act in a given situation results in a unique level of act-utility u and external harm h, we are describing as distinct acts, for example, lies that might have different consequences. (Later, we will analyze the situation in which non-identical acts, such as all lies or a certain type of lie, are grouped together.)

Assume that society may instill guilt g(u,h) $\geq$ 0 with regard to the commission of an act (u,h) in a particular situation. By this, we mean that a person in that situation will experience guilt — that is, suffer disutility — of g(u,h) if and only if he commits the act (u,h).[7] (As we note in the introduction and

---

[6]More generally, an individual who finds himself in a situation s may choose from a set of n(s) acts $a_1(s)$, $a_2(s),...,a_{n(s)}(s)$, 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 call the act, and the other no act.

[7]We also note that, in principle, society might be able to instill guilt g $\geq$ 0 if a person does *not* commit an act (u,h). We could consider this possibility formally, and we could show that it is suboptimal, so it would never be done. In particular, if acting is first best, this can be achieved without guilt, for then u > h, which implies that u > 0, so that the person will act in the absence of guilt; thus, it can only lower social welfare to incur costs to instill guilt (and possibly impose it) for not acting. If not acting is first best, that is because u < h. In this case, it is always

discuss in subsection 5.4, it is useful to keep in mind when interpreting g(u,h) not only the internally generated experience of feeling the emotion of guilt, but also the disutility associated with disapprobation or blame expressed by others.)

Similarly, assume that society may instill virtue v(u,h) $ 0 for not committing an act (u,h). Virtue has the property that a person obtains utility of v(u,h) if and only if he does not commit the act (u,h).[8] (We note that although we call this source of utility virtue, it can be interpreted not only as the internally-generated positive emotion, but also as the utility from approbation or praise that an individual receives from others.)

The prospect of guilt or virtue can lead an individual to change his behavior.[9] In the absence of

---

better to set g equal to zero for not acting: If g is positive for not acting and it is lowered to 0, the incentive not to act can only be raised, and the costs of instilling g are saved as well as the possibility that guilt will be suffered.

[8]Also, we observe that, in principle, society may instill virtue if a person commits an act. We assume that this does not occur; the implications of allowing for this possibility will be apparent from the discussion in section 4 and are mentioned in note 22.

[9]It is not important for our analysis how individuals actually think about guilt and virtue or whether these feelings, at root, have a common denominator with other sources of utility. For example, it may be that some individuals ask themselves whether an act would be right or wrong, and generally do what is right, without explicitly contemplating that, if they behaved otherwise, they would experience guilt. Such a decisionmaking process can be imagined to be the outcome of prior inculcation and of experience that ultimately becomes crystalized in the form of habit (a theme of Hume (1751), Mill (1861), and Darwin (1874)), or it may be viewed as some other sort of reduced-form internal deliberation in which the role of feelings of guilt and virtue is implicit. It is sufficient for our purposes that individuals behave as if they were motivated by the prospect of guilt and virtue.

It is also unimportant for our purposes whether, as some suggest, moral considerations are qualitatively different from ordinary sources of utility. All that matters here is that individuals make tradeoffs, so that they are likely to refrain from committing a wrongful act if its act-utility is low and its degree of wrongfulness is high, whereas, conversely, if the act-utility is great and the act is only trivially wrong, they would commit the act. Under this interpretation, a unit of g is simply a measure of the degree of wrongfulness associated with the individual being just indifferent with regard to committing a wrongful act producing that level of act-utility.

Another interpretation of the influence of morality on decisionmaking is that morality (whether viewed in a Kantian manner, as a matter of divine commands, or otherwise) is based in decisionmaking processes in our brain that override our ordinary decisionmaking that is based on a balance of pleasure and pain (just as, for example, the brain may send signals to a gland involuntarily, and, in particular, without regard to whether the consequence of sending the signal will increase our utility). Clearly, such matters may be illuminated by neurological study in addition to (or instead of) philosophical inquiry. See, for example, Berridge (1996). Moreover, they raise interesting questions about the meaning of the concept of well-being that underlies the welfare economic approach (notably, whether preferences revealed by certain behaviors are pertinent for normative assessment). Nevertheless, as long as the strength of such other influences can vary and as long as the likelihood that such influences will override the ordinary utility calculus depends on the magnitude of preference indicated by that calculus, the implications for behavior will be much the same as what we present in the text.

In sum, for descriptive purposes, one can, essentially tautologically, define utility as the resultant balance of all relevant forces that affect behavior. (See also our discussion in subsection 5.8 of self-interest as a motivation for moral behavior.) Our assumption, then, is simply that, in addition to narrow self-interest, there may also be "moral" forces that influence individuals' behavior.

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, then 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 will assume that there is a cost (increasing at the margin) of instilling guilt and of instilling virtue. In our first case, in which guilt and virtue may be assigned specifically to each possible situation, such costs will be associated with each situation. In our second case, in which guilt and virtue must be uniform across each set of acts, such costs will be associated with each set of acts.

We will also assume that there is a constraint on the actual experiencing of guilt, namely, that the expected value of experienced guilt cannot exceed an amount $G \geq 0$. Likewise, we will assume that the expected value of experienced virtue cannot exceed an amount $V \geq 0$. The motivation for these assumptions, as we note in the introduction and revisit in subsection 5.3, is that our capacity actually to experience the emotions of guilt or of virtue is limited; 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]

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. Explicit expressions for social welfare will be given below.

The social problem is to instill guilt and virtue so as to maximize social welfare, subject to the constraints on the realization of guilt and virtue. The social problem will be considered first for guilt alone in section 3 and then for guilt and virtue in section 4. In each of these sections, we will consider initially the social problem when guilt and possibly virtue can be selected individually for each act (u, h) and then that when they can be selected only for sets of acts.[12]

---

[10]To avoid having to make tedious qualifications to our analysis and statements of conclusions, we will assume that, when $u = 0$, the person will not act, and we will make similar assumptions about cases of indifference later without further comment.

[11]See, for example, Frederick and Loewenstein (1999). We note that, although there is a substantial regularity to the tendency of mental reactions to stimuli to fall as the stimuli are repeated, there are some exceptions.

[12]There is, however, a formal similarity between some of our analysis involving the use of guilt and that in the literature on optimal law enforcement that addresses the use of socially costly sanctions. See, for example, Polinsky and Shavell (1984), Shavell (1987), and Kaplow (1990).

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. Moral Rules Enforced by Guilt

3.1. Specific moral rules. — In the case under consideration, we assume that guilt g(u,h) may be instilled independently for acting in each situation (u,h) at cost $\psi(g)$, where $\psi(0) = 0$ and, for $g > 0$, $\psi'(g) > 0$ and $\psi''(g) \geq 0$.[13] Hence, the social problem is to assign guilt g(u,h) to each act (u,h) so as to maximize social welfare, subject to the constraint that realized guilt not exceed G. It will be convenient to let A denote the set of acts that are committed, that is, $A = \{(u,h) \mid S^*u > g(u,h)\}$. Therefore, we can write the social welfare maximization problem as choosing the function $g(\cdot,\cdot) \geq 0$ to maximize morally inclusive social welfare[14]

$$(3.1) \quad \int\int_A \big(u - h - g(u,h)\big) f(u,h)\,du\,dh \; - \; \int\int_S \alpha(g(u,h))\,du\,dh,$$

subject to the constraint

$$(3.2) \quad \int\int_A g(u,h) f(u,h)\,du\,dh \; \leq \; G.$$

Note in (3.1) that the first integration is over acts (u,h) that are committed; that when an act is committed, u - h - g(u,h) is the effect on social welfare, since both act-utility and guilt are experienced by the individual committing the act and since the externality occurs; and that the likelihood of the situation (u,h) arising is f(u,h). The second integration reflects the cost $\psi(g(u,h))$ of instilling guilt for each act (u,h). (Because we assume that the inculcation cost for an act is borne up front and thus its level is independent of how often the situation in which the act might be committed arises and of

---

[13] We could allow the inculcation cost function $\psi(\cdot)$ to depend on the particular act, but this would not materially affect our analysis.

[14] Expression (3.1) may naturally be interpreted as the welfare of a representative individual. Alternatively, one may interpret (3.1) as the average welfare of a group of possibly heterogeneous individuals, an extension that we discuss in subsection 5.6 (in which case the constraint (3.2) would need to be modeled differently, see note 55). Moreover, a rigorous interpretation of our constraint (3.2) requires that f(u,h) be interpreted as the fraction of time that an individual will spend in each situation (u,h) rather than as a probability. However, the aforementioned extension allowing for heterogeneity (which formally includes the case in which ex ante identical individuals have different experiences ex post, when uncertainty about the situations in which they will find themselves is resolved) and changing how the constraint (3.2) is modeled would permit f(u,h) to be interpreted as a probability.

whether the act is committed in that situation, there is no weighting by the density f(u,h), and the integration is over the entire set of acts S.) Expression (3.2) states that the expected value of experienced guilt cannot exceed G.

Let g*(u,h) denote the optimal g(∅,∅). We have the following result, which is proved in the Appendix.

Proposition 1. Assume that guilt can be instilled separately for each act. Then, for each act (u,h):[15]

a. positive guilt is instilled only if not acting is first best, and, when guilt is instilled, it equals the minimum necessary to discourage the act; that is, if g(u,h) > 0, then u < h and g*(u,h) = u;*

b. guilt is never actually experienced;

c. the only possible deviation from first-best behavior is the commission of undesirable acts; and

d. it is optimal to instill guilt with respect to a situation in which u > 0 if and only if

$$(3.3) \quad \alpha(u) < (h-u)f(u,h).$$

Notes: Part (a) is readily explained. On one hand, if an act is first best, guilt will not be optimal to instill, for guilt can only discourage the act, guilt involves an expense to instill, and, if g < u, guilt will be experienced, imposing a further cost. On the other hand, if an act is not first best, guilt cannot be useful to instill if it fails to discourage the act, for then guilt is suffered, involving a social cost, and the expense of instilling guilt is incurred. Hence, if an act is not first best, guilt can be optimal to instill only if it discourages the undesirable act. Finally, given that this is so, it must be optimal to spend the least to accomplish this (to minimize inculcation costs), meaning that g will equal u, just offsetting the gain from acting.

Part (b) follows from (a): Since positive guilt is employed only when it successfully deters undesirable acts, it is never experienced. Notice that this means that the constraint (3.2) on the actual experiencing of guilt is irrelevant in the present formulation of the problem. Part (c) follows from (a), which implies that guilt is not employed for desirable acts. Part (d) states that, when guilt is optimally employed, the cost of instilling guilt, which is borne ex ante, must be less than the expected net gains from deterring an undesirable act.

Relationship of the optimal level of guilt g to u, h, ″, and f.* An increase in the utility u

[15]For expositional convenience, we state that our conclusions must hold for each act even though they need not hold with respect to any set of acts of measure zero.

from an undesirable act — an act for which u < h — raises the optimal level of guilt because g* = u, until u becomes so high that guilt is no longer optimal to instill. The other parameters do not affect g* if it is optimal to instill guilt (because g* always just equals u), but they do affect whether it is optimal to instill guilt. Instilling guilt is favored by a higher level of harm h, lower inculcation costs θ, and a greater frequency of the act f(u,h). (The reason for the latter effect is that the inculcation cost θ is incurred once — it is a fixed cost of sorts, as we model it — whereas the situation in which the act may be committed occurs with frequency f(u,h).)

3.2. General moral rules. — Now let us assume that guilt cannot be inculcated independently for each possible act (u,h). Instead, guilt is constrained to be the same for all acts within each of n subsets S_i that partition the universe S of situations.[16] Let g_i denote the uniform level of guilt for acts within S_i.

The motivation for the assumption that guilt is constrained to be constant within S_i is that it is difficult if not impossible to instill guilt at too refined a level, as we mention in the introduction. Another way of expressing this point is that it would be so costly to instill guilt for each conceivable situation as to render this idea fanciful. We further assume that the subsets S_i are exogenously determined. However, a more articulated theory than the one we are now considering might determine the S_i on the basis of cost and of certain psychological factors of similitude among situations; see subsection 5.5 below.

Given the assumptions that we have made, the social problem becomes to choose $g_i \geq 0$ on each subset S_i optimally, subject to the constraint on the realization of guilt. Specifically, 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. Let $\theta_i(g_i)$ denote the cost of instilling guilt g_i for choosing acts within this subset, and assume that the derivatives of $\theta_i(g_i)$ have the same properties as those of $\theta(g)$. Then morally inclusive social welfare is

$$(3.4) \quad \sum_{i=1}^{n} W_i(g_i),$$

[16]We observe that our earlier assumption that each act causes harm $h \geq 0$ involves an implicit restriction when guilt must be uniform for all situations in S_i. In particular, before, we noted that we could incorporate into our model acts that cause a positive externality by labeling such acts as not acting. Now, however, that cannot be done because the make-up of a set S_i is understood to be determined by psychological links that tie together certain acts across situations in S_i. For example, lying in different states might be a natural grouping, so that, if a particular lie causes a positive externality, we cannot simply relabel this lie as not acting. In other words, our assumption implies that no lies create positive externalities (or, we could relabel the entire set if no lies create negative externalities). However, were we to relax the assumption that $h \geq 0$, there would be no difference in the qualitative nature of our conclusions. Allowing for $h < 0$ would increase the potential degree of heterogeneity within a subset S_i, which would tend to reduce the value of inculcating guilt (or, in subsection 4.2, virtue as well), but would not otherwise affect our analysis.

where

$$(3.5)\quad W_i(g_i) = p_i \int_0^{\infty} \int_{g_i}^{\infty} (u-h-g_i) f_i(u,h)\,du\,dh - \alpha_i(g_i).$$

To explain, individuals commit acts in set S_i when $u > g_i$, in which case the effect on social welfare is $u - h - g_i$ (as in the explanation of expression (3.1)), and the cost of instilling g_i is $\alpha_i(g_i)$. The constraint on the actual realization of guilt is

$$(3.6)\quad \sum_{i=1}^{n} y_i(g_i) \le G,$$

where

$$(3.7)\quad y_i(g_i) = p_i \int_0^{\infty} \int_{g_i}^{\infty} g_i f_i(u,h)\,du\,dh = p_i g_i \left(1 - F_i(g_i)\right),$$

where $F_i(g_i)$ is the frequency with which $u < g_i$ on the set S_i (and thus $1-F_i(g_i)$ in (3.7) is the fraction of acts in S_i that are not deterred).

The Lagrangian for the problem of maximizing welfare (3.4) subject to the constraint (3.6) is

$$(3.8)\quad \sum_{i=1}^{n} W_i(g_i) - \lambda \left[\sum_{i=1}^{n} y_i(g_i) - G \right],$$

where λ is the Lagrange multiplier of the constraint — that is, λ is the shadow price or cost associated with the use of additional units of experienced guilt to control acts in S_i when the constraint is binding.

The first-order condition if $g_i^* > 0$ is[17]

$$(3.9)\quad p_i\left[\int_0^{\infty}\left(h+\lambda g_i\right)f_i(g_i,h)dh \;-\; (1+\lambda)\left(1-F_i(g_i)\right)\right] \;=\; \alpha_i'(g_i).$$

On the left side of (3.9), the integral term reflects the *marginal* benefit of deterring additional acts. When g_i is raised slightly, the marginal individuals who are deterred are those for whom $u = g_i$; hence, with regard to the utility experienced by marginal individuals, deterrence has no effect on social welfare. However, when an individual is deterred, the external harm h is also avoided; moreover, when an individual 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 value per unit of 8. Each of these marginal benefits is weighted by $f_i(g_i,h)$, which is the density of individuals deterred at the margin.

The second term on the left side of (3.9) is the *inframarginal* effect on welfare of raising g_i. For those individuals who are not deterred, whose relative proportion in the set S_i is $1-F_i(g_i)$, there are two costs of raising g_i: They suffer an additional unit of guilt (this is the 1 in the 1+8 term), and an additional unit of the constrained pool of guilt is used (which has a shadow price of 8).

These two effects, the marginal (or deterrence) effect and the inframarginal effect, are equated with the direct marginal cost of instilling a higher level of guilt, "$_i$N(g_i).

We now offer some remarks about the implications of (3.9). First, g_i^* may be less than the expected level of harm (associated with the acts of marginal individuals, those for whom $u = g_i$). This is in contrast to the point that the optimal Pigouvian tax equals the expected harm.[18] There are three reasons why the optimal level of guilt may be lower than expected harm. (1) Guilt is costly to instill. (2) For individuals who are undeterred, guilt is experienced, which in turn reduces utility. (3) Guilt is scarce; if the constraint on the use of guilt is binding, a lower level of guilt raises welfare on that account. One implication of these points is that, as in subsection 3.1, it clearly is possible that $g_i^* = 0$.

Second, it is also possible that g_i^* exceeds the expected harm. (1) Because guilt is socially costly when experienced, it may be optimal to deter some acts that it would be first best to commit (i.e.,

---

[17]As will be apparent from the discussion to follow, $g_i^* = 0$ is possible. In addition, this first-order condition is not a sufficient condition for a global optimum.

[18]This is 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.

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.[19]) To consider a simple (discrete) case, suppose that the only individuals for whom u > h have a u that is just slightly above h. If the marginal cost of raising g_i is not too large, then it may be optimal to deter everyone 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 exceeded by the benefit that consists of avoiding the utility loss g_i when these individuals are deterred, and by enough to exceed the additional inculcation cost. (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.) (2) Furthermore, because raising g_i can reduce the total amount of guilt instilled, it may relax the constraint, which is valuable when the constraint is binding.

Another point about the optimum concerns the use of guilt across sets of acts in the case in which the constraint on the use of guilt is binding. This condition can be stated as

$$(3.10) \quad \frac{\partial W_i(g_i)/\partial g_i}{\partial y_i(g_i)/\partial g_i} = \lambda, \textit{ for all } i \textit{ for which } g_i > 0.$$

Expression (3.10) requires that the marginal welfare benefit *per unit of guilt that will be experienced* be equal for all sets of acts where guilt is used. In examining (3.10), we emphasize the interpretation of the denominator because it refers not to the marginal unit of guilt that is inculcated, g_i itself, but rather to the impact of inculcating another unit of guilt on the amount of guilt that is expected to be experienced, $y_i(g_i)$. From our discussion of expression (3.9), it is clear that inculcating additional guilt could raise or lower the amount of guilt that will actually be experienced. Indeed, it is possible that the optimum will be such that $\partial y_i(g_i^*)/\partial g_i < 0$, which in turn, from (3.10), implies that $\partial W_i(g_i^*)/\partial g_i < 0$. (This would be in the case where, at the margin, raising g_i reduces welfare with regard to situations in S_i but is nevertheless desirable because of the extent to which it relaxes the constraint on the use of guilt.)

Let us now state conclusions about the present case of general moral rules, which are proved in the Appendix.

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

Proposition 2. Assume that guilt can only be instilled in a uniform manner within each of the subsets S_i that partition S. Then, within each subset S_i:

a. positive guilt is instilled only if not acting is first best for some acts in S_i; that is, if $g_i^* > 0$, then $u < h$ for some $(u,h) \in S_i$;

b. if $g_i^* > 0$, guilt may sometimes be experienced;

c. both types of deviations from first-best behavior are possible: the commission of undesirable acts and the deterrence of desirable acts; and

d. if $g_i^* > 0$, g_i^* satisfies (3.9).

Notes: Part (a) is obvious because, if all acts in S_i are desirable, then instilling guilt could only deter desirable acts and would waste inculcation costs; also, guilt might be experienced (causing further disutility), and, if experienced, would involve an additional cost if the constraint on the use of guilt is binding.

Part (b) is true because S_i is a set of potentially heterogeneous acts: If guilt is instilled, it will be experienced whenever $u > g_i^*$, and since the distribution of the u's is not restricted, it is clearly possible that some individuals may experience guilt. (One can consider a case in which using some guilt is very valuable because h is large and a small level of guilt will deter almost everyone, but there is a tiny fraction of individuals whose u is very large and it would be extremely expensive in terms of inculcation costs to raise g high enough to deter them — and deterring them may directly be undesirable as well if their u's are high enough.)

The claim in part (c) that some undesirable acts may not be deterred is obvious, as in subsection 3.1, for inculcation costs could be large (and p_i small, and so forth). Unlike the result in subsection 3.1, however, some desirable acts now may be deterred. There are two reasons. First, when $g_i^* > 0$, no matter how low g_i^* is, it is possible that some desirable acts will be deterred because there may be some individuals for whom $u < g_i^*$ (and thus they are deterred) but for whom it is also true that $h < u$. This is an immediate consequence of the grouping of potentially heterogeneous acts in S_i. Second, as explained in the discussion of the first-order condition (3.9), it is also true that g_i^* could be very high, even above expected harm, because deterring first-best acts has two possible benefits: Reducing disutility from individuals actually experiencing guilt and relaxing the constraint on the use of guilt.

Part (d) is true because (3.9) is the first-order condition, which is a necessary condition for $g^* > 0$ to be optimal.

4. Moral Rules Enforced by Guilt and Virtue

4.1. Specific moral rules. — Returning to the case in which each situation (u,h) can be treated separately from any other, we now assume that both guilt and virtue may be instilled independently for each situation (u,h). Let $\$(v(u,h))$ denote the cost of inculcating virtue v(u,h) for situation (u,h), where $\$(0) = 0$ and, for $v > 0$, $\$'(v) > 0$ and $\$''(v) \geq 0$.

The social problem is to assign guilt g(u,h) to acts (u,h) and virtue v(u,h) for not committing acts (u,h) so as to maximize social welfare, subject to the constraints that realized guilt not exceed G and that realized virtue not exceed V. As before, let A denote the set of situations in which acts are committed, so we have A={(u,h)OS*u-g(u,h)>v(u,h)}, and we now let N denote the set of situations in which individuals do not commit the act, that is, N={(u,h)OS*u-g(u,h)#v(u,h)}. (The only difference from before is that, here, individuals commit an act when u-g > v rather than when u-g > 0.) Accordingly, the problem to be solved can be stated as choosing the functions g(@,@) $ 0 and v(@,@) $ 0 to maximize morally inclusive social welfare

$$(4.1)\quad \int\int_A \big(u-h-g(u,h)\big)f(u,h)dudh \;+\; \int\int_N v(u,h)f(u,h)dudh$$

$$-\int\int_S \big(\alpha(g(u,h))+\beta(v(u,h))\big)f(u,h)dudh,$$

subject to the constraints that

$$(4.2)\quad \int\int_A g(u,h)f(u,h)dudh \;\le\; G, \text{ and}$$

$$(4.3)\quad \int\int_N v(u,h)f(u,h)dudh \;\le\; V.$$

The first term in expression (4.1) is identical to the first term in expression (3.1) for the case in which only guilt can be instilled. It represents the welfare effects for acts that are committed — act-utility and guilt are experienced and external harm is caused — weighted by the frequency of the situations. Expression (4.1) includes a second term that has no analogue in (3.1); it is the welfare effects of the experiencing of virtue for those who do not commit acts, weighted by the frequency of these situations. The third term in (4.1), inculcation costs, now includes the cost of inculcating virtue as well as that of inculcating guilt. Finally, in addition to the constraint (4.2) on the realization of guilt (which is identical to the constraint (3.2)), we now have a corresponding constraint (4.3) on the experiencing of virtue.

The Lagrangian for the foregoing problem is

$$(4.4)\quad \int\int_{A}\big(u-h-g(u,h)\big)f(u,h)\,du\,dh \;+\; \int\int_{N} v(u,h)f(u,h)\,du\,dh$$

$$-\int\int_{S}\big(\alpha(g(u,h))+\beta(v(u,h))\big)f(u,h)\,du\,dh$$

$$-\lambda\left[\int\int_{A} g(u,h)f(u,h)\,du\,dh \;-\; G\right] \;-\; \mu\left[\int\int_{N} v(u,h)f(u,h)\,du\,dh \;-\; V\right],$$

where, as before, λ is the Lagrange multiplier for the constraint on the use of guilt and where μ is the Lagrange multiplier for the constraint on the use of virtue.

Before discussing the solution to this problem, let us make some general observations about it. First, virtue is not only an additional incentive, and possibly of value for that reason; it is also a potential substitute for guilt. Individuals are deterred from committing the act when $u \# g(u,h) + v(u,h)$, so, as far as behavior is concerned, guilt and virtue can be used interchangeably. Now, if the only further consideration were the marginal costs of inculcating guilt and virtue, our results would be essentially the same as they were in subsection 3.1, except that guilt and virtue would be used together, in a manner that minimized inculcation costs, to control (only undesirable) acts that are worth controlling. The problem, however, is more complicated.

The reason is that guilt and virtue differ regarding whether they are experienced when optimally employed. Recall that when guilt is successfully used as an incentive, it does not directly affect utility — it produces disutility if it is experienced, but when optimally employed in this case, it is not experienced. In contrast, when virtue is successfully used as an incentive, it directly contributes to utility — when virtue is successful in controlling behavior, it *is* experienced, unlike in the case of guilt. This suggests that there is an advantage of virtue as an incentive device. This also means, perhaps surprisingly, that virtue could be desirable to instill purely as a means of producing utility, if it is cheap enough to instill and if it is not needed as an incentive elsewhere. Yet when virtue is used as an incentive, just because it is enjoyed, it reduces the stock of virtue available to be used to control other behavior, whereas when

guilt is employed successfully as an incentive, it is not realized, so does not deplete the stock of guilt.[20] This latter factor works against virtue and in favor of guilt as an incentive device. From these remarks, one can see that determining the optimal use of virtue and guilt will be somewhat complex.

Allowing g*(u,h) to denote the optimal g(,) as before, and letting v*(u,h) denote the optimal v(,), we now state the following results, which are proved in the Appendix.

Proposition 3. Assume that guilt and virtue can be instilled separately for each act. Assume further that, at the optimum, $(0) > (1-:)f(u,h) for an act (u,h). Then, for that act:[21]

a. positive guilt or virtue is instilled only if not acting is first best, and, when guilt or virtue is instilled, the sum of guilt and virtue equals the minimum necessary to discourage the act; that is, if g(u,h) > 0 or v*(u,h) > 0, then u < h and g*(u,h) + v*(u,h) = u;*

b. guilt is never actually experienced, and virtue is always experienced whenever v(u,h) > 0 and situation (u,h) arises;*

c. the only possible deviation from first-best behavior is the commission of undesirable acts; and

d. it is optimal to instill guilt or virtue with respect to a situation in which u > 0 if and only if, at the g(u,h) and v(u,h) that maximize the Lagrangian (4.4) conditional on g(u,h) + v(u,h) = u,

$$(4.5)\quad \alpha(g(u,h)) + \beta(v(u,h)) < \langle h + (1-\mu)v(u,h) - u\rangle f(u,h).$$

Proposition 4. In general (that is, without restrictions on $ or on the value of : at the optimum), none of the results of proposition 3 necessarily hold except (b), that guilt is never experienced, and virtue is always experienced whenever v(u,h) > 0 and situation (u,h) arises.*

Notes: We begin by elaborating on the meaning of the assumption in proposition 3 that, at the optimum, $(0) > (1-:)f(u,h). (We comment briefly on the plausibility of this assumption in subsection 5.3.) This restriction is a sufficient condition for virtue never to be employed merely for the benefit of individuals' enjoying the experience of virtue; that is, the condition guarantees that virtue will be used only when it helps to improve behavior.

The left side of this condition is the marginal inculcation cost when virtue is equal to zero. Clearly, if this cost is sufficiently high, virtue will be used only when it can improve behavior. In particular, note that $(0) > f(u,h) is sufficient, which is to say that, for acts that are not very likely to

---

[20]Thus, as in subsection 3.1, the constraint on the use of guilt, here (4.2), is never binding.

[21]As with proposition 1 (see note 15), this result need not hold with respect to any set of acts of measure zero.

arise, it will never be optimal to use virtue just for the sake of the utility benefit of experiencing virtue. When virtue is inculcated for a situation (u,h), the situation in which it may be experienced arises with frequency f(u,h), and the maximum benefit in that situation is one — the benefit is less, 1-∶, if the virtue constraint is binding, in which case the shadow price ∶ of using virtue is positive. Observe that the inculcation cost is a fixed cost of sorts, whereas all benefits (from controlling behavior or from simply having individuals feel virtuous) arise only if the situation arises (and, moreover, individuals do not commit the act).

The other pertinent consideration is the constraint on the use of virtue. As this constraint becomes tighter, it is more likely that it is not optimal to use virtue unless it helps to control behavior. (As noted, 1-∶ is the benefit of experiencing virtue less the shadow price of using virtue in light of the constraint on its use.) The intuition is that, when virtue is scarce, using more virtue to control one act can never increase the total of virtue experienced, for to satisfy the constraint it will be necessary to use less virtue elsewhere. It will be optimal to allocate scarce virtue across acts so that each unit of virtue used produces the maximum possible benefit. Since the utility benefit of experiencing virtue is the same across acts, virtue will be concentrated where the behavioral benefit is relatively greater (and also where the marginal inculcation cost is relatively lower). A case of particular interest is that in which the virtue constraint is significantly binding, namely, when ∶ > 1. It is apparent that in this case the assumption in proposition 3 is always satisfied, regardless of how low is the inculcation cost or how high is the frequency of the act.

Continuing with proposition 3, it is clear that the results are very similar to those in proposition 1, where only guilt could be employed. Regarding part (a), when virtue is sufficiently costly to use, virtue, like guilt, will only be used when acts are undesirable (u < h), and when they can successfully be deterred, and no excess virtue will be used to deter acts. Part (b) is analogous to proposition 1, except that virtue, when used, is experienced. The reason is that part (a) indicates that virtue is used only when it deters acts, but since not committing an undesirable act is the choice that does result in virtue being felt by the individual (in contrast to guilt, which is associated with committing the act, which is here taken to be deterred), virtue is experienced, unlike guilt. Part (c), as before, follows from part (a). Finally, the condition in part (d) now reflects both that there is a cost of instilling virtue, as well as of instilling guilt, and also that virtue is experienced when used, thus producing a net benefit if ∶ < 1 and a further cost if ∶ > 1.

Proposition 4 is explained by the possibility that virtue, if sufficiently cheap to inculcate and if not very scarce, may be optimally employed simply so that individuals can benefit from experiencing virtue. To see the intuition, consider an extreme case, in which there is no undesirable behavior to control (u > h for all situations in S), the marginal inculcation cost $ℕ(0) is very low, and V is large. Here, particularly for acts that are only marginally desirable, it may well be optimal to inculcate virtue to induce individuals not to commit acts. This would violate (a) of proposition 3 because virtue would be instilled even though not acting is not first best. Moreover, if the marginal inculcation cost was low, one might employ more virtue than necessary to induce individuals not to commit the act. With regard to (b) of proposition 3, it is still not optimal for guilt ever to be experienced (for, as before, when it is, the

act is not deterred, so instilling guilt involves costs and no benefits). In addition, that virtue may be cheap to instill and not very constrained still does not make it worthwhile to instill virtue if the act will be committed in any event, for in that case the virtue is not actually experienced. For (c) of proposition 3, the foregoing discussion illustrates the possibility that one would want to induce individuals to abstain from committing desirable acts, in order that they may thereby experience virtue. Finally, for (d), condition (4.5) need not hold because it is possible that virtue should be inculcated even when the act would not be committed in any event; moreover, when virtue is used to deter an act, it is no longer assured that no more virtue than necessary will be used.[22]

4.2. General moral rules. — Now let us assume, as in subsection 3.2, that guilt and virtue cannot be inculcated independently for each possible act (u,h). Instead, guilt and virtue are constrained to be the same for all acts within each of n subsets S_i that partition the universe S of situations. Let g_i and v_i denote the uniform levels of guilt and virtue for acts within S_i.

Given these assumptions, the social problem is analogous to that in subsection 3.2: Choose $g_i \geq 0$ and $v_i \geq 0$ on each subset S_i to maximize welfare, subject to the constraints on the realization of guilt and virtue. As before, 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 within S_i. Let $\theta_i(g_i)$ and $\psi_i(v_i)$ denote the costs of instilling guilt g_i and virtue v_i for acts within subset S_i, and assume that the derivatives of $\theta_i(g_i)$ and $\psi_i(v_i)$ have the same properties as those of $\theta(g)$ and $\psi(v)$. Then morally inclusive social welfare is

$$(4.6) \quad \sum_{i=1}^{n} W_i(g_i, v_i),$$

where

[22]When virtue is cheap to inculcate and not very scarce, it is possible that it would be optimal to inculcate virtue for committing an act that causes external harm (a possibility that we rule out earlier, by assumption). This could arise if it was cheaper to inculcate virtue for the more harmful choice or if most would engage in the more harmful choice, so that virtue would be experienced by more individuals if it were thus inculcated.

$$(4.7)\quad W_i(g_i,v_i) = p_i\left[\int_0^{\infty}\int_{g_i+v_i}^{\infty}\left(u-h-g_i\right)f_i(u,h)\,du\,dh+\int_0^{\infty}\int_0^{g_i+v_i}v_i f_i(u,h)\,du\,dh\right]$$

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

To explain, individuals commit acts in set S_i when $u - g_i > v_i$, or $u > g_i+v_i$, in which case the effect on social welfare is $u - h - g_i$ (as in the explanation of expression (3.5)); 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 actual realization of guilt and virtue are

$$(4.8)\quad \sum_{i=1}^{n} y_i(g_i,v_i) \le G \text{ and}$$

$$(4.9)\quad \sum_{i=1}^{n} z_i(g_i,v_i) \le V,$$

where

$$(4.10)\quad y_i(g_i,v_i) = p_i\int_0^{\infty}\int_{g_i+v_i}^{\infty} g_i f_i(u,h)\,du\,dh = p_i g_i\left(1-F_i(g_i+v_i)\right) \text{ and}$$

$$(4.11)\quad z_i(g_i,v_i) = p_i\int_0^{\infty}\int_0^{g_i+v_i} v_i f_i(u,h)\,du\,dh = p_i v_i F_i(g_i+v_i).$$

The Lagrangian for the problem of maximizing welfare (4.6) subject to the constraints (4.8) and (4.9) is

$$(4.12)\quad \sum_{i=1}^{n} W_i(g_i,v_i) \;-\; \lambda\left[\sum_{i=1}^{n} y_i(g_i,v_i) \;-\; G\right] \;-\; \mu\left[\sum_{i=1}^{n} z_i(g_i,v_i) \;-\; V\right].$$

The first-order condition if $g_i^* > 0$ is[23]

$$(4.13)\quad p_i\left[\int_0^{\infty}\left(h+\lambda g_i-\mu v_i\right)f_i(g_i+v_i,h)\,dh \;-\; (1+\lambda)\left(1-F_i(g_i+v_i)\right)\right] \;=\; \alpha_i'(g_i),$$

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

$$(4.14)\quad p_i\left[\int_0^{\infty}\left(h+\lambda g_i-\mu v_i\right)f_i(g_i+v_i,h)\,dh \;+\; (1-\mu)F_i(g_i+v_i)\right] \;=\; \beta_i'(v_i).$$

The first-order condition for g_i^* (4.13) is the same as that in subsection 3.2 (expression (3.9)), except that：v_i is subtracted in the integrand, reflecting the fact that, when individuals are deterred from committing acts in S_i, they now experience v_i, which has a shadow cost if the constraint on the use of

[23]As in subsection 3.2, a corner solution (at 0) is possible and the first-order conditions are necessary but not sufficient for a global optimum.

virtue is binding. In addition, the marginal individual now has act-utility of g_i+v_i, as explained above. The first-order condition for v_i^* (4.14) is analogous to that for g_i^*. The difference is that, in the second term, the inframarginal effect, there is a utility gain for those not committing the act, for they experience more virtue when v_i is increased, but the shadow price of the virtue constraint ： must be deducted from this benefit.

In summary, for both guilt and virtue, there are two types of effects. There are marginal effects, consisting of reduction of the externality and benefits or costs associated with the constraints (when an individual is deterred, there is a benefit from using less guilt and a cost from using more virtue). There are also inframarginal effects with regard to those whose behavior is unchanged; those who continue to commit the act experience more guilt and those who continue to be deterred experience more virtue, as the case may be. (For both first-order conditions, the direct utility effects on the marginal individuals equal zero, for the marginal individuals are those for whom $u = g_i+v_i$.) The sum of these two effects, the marginal (or deterrence) effect and the inframarginal effect, are equated with the direct marginal costs of instilling a higher level of guilt or virtue.

The interpretation of (4.13) is essentially the same as that of (3.9). As noted, the only real difference is that the marginal deterrence benefit is now lower, because individuals who are deterred experience virtue, which is costly when the constraint is binding. (Also, the previous observation that g_i^* may exceed expected harm now becomes the point that $g_i^*+v_i^*$ may exceed expected harm.)

With regard to the interpretation of the first-order condition for virtue (4.14), we note, as stated, that the second term may be positive (it will be so if, at the optimum, the constraint on virtue is not binding or if it is binding but ： < 1), indicating, as in subsection 4.1, that there can be a direct benefit — independent of controlling behavior — from the use of virtue. However, if we confine attention to cases in which the constraint is sufficiently binding (： > 1 at the optimum), virtue will be used only when it is valuable in controlling behavior.

Comparing (4.13) and (4.14), it is possible to make a statement about whether it is optimal to primarily (or exclusively) rely on guilt or on virtue in controlling behavior in a set S_i. The marginal benefits of using guilt and virtue (the first terms on the left sides of (4.13) and (4.14)) are identical, reflecting the fact that they are interchangeable as deterrents. The marginal inculcation costs (the right sides of (4.13) and (4.14)) are symmetric, so this consideration favors using whichever moral sanction/reward has the lower marginal inculcation cost. At this point, there is no qualitative difference between the desirability of virtue and guilt as incentives.

However, when we consider the inframarginal effects (the second terms on the left sides of (4.13) and (4.14)), one does see a qualitative difference. In what we take as our benchmark, the case in which ： > 1 at the optimum, both second terms are negative, indicating that greater actual use of both guilt and virtue is costly. One difference is in costs per unit used, 1+8 for guilt and ：-1 for virtue. The other difference concerns how much is used, the fraction $1-F_i(g_i+v_i)$ for guilt and the fraction $F_i(g_i+v_i)$ for virtue. 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 used, whereas a significant amount of virtue will be used (each per unit inculcated). Accordingly, when most acts will be deterred, it will tend on this account to be optimal to use guilt and not virtue, ceteris paribus. Likewise, when few individuals will be deterred from committing acts in S_i, so F_i is small, it will tend to be optimal to use virtue and not guilt.[24] And, because the effect of raising g_i or v_i on inframarginal costs can 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.

A final implication of the optimization process concerns the use of guilt and virtue across sets of acts, in the cases in which the constraints on the use of guilt or virtue are binding. These conditions can be stated as

$$(4.15) \quad \frac{\partial W_i(g_i,v_i)/\partial g_i}{\partial y_i(g_i,v_i)/\partial g_i} = \lambda, \textit{ for all } i \textit{ for which } g_i > 0, \textit{ and}$$

$$(4.16) \quad \frac{\partial W_i(g_i,v_i)/\partial v_i}{\partial z_i(g_i,v_i)/\partial v_i} = \mu, \textit{ for all } i \textit{ for which } v_i > 0.$$

As we discuss in subsection 3.2, these conditions require that the marginal welfare benefit per unit of guilt or virtue that will be experienced be equal for all sets of acts where guilt or virtue is used. As noted, the denominators refer not to the marginal unit of guilt or virtue that is inculcated, g_i or v_i itself, but rather to the effect of inculcating another unit of guilt or virtue on the amount of guilt or virtue that is expected to be experienced, $y_i(g_i,v_i)$ or $z_i(g_i,v_i)$. And, from our analysis of expressions (4.13) and (4.14), it is clear that inculcating additional guilt or virtue could raise or lower the amount of guilt or virtue that will actually be experienced.

Let us now state our conclusions, which are proved in the Appendix, for the case in which both guilt and virtue may be used to enforce general moral rules.

[24]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.

Proposition 5. Assume that guilt and virtue can only be instilled in a uniform manner within each of the subsets S_i that partition S. Assume further that, at the optimum, $\$_i\mathbb{N}(0) > (1-:)p_i$, for a given i. Then, within that subset S_i:

a. positive guilt and/or positive virtue are instilled only if not acting is first best for some acts in S_i; that is, if $g_i^* > 0$ or $v_i^* > 0$, then $u < h$ for some $(u,h) \mathbf{O} S_i$;

b. 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;

c. both types of deviations from first-best behavior are possible: the commission of undesirable acts and the deterrence of desirable acts; and

d. if $g_i^* > 0$, g_i^* satisfies (4.13); if $v_i^* > 0$, v_i^* satisfies (4.14).

Proposition 6. In general (that is, without restrictions on $\$$ or on the value of : at the optimum), result (a) of proposition 5 need not hold.

Notes: The reasoning behind proposition 5 parallels that in subsection 3.2 for proposition 2, concerning the use of guilt alone. For part (a), the assumption is sufficient to indicate that virtue will never be used except when there is a benefit of controlling behavior. For part (b), due to the grouping of acts, just as guilt may sometimes be experienced (that is, guilt may be useful even if it is not effective for every act in S_i), so virtue may not always be experienced (because it may be useful even if it is not effective for every act in S_i). Parts (c) and (d) are as before.

Proposition 6 indicates that, if virtue is sufficiently cheap to instill and if the constraint on the use of virtue is not very binding, it may sometimes be desirable to instill virtue purely for the benefit of experiencing it, which could even include a set of acts in which acting is always desirable. (See the discussion of proposition 4 in subsection 4.1.)

5. Interpretation, Extensions, and Discussion

In this section, we draw on our model in an attempt to explain the observed use of guilt and virtue in the enforcement of moral rules, discuss the basis for a number of our assumptions, and address further issues that we believe are illuminated by our analysis. We acknowledge at the outset that many of our remarks are speculative in nature.

5.1. The use of guilt and virtue. — Our results indicate how guilt and virtue would be used to enforce moral rules if, in fact, moral rules were designed to maximize social welfare. Some implications are straightforward and seem in accordance with what we observe. Notably, guilt and virtue are employed to prevent acts that typically reduce welfare (lying, breaking promises, harming others) and to encourage acts that benefit others (rescuing someone in distress), and the magnitude of guilt and virtue will tend to be correlated with the size of the negative or positive effect of the acts on others. Of course, it is not always true that a given type of act is invariably desirable or undesirable; for example, some lies may be desirable. Nevertheless, it has long been observed by philosophers and others that we still may feel guilty when we tell such a lie. This too is implied by our analysis because guilt and

virtue often must be inculcated for categories of acts, fine-tuning being costly or impossible.

Our analysis also has some less straightforward implications. The most significant of these concern the fact that guilt and virtue will actually be experienced by individuals who commit acts with which guilt and virtue are associated. In particular, if guilt and virtue will succeed in inducing most individuals to act in a socially desirable manner, we would expect to see guilt used, for there will be little social cost of guilt being experienced (whether in terms of disutility from feeling guilty or in terms of drawing excessively on a limited pool of potentially available guilt). In contrast, if virtue were used in such situations, there would be a substantial draw on the limited pool of potentially available virtue, leaving much less to control other types of behavior.[25] Similarly, when few individuals can be induced to act optimally, we would expect to see virtue used. This would draw little on the pool of potential virtue, whereas the use of guilt would result in many suffering disutility and would draw heavily from the pool of guilt.

These ideas do seem to be in accord with the observed use of guilt and virtue. On one hand, individuals who do a range of undesirable acts, from cutting in line to physically assaulting those with whom they have disagreements, generally feel guilty, and indeed those are acts that most individuals are successfully deterred from committing most of the time. But individuals do not, it seems, feel especially virtuous when they abstain from such acts, since it is expected that everyone will do so.[26] On the other hand, individuals who rescue others at great personal sacrifice and those who devote their lives, say, to helping the poor in less developed countries, feel virtuous, whereas the substantial majority of us who do not routinely give most of our time or resources to helping strangers (and could not readily be induced to do so) do not generally feel terribly guilty. An implication related to the foregoing is that we do not ordinarily see significant use of both guilt and virtue with regard to the same decision, which also is suggested by our analysis. In all, it appears that our model helps to understand why guilt may be associated with some acts and virtue with others, a distinction that does not readily seem capable of explanation on other grounds.[27]

---

[25]In making this statement, we are assuming that the constraint on the pool of available virtue is binding (and has a significant shadow price), which seems plausible for the reasons we elaborate in subsection 5.3.

[26]There are, of course exceptions. For example, one who abstains despite unusual provocation may feel virtuous or be subject to approbation from others. But, interestingly, this is precisely a type of situation in which most individuals would not act optimally, so this apparent exception is itself consistent with our analysis. As an illustration, there is often an exception to the moral injunction against aggression for cases of self defense. This rule — in addition to the benefit of the prospect of retaliation in deterring aggression — has the additional advantage that, since most individuals will not be able to restrain themselves in certain settings, a needless use of guilt is avoided.

[27]The simplest alternative theory would be that significant virtue is associated with most (all) good choices among actions and significant guilt with most (all) bad choices. But if this were generally the case, then abstaining from bad behavior in everyday situations in which almost everyone would abstain would be associated with individuals' feeling highly virtuous, and failing to behave in a manner that raises total welfare at significant personal sacrifice when most others would similarly fail to do so would be associated with substantial guilt. Neither seems

Additionally, we note that guilt seems more often to be on people's minds than is virtue. To explain this, we observe initially that, because both guilt and virtue are actually experienced only when individuals behave atypically, it is the contemplation of acts that are not ultimately committed that is most relevant here. And, with regard to contemplated acts, guilt discourages acts from which individuals would obtain direct utility and thus would otherwise commit; because such temptation may be frequent (for example, one may often be tempted to lie), the prospect of guilt would often come to mind. In contrast, virtue encourages acts that individuals would not otherwise commit — acts from which direct utility would be negative (such as acts involving self-sacrifice to aid others); because most individuals may not frequently be inclined to contemplate committing such acts, virtuous feelings would seldom be pondered.[28]

Another implication is that, the greater is the density of individuals whose utility from committing acts is close to the level of guilt or virtue being employed, the greater is the benefit of raising the level of guilt or virtue in inducing more individuals to behave in a socially desirable manner. This implies, for example, that behaviors that are more automatic, less conscious, will be less subject to the use of moral emotions, which also seems to be true.[29] (An exception would be where behavior is automatic due to habit formation, when following one's moral emotions is part of what produced the habit in the first instance.)

Furthermore, our analysis may help to explain variations in moral rules across cultures, over time, and within societies (comparing different groups).[30] Such variation has long been recognized and is said to pose difficulties for some normative theories. As a positive matter, however, our analysis suggests that, even if moral rules in different settings were optimal, their content — whether and the extent to which there would be guilt and virtue associated with assorted types of behavior — may well differ. Methods of inculcation as well as the identity and thus the interests of inculcators will vary (role of organized religion, form of government, existence of formalized education, mobility across communities), as will the relative frequency of different types of situations and the harm and utility associated with acts in those situations. And, even if two types of behavior were identical in frequency, harm, and utility in two settings, optimal moral rules may nonetheless differ, for example, if the guilt or

---

generally to be true.

[28]Another explanation is that the cost of inculcating virtue may be much higher than that of inculcating guilt or that the constraint on the use of virtue may be much tighter than that on using guilt, so there is simply less virtue employed in enforcing moral rules.

[29]For example, the insane are understood to be exempt from most moral sanctions, as are epileptics in certain situations. A more complicated case involves the treatment of children. On one hand, due to their less developed ability to conform their behavior to moral rules, we tend to excuse them. On the other hand, one cannot wholly refrain from applying moral rules if one is hoping to inculcate the rules. An implication is that one should defer the use of moral sanctions, especially guilt, until children reach an age where there is a reasonable prospect of achieving success over a period of time that is not unduly prolonged.

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

virtue constraints are more tightly binding in one society due to the greater need to use the moral emotions to regulate some other type of behavior. That a consequentialist moral system is consistent with variations in moral rules across societies is, of course, not a novel suggestion, for it is recognized that consequentialist morality is contingent upon the circumstances under consideration.

It would be difficult to pursue the foregoing descriptive claims for two reasons. First, the ability to measure the relevant phenomena is limited and further speculation at this stage in the investigation of the subject seems premature. Second, a number of complications discussed in some of the subsections to follow also bear on the observed use of guilt and virtue. Most important is that the optimality of actual moral rules is hardly assured and that the rules that tend to emerge may promote an objective that is different from social welfare.

5.2. Evolution and inculcation. — Both evolution and inculcation (nature and nurture) seem to play important roles in determining the use of guilt and virtue in enforcing moral rules.[31] 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 — obviously has an evolutionary origin, just as does any other capacity we might have. See, for example, Darwin (1874), E.O. Wilson (1975), and Izard (1991).[32] 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.

We observe that our capacity to feel guilt and virtue is a flexible one. Such flexibility would seem to confer an evolutionary advantage because it allows adaptation to changed circumstances. In addition, flexibility is certainly 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 even 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

---

[31]Many of the ideas discussed in this subsection are developed in the literature cited in note 3. The interaction of evolution and inculcation in determining behavior is further explored in Barash (1982) and Tooby and Cosmides (1990).

[32]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 that correspond to different emotions are universal in humans and evident in some other species and hence must have an evolutionary origin (which seems necessarily to imply that the emotions being expressed must too have an evolutionary origin). Thus, although critics of sociobiology, such as Lewontin, Rose, and Kamin (1984), would give a heavier role to cultural determinants in most domains, it seems difficult to deny that biology has an important role at least in explaining the existence of features of the human brain that enable us to experience moral emotions.

over time (which seems greatly to exceed the rate of biological evolution).[33]

It seems plausible that some of our more particular feelings of guilt and virtue are not entirely the product of inculcation. Consider, for example, the guilt that we associate with stealing. No doubt society instills guilt in this case, but it also seems possible that some of the guilt we feel is a product of evolution in the biological sense. A hard-wired reluctance to steal may well help to overcome acquisitive urges that, if acted upon, would be met with retaliation, which can prove very costly to the aggressor.

The extent to which moral rules and associated feelings of guilt and virtue are inculcated rather than purely evolved has normative and positive implications. Regarding normative implications, to the extent that morality can be instilled, society can attempt consciously to design policy to adjust our moral system to maximize social welfare, whereas if morality were essentially hardwired, little could be done.

Regarding positive implications, we offer two comments. The first concerns what is being 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.[34] In examining various questions — such as how bad it is to tell a lie or how good it is to help others in particular circumstances — it seems clear that the answers may be different if the goal is different. 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. On the other hand, 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 and thus be more adaptive to modern circumstances.

Second, the tendency for moral systems to be optimal — with reference to whatever is being maximized — may differ depending upon the relative importance of evolution and inculcation. Neither process assures optimal results. With evolution, there is the familiar point that selection is fundamentally at the level of individual genes, so, for example, traits that would benefit a group as a whole do not tend to emerge (although they may to some extent through kin selection, reciprocity, and so forth). Also,

---

[33]See, for example, Nisbett and Cohen (1996), who identify cultural differences regarding what they refer to as a "culture of honor" between inhabitants of northern and southern states. Izard (1991) elaborates the view that the capacity to experience guilt and shame has an evolutionary origin but that the association between these emotions and particular acts is produced by internalization as a consequence of parental activity and other forms of social learning and thus varies across societies. See also Tangney and Fischer (1995).

[34]There are, of course, limits to the latter, in that societies less successful in ensuring survival, especially in competition with other societies, will tend to die out.

with evolution, there must be a feasible path for a desirable trait to emerge. 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. Likewise, when there are multiple inculcators, each may impose externalities on others through excessive use of the scarce capacity to experience guilt and virtue, a sort of common pool problem.

In most of our discussion, we take the view that guilt and virtue are to a substantial extent the product of inculcation and thus may be regarded as welfare-maximizing (although there remains the question of which groups' welfare is likely to be maximized). To a degree, the moral system that we observe seems consistent with this view. There do, however, seem to be important aspects that may well have evolutionary explanations. Notably, there is a strong tendency to limit altruism, and related feelings of virtue, to one's kin. (Here, the explanation may be mixed, for even if moral emotions are largely a product of inculcation, one's parents and other relatives may have a disproportionate influence on the inculcation process.)

Nevertheless, the capacities for experiencing guilt and virtue are, as noted, to an important extent the product of evolution and thus are not designed to maximize welfare. Notably, from a welfare-maximizing view, a large capacity for feelings of virtue would be highly desirable: Not only could one use the large reservoir of virtue better to control behavior, but there is also the direct utility benefit from experiencing virtue. Indeed, would it not be a wonderful world if we could all feel incredibly virtuous every time we did not cut in line or each instance in which we refrained from punching someone who was rude? As a matter of survival, however, virtue and guilt may be equally useful, for all that matters is controlling behavior, not whether it is controlled by the prospect of rewards or of punishments.[35] Moreover, if there are limits on the extent to which capacities for moral emotions can be developed, then having a modest pool each for guilt and for virtue rather than, say, only guilt or only virtue (with a total pool of equal size), has the benefit previously described: When individuals can usually be induced to refrain from an undesirable type of act, then little guilt needs to be used to accomplish this (the threat suffices for most); but when individuals cannot usually be induced to commit a good type of act, virtue is superior because much less of it needs to be used.[36]

5.3. Inculcation costs and constraints on the use of guilt and virtue. — In our model, we assume that there is an increasing marginal cost of inculcating guilt or virtue for a particular set of acts.

---

[35]If guilt were so extensive and so often experienced that the level of emotional pain induced individuals to commit suicide, the situation would be different, but, short of that, there seems to be little difference between guilt and virtue in this respect.

[36]Other factors would seem to have an evolutionary explanation. Notably, guilt and virtue are part of a larger system of emotions that serves many functions; moral emotions are plausibly an application of this more general system and thus would have its attributes even if they might not be ideal for the task of enforcing moral rules. For example, there may be limits on the extent of our emotions, and extreme emotions might be reserved for acts more directly related to our survival (reproduction, caring for offspring, self-protection).

For inculcation, the explanation for the existence of a cost is straightforward: It presumably takes time and effort to inculcate guilt or virtue. Our supposition that the required investment is subject to diminishing returns is more speculative, but seems plausible. Regardless of this particular feature, we note that the inculcation cost technology could take a number of forms.[37] (For example, it may be the total of guilt and virtue inculcated that determines the cost, without regard to how much of each is used, as we assumed; or more frequently occurring acts might have lower costs of inculcation because there are more learning opportunities.) Although we focus on inculcation, it is also the case that the effects of our increasing marginal cost assumption can be motivated by evolutionary considerations, because there is in a sense a 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.

One particular aspect of inculcation costs that we did not model is that there may be scarcity or crowding out across sets of acts; that is, it may well be that the more time one spends inculcating guilt or virtue for some types of acts, the less time will be available or the less effective the time will be 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 use of guilt and virtue, we do not believe that incorporating this form of interaction in the cost of inculcation would have changed our results significantly.[38]

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.[39] All of this seems plausible to us, as a matter of

---

[37]In fact, we only use the assumption of diminishing returns in propositions 3 and 5, where we refer to $N(0) in stating sufficient conditions for the stated results. If we did not assume that $O(0) $ 0, we could have stated a more cumbersome sufficient condition regarding $.

[38]The main effect of using a technology under which increasing the use of guilt or virtue for one set of acts raises the marginal cost of using guilt or virtue for all other sets of acts would have been to make it optimal to use less guilt and virtue. But since we do not specify how high is the marginal cost of instilling guilt and virtue for each act or set 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 set of acts, the analysis would be much the same.

[39]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. Since the guilt constraint did not play as significant a role (qualitatively) as the virtue constraint and since, with many sets of acts, this seems a modest consideration (one may simply need to stop short of the guilt constraint by enough to leave room to deter the marginal act), we do

introspection and observation, though we are not yet aware of what research may exist that would allow a more precise statement of the phenomenon in the present context.[40]

We note that a more accurate model of this feature of human nature would not, as ours does, employ a simple constraint on the total amount of guilt or virtue that can be experienced. Rather, it seems more plausible to assume that, as the total amount of experienced guilt or virtue increases, its marginal effectiveness diminishes.[41] If one takes this approach, the primary effect on the analysis would be as follows: where in our existing model (such as in first-order conditions (4.13) and (4.14)) there appear the shadow prices of the constraints (which increase as the constraints become more binding), one now would have 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). Under such a formulation, it seems that similar qualitative conclusions could be obtained. 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 affect welfare (negatively and positively, respectively) would have a more clearly identifiable effect on the optimum: ceteris paribus, this consideration favors using somewhat less guilt and more virtue than otherwise.

We also 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 of any such emotion one could feel. Had we included such a limitation in our model, the results would not be greatly affected because we

---

not pursue this complication further. Nevertheless, 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.

[40]The present discussion motivates the constraints on the total extent to which guilt and virtue can be experienced by reference to our internal capacity to experience any emotions. In subsection 5.4, where we explicitly introduce external sanctions and rewards, 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, a zero-sum phenomenon.

[41]Consider, for example, the following model. The term g_i continues to indicate how much guilt is inculcated for acts 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 set S_i, G would increase, which will decrease the effectiveness of guilt in controlling all behavior. (This model is more complicated than one in which there is simply a constraint. The first-order condition for the model is similar to (3.9), where the main difference is that described in the text to follow.)

already assume that there are increasing marginal inculcation costs on a per act basis.[42]

Finally, we remark on the plausibility of the assumption that, at the optimum, $\$_i\mathbb{N}(0) > (1-:)p_i$, which we used in proposition 5 (and an analogous assumption was used in proposition 3) to rule out the possibility that it could be optimal to use virtue solely so that individuals' utility will increase on account of experiencing it. As previously discussed, this assumption depends on marginal inculcation costs being sufficiently high or the constraint on the use of virtue being sufficiently binding. The former assumption seems plausible in that inculcating any moral lesson, even a minor one, seems difficult and time-consuming, so, for example, parents would not be likely to undertake the effort to teach their children that committing some act was virtuous solely so that the children could experience the pleasure of feeling virtuous from committing it. Moreover, the constraint on the use of virtue does seem substantially binding, for we (whether as parents or society as a whole) hardly are able to induce most individuals to do almost any good act whenever it would be desirable to do so. (To an extent, however, it is difficult to determine which of these assumptions, or combination thereof, explains what we observe, for either would be sufficient.)

5.4. Internal versus external sanctions and rewards. — Our model and most of our discussion refers to feelings of guilt and virtue, which are generally understood as internal punishments and rewards for following moral rules. There are, as noted in the introduction, corresponding external sanctions and rewards as well, disapprobation or blame, and approbation or praise.[43] Until now, we have loosely suggested that these external sanctions (hereinafter taken to include rewards) are encompassed by our analysis; hence, if guilt is optimally associated with a particular set of acts, so would disapprobation or blame.

Despite the similarities between internal and external sanctions, 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) and often of individuals with little or no direct relationship to the victim (beneficiary). 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.[44] The third element seems quite closely related to the internal sanctions and rewards

---

[42]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 act* and another (the pool constraint) that is *across acts*. Hence, our particular assumptions capture aspects of a number of plausible features that could have been included separately.

[43]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.

[44]There are also mixed or intermediate cases. For example, one might feel ashamed, and thus suffer a decline in utility, if others find out about one's act, without others having to engage in any particular behavior (such as

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 granted.[45] 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.[46] 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).

We could explicitly model disapprobation and approbation as follows. First, we would define such behavior as involving additional sets of acts, which themselves might have guilt or virtue (or other moral emotions, such as disgust or a sense of delight with regard to others' behavior) 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.[47] 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, guilt and virtue in this setting (or whatever moral emotion was employed) would not only affect behavior but would also sometimes be experienced, which itself would affect social welfare. And disapprobation and approbation would sometimes be experienced, which would affect the utility of those engaging in the initial behavior regulated by moral rules and would also involve 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

---

expressing disapprobation) in response to their learning about the act.

[45]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 5.6.)

[46]These phenomena need not, of course, be the same, and there are independent moral rules that govern expressing approval or disapproval of others' behavior, such as rules about when one should mind one's own business.

[47]Moreover, society might use external sanctions to enforce third parties' enforcement against primary behavior, and so forth. See, for example, Axelrod (1986) and Pettit (1990).

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.

In sum, although it would be an oversimplification simply to treat internal and external sanctions as if they were the same, there are important similarities in how they should be analyzed. We believe, therefore, that there is some basis for our preliminary conjecture that implications of our analysis for the use of guilt and virtue will often be suggestive with regard to disapprobation and approbation.

5.5. Grouping of acts. — Motivation. — In the second version of our model, we assume that certain acts are naturally (and exogenously) grouped into distinct sets, 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 set. Thus, one might inculcate guilt for telling lies, breaking promises, assaulting others, and so forth, each on a wholesale basis. It may not be easy or even possible to inculcate guilt separately for each of the infinitude of possible situations in which one might tell a lie. As a consequence, moral rules have the familiar characteristic that they sometimes seem to be in error. One might, for example, feel guilty for telling a lie even when telling the lie would be desirable in the particular, atypical situation at hand.

Although the view that acts fall naturally into distinct categories is an oversimplification, as we will elaborate in a moment, we first reflect on why there is an important element of truth to the view that acts tend to be clustered in the manner just described. Many of the reasons presumably have to do with the organization of our brain.[48] An important aspect of the phenomenon involves perception, for the actor 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. Likewise, there is, as noted, scarcity in the evolutionary process that limits how such systems — and particular moral rules, to the extent they are evolved rather than purely inculcated — 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.

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

Another important set of reasons that acts tend to be grouped for purposes of moral rules concerns the rules' application. 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.

Endogeneity of groupings; rules and exceptions. — In all, there are strong reasons that we are led to group acts in various ways. Independently of the extent to which we must inculcate guilt and virtue on a categorical basis, there are clear advantages of choosing to do so on account of inculcation costs: One would expect there to be significant economies involved in inculcating these emotions with respect to groups of similar acts. That is, it may be less costly, we suspect much 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 will justify the practice. Moreover, given that the optimality of inculcating guilt or virtue 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), 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.

Inevitably, the natural groupings of acts will not be ideally suited for the particular purposes of regulating externality-causing behavior. Of course, our minds have a good deal of flexibility and are susceptible to some forms of reprogramming. Hence, if some natural category for which we would like to inculcate guilt is overinclusive — perhaps an important subset of acts in the category is desirable or is difficult to deter — we might expend additional resources to inculcate an exception. That is, the boundaries of the sets of acts in our model could be made endogenous. Thus, self-defense in certain types of circumstances might be excepted from the prohibition on aggression.

One might suppose instead that society could simply inculcate guilt over a smaller set — aggression that is not in self-defense — rather than inculcating guilt over the broader set and then expending additional effort to inculcate an exception. Whether this is feasible, we submit, is largely exogenously determined; sometimes there will be a narrower natural set that is rather homogenous regarding the optimal level of guilt or virtue to instill, and sometimes the natural set will be broader and quite heterogenous in this regard. Other times, there may be a choice whether to inculcate guilt or virtue situation by situation (more realistically, small cluster by small cluster) or to inculcate over a larger

set. The larger set may not allow as precise a match of guilt and virtue to particular situations, but the scale economies realized through more wholesale inculcation may warrant the use of grosser classifications.[49]

Overlapping groups. — Another important feature of groups of situations not captured in our model is that the groups may overlap in reality. For example, there could be one group of acts — pushing another individual out of one's way — that is subject to guilt and another group of acts — aiding others in distress — that is subject to virtue. But this raises the question of what happens when one pushes someone out of one's way in order to help someone else who is in distress. One possibility is that we simply combine all sources of guilt and virtue in making our decision. Hence, our prospective rescuer may help the person in distress and thereby feel virtuous, but still feel guilty for having pushed someone out of the way.[50] 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. Another possibility is that certain emotions may trump or at least dull others, so perhaps our rescuer would not feel guilty after all under these circumstances. How such overlaps and conflicts are resolved is an empirical question about the nature of the categories in our minds and the manner in which our emotions actually function. To an extent, the outcome may also be socially determined, for society could choose to inculcate an exception to one or another moral rule in cases of conflict, and sometimes this seems to be done. Obviously, one could model overlapping categories using our basic approach, with qualitatively similar results. One of the main conclusions would be that, when categories overlap, guilt or virtue may be used even more often in situations in which that would be unnecessary or undesirable (compared to a setting in which specific moral rules were feasible).

We now consider one particular manner in which sets of situations may overlap: In addition to moral rules for particular types of acts (such as lying or stealing) there exist moral rules that apply very broadly, notably, the Golden Rule, which enjoins individuals always to take into account the effects of their behavior on others.[51] One can understand such a rule as associating guilt with all undesirable acts

---

[49]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. Thus, one 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 set.)

[50]Brandt (1996) and Ross (1930) suggest that, when individuals follow the stronger moral obligation, they nevertheless feel compunction about having neglected the weaker obligation.

[51]Similar analysis would apply to intermediate cases, such as a rule enjoining all acts that harm others or all acts that intentionally harm others.

and/or virtue with all desirable acts, perhaps with the level of guilt or virtue rising with the extent of negative or positive externality. It seems clear that such broad rules do exist, although it is equally clear that they exist along side the other sort of categorical rules we have been discussing thus far and not in lieu thereof. This raises the question of why society does not simply inculcate the Golden Rule or some variant, eschewing all other rules, and thereby enjoin all individuals always to act in a socially optimal manner. Reflecting on the factors we have previously discussed, there are good reasons why this is not how moral systems operate: 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).[52] 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. Thus, even with the Golden Rule in force, many individuals would still commit undesirable acts, which would consume the scarce pool of guilt, making it more difficult to control other acts that it may be more important to deter; likewise, if virtue were instilled for all good acts, virtue would quickly be consumed on routine good behavior, leaving little to encourage certain types of behavior that may be particularly valuable. In sum, broad rules like the Golden Rule, as a supplement to more specific (but still fairly broad) rules are likely to be valuable precisely because of their breadth (they may cover acts that fall in the gaps between other moral rules) and their flexibility (they are directly sensitive to the externalities in particular situations). Nevertheless, due to their limitations, they optimally would 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 throughout our discussion.

That groupings sometimes overlap seems important in explaining aspects of guilt and virtue that we observe. Notably, sometimes individuals do feel conflicted about what behavior is morally correct, and, moreover, conflicts often seem to arise in instances in which two or more moral rules plausibly apply in the same situation. Both the existence of groupings and their possible overlap is also highly relevant to philosophical assessments of morality, which often draw heavily on our moral instincts and intuitions for insight. Our discussion reinforces the suggestion of those who have advanced two-level moral theories, such as Hume (1739, 1751), Austin (1832), Mill (1861), Sidgwick (1907), and Hare (1981), that moral rules may well condemn or endorse acts that are not in themselves bad or good, respectively, because they fall into broader categories for which it is generally — that is, on average — true that the included acts are bad or good.[53] Likewise, many philosophical discussions are concerned with cases in which our moral intuitions seem to be in conflict, such as in cases in which one must inflict

---

[52]These and related factors have been emphasized by philosophers. See, for example, Smith (1790), Austin (1832), Brandt (1979, 1996), Hare (1981), Mackie (1985), and Sartorius (1972). In addition, Cosmides and Tooby (1994) suggest that the human mind is better at specialized than general problem solving, suggesting that we are more capable of properly applying rules targeted to particular contexts than a broad command like the Golden Rule.

[53]Psychologists have also 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).

harm on one individual in order to help others (who are greater in number or who are affected to a greater extent).[54] Many such discussions fail to acknowledge that, even without regard to overlap, the grouping of situations implies that some behavior will be subject to feelings of guilt or virtue even when, if the act were viewed in isolation, that would be inappropriate. Moreover, when there is overlap, it does not follow that whichever category seems to exert a stronger pull on our intuition — perhaps the category for which the absolute magnitude of the moral emotion is greater — is the one whose rule would lead to first-best behavior. For example, helping others may have little virtue attached to it because virtue is costly to instill, because few would in fact help others, or because typical instances of helping others do not involve nearly as great a benefit as would the particular act in question; but none of these reasons suggest that commission of the particular act, which may also be in another category to which guilt is assigned, would be socially undesirable.

5.6. 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.

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 acts would have to 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 for each individual.)[55]

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 the latter type of variation is substantial.) Thus, to the extent that one can speak of a social

---

[54]See, for example, the discussion in subsection 5.8 on the act/omission distinction and related doctrines. Note that if a command akin to the Golden Rule is inculcated, as discussed in the preceding paragraph, then such conflicts among our moral intuitions will arise whenever a categorical moral rule requires welfare-reducing behavior, a phenomenon that does seem to fit many philosophers' arguments regarding consequentialism.

[55]As noted in subsection 5.3, a more realistic way to think of the limitations on the capacity of individuals to experience guilt and virtue (but one which would not qualitatively change our analysis) is not that there is a literal limit, but rather that, as more guilt and virtue are experienced, the less is the impact that they have on utility and thus on behavior. Such a formulation would be more appropriate if modeling heterogenous individuals, since it allows for different individuals to experience different levels of guilt and virtue — which would be the primary qualitative difference between such a model and the one we analyze. (As previously discussed, however, to the extent that constraints on the use of guilt and virtue refer to the imposition of external sanctions and rewards by third parties, our model's use of a single, aggregate constraint may be more apt.)

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 acts 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. Individual heterogeneity combined with the clustering of acts helps to explain why guilt is sometimes experienced and why even modest levels of inculcated virtue will induce some individuals to do desirable acts that most individuals could not be induced to commit even by the prospect of great rewards.

Heterogeneity in the extent to which guilt and virtue are experienced helps to explain other features of observed behavior.[56] Clearly, there are many undesirable acts that very few individuals would commit, and we sometimes classify individuals who would commit such acts as psychopaths. One possibility is that these individuals have little capacity for feeling guilt. At the other extreme, there are a handful of individuals — such as Mother Theresa — who seem unusually willing to make significant personal sacrifices to help others. One might suppose that such individuals either experience less direct disutility from self-sacrifice or experience stronger feelings of virtue from committing such acts. Finally, we observe that heterogeneity, particularly with regard to different experiences of inculcation, helps to explain the moral disagreement that we observe among individuals in a given society.

5.7. Prudence. — Many acts involve no (or only trivial) 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

---

[56]Additionally, as suggested in note 45, 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). This, in turn, can explain certain signalling behavior and, relatedly, our tendency to make associational decisions based on what may otherwise seem to be irrelevant characteristics, such as whether a prospective business associate is philanthropic or sexually abuses subordinates. See, for example, Posner (2000).

Yet another possibility is that a model with heterogeneity could address how others' behavior influences an individual's susceptibility to the moral emotions. For example, it may be more difficult to inculcate or maintain the effectiveness of guilt for committing an act — as well as a social practice of expressing disapprobation — if too many other individuals commit the act.

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 so may feel guilty.

Can one offer a consequentialist explanation for the use of the moral sentiments for the regulation of self-regarding behavior? One possibility is that externalities are 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. Indeed, some level of general altruism may be supported by the moral sentiments themselves. Moreover, parents will feel altruistically toward their children and thus be motivated to use available means, including the inculcation of moral rules, to encourage more prudent behavior.

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. As Schelling (1984), Thaler and Shefrin (1981), and others have suggested, 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. Under such a formulation, the behavior of the present self does create an externality, and hence our analysis suggesting the potential benefits of employing guilt and virtue could be applied. As a consequence, we do not regard subjecting such personal choices to the same type of moral mechanisms used for activity affecting others as inconsistent with our analysis of moral rules. It remains, however, to consider the extent to which the actual association of moral emotions to matters of prudence is consistent with the implications of our model.

5.8. Relationship to literature on moral philosophy. — Right and wrong versus social optimality.[57] — Most twentieth-century moral philosophers do not view moral rules as a system that is supposed to maximize welfare, but rather tend to see moral rules as indicating which acts are intrinsically right or wrong. Indeed, such philosophers frequently offer examples in which our moral instincts and intuitions deem 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 utility. These scholars, along with some modern writers, advance what is now described as a two-level view of morality.[58] At the first (higher)

---

[57]For further elaboration, see Kaplow and Shavell (2002).

[58]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

level is the ultimate criterion of judgment, which for them is social utility. (This corresponds to the social welfare function in our model.) At the second (lower) level are the moral rules that are supposed to guide behavior. (These correspond to the sets of acts in our model, and the corresponding uniform levels of guilt and virtue applying to acts in each set.) Implicit in their analysis is what economists would recognize as a standard problem of constrained maximization. There are taken to be a limited number of moral rules that might be chosen (the contours corresponding, roughly, to the categories of acts in our model), and the challenge is to select those rules that maximize welfare.

Because the choices are limited — due to given facts of human nature — this problem is of a second-best character. Accordingly, one does not expect any of the rules 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; likewise, some "right" acts may be welfare-reducing. Relatedly, under these two-level theories, unlike under many contemporary moral theories, blame and praise (externally administered analogues to guilt and virtue, which receive less attention) are viewed instrumentally: Whether an act should be deemed blameworthy, it is argued, should depend not on intrinsic features of the act or even on whether the act produces undesirable consequences, but rather on whether the practice of blaming those who commit the act will itself promote welfare. Thus, as noted, blame might be associated with some welfare-promoting acts if the practice of blaming acts of that general type is, as a whole, socially desirable.

This approach offers one way to reconcile our moral instincts and intuitions with the view that our moral system tends to advance welfare. As noted, because the moral rules must be categorical, it is inevitable that sometimes they will deem wrong a particular act that in fact would increase welfare. (Thus, the act 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.) Actually, given that moral rules need to operate at a level of groups of acts, which are inevitably heterogenous, it would be surprising if there did not exist such cases. Accordingly, the occasional failure of our intuitive sense of right and wrong to reflect first-best choices of acts can be viewed as simply as an ordinary feature of a (second-best) optimal moral system rather than as a deep problem with consequentialist moral theories.

Despite this close affinity between our view and that of the philosophers who have developed two-level moral theories, there also are important differences. One is that, to a large extent, they asked if the existing list of moral rules — those that were most prominent and widely recognized — advanced welfare (to which they answered affirmatively). They did not generally organize their analysis by looking at particular acts or natural clusters of acts and asking, for each, what rule would best advance

---

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 (1979, 1996), Hare (1981), Harrod (1936), Rawls (1955), and Sartorius (1972).

welfare. (That is, they did not systematically seek to consider the full range of behavior, rather than that presumed to be subject to moral rules, and they did not ask what rule maximized welfare so much as whether the existing rule promoted welfare to an extent.) The greatest difference, however, is that the question of how best to deploy moral sanctions and rewards was not much addressed. In particular, they did not focus on whether guilt or virtue, or some combination of the two, was best to use, and or on whether the observed use of guilt and virtue (and disapprobation and approval) corresponded to that which would be socially optimal. Moreover, they did not, for the most part, take into account that the experience of guilt and virtue is part of utility and, accordingly, will influence how the moral emotions should best be employed.[59] Nor did they make explicit all of their assumptions about human nature and systematically trace the implications.

Acting from self-interest versus acting from moral obligation. — Another strand of philosophical literature that relates to our analysis — one that in modern times sometimes addresses economics directly — is that focusing on the concept of self-interest. Many philosophers — notably, Kant (1785) — and some economists are emphatic that acting out of a sense of obligation or duty is distinct from acting out of self-interest.[60] This, of course, raises the question that occupied such philosophers as Hume, Mill, and Sidgwick: If an act is against self-interest and is nevertheless committed because it is morally the right thing to do, what is it that, as a positive matter, can explain such behavior? Of course, an important possibility is that the moral sentiments — feelings of guilt and of virtue and, relatedly, concern for the disapprobation or approval of others — provide the explanation. When the utility effects of the moral emotions outweigh the utility associated with the act per se, individuals will behave differently. Whether this is described as a part of self-interested behavior or whether we choose to characterize such acts as motivated by duty or obligation rather than self-interest is largely a semantic dispute. Likewise, the very use of terms like guilt and virtue to capture whatever it is that motivates individuals to follow moral rules when doing so would otherwise be against their narrowly defined self-interest is, in an important respect, tautological. (See our previous discussion in note 9.)[61]

---

[59]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 formulation of moral rules.

[60]See, for example, Smith's (1790) discussion of Mandeville and Hobbes, and also Hutcheson's (1725-1755) attempts to distinguish acts based on self-interest from those based on obligation or benevolence. For modern examples, see Anderson (2000), Scheffler (1992), and Sen (1977).

[61]In reading the philosophical literature, it appears that many writers seem to believe that individuals who "do their duty" do not obtain pleasure thereby, but rather they feel compelled to do the morally correct act. The assumption is that proponents of the self-interest view rest their arguments on positive utility, such as from altruistic feelings toward others or simply from the feeling of virtue associated with doing the right thing. Against this argument, it is suggested (plausibly, in our view) that individuals would in fact have preferred that the situation, in which they have to sacrifice "ordinary" utility in order to comply with the dictates of morality, had never arisen. Thus, it is argued that it is not utility from doing one's duty that motivates moral behavior. This response, however, ignores an alternative, plausible interpretation: Perhaps it is not pleasure that motivates moral behavior in such situations, but rather the desire to avoid pain; that is, the moral rules in question may be enforced by guilt rather

On the independent importance of acting morally. — A common objection to consequentialist accounts of morality is that they conflict with our intuition that compliance with the dictates of morality has weight independent of the consequences of our actions. For example, Ross (1930) suggests that an individual following a consequentialist morality would be indifferent to whether a promise should be kept when the balance of benefits and harm was precisely equal, whereas our moral intuition is that at least some weight should be accorded to keeping the promise. Arguments about whether consequentialism can account for our instinct that promise-keeping is independently important often involve consequentialists identifying indirect consequences of breaking promises (such as by setting a bad example that will affect others' behavior) and critics posing hypothetical examples in which such effects are absent (a promise to a dying person that will never become known to anyone else) or suggesting that such additional effects be subsumed in the balance of benefits and harm and asking whether our intuition about promise-keeping still seems to carry weight.

Without entering into the particulars of prior debates, such as that about promise-keeping, we nevertheless observe that our framework of analysis has a straightforward implication regarding whether consequentialism can account for the moral intuition that morality has independent weight. Specifically, in the consequentialist moral system that we have described, anything morally prohibited is associated with feelings of guilt (and possibly shame and the prospect of being subject to disapprobation) and anything morally encouraged is associated with feelings of virtue (and an expectation of approbation). If this scheme has even modest descriptive accuracy, as we suggest it does in subsection 5.1, then an explanation has been offered: The tendency for moral behavior to be associated with moral sentiments is not contingent on an independent consequentialist assessment of such behavior, but rather is associated with all behavior subject to the pertinent moral rule. Moreover, when moral rules operate at a categorical level, all acts in the relevant category — such as that consisting of situations in which we may decide whether to keep a promise — will lead, say, to the experiencing of guilt (even if the act happens to be, on balance, socially neutral or desirable in its consequences). Indeed, many who have written about the moral sentiments in times past, such as Sidgwick (1907), suggest a link between moral sentiments and moral instincts and intuitions. And to the extent that moral emotions tend to have an autonomous, not entirely conscious character, being triggered by particular actions and being anticipated by the mere contemplation thereof, their phenomenology does seem similar to that of the intuitions to which philosophers refer.

The problem of unlimited individual obligations under consequentialism. — It is commonly objected that a consequentialist or utilitarian moral criterion is too demanding: Everyone would have to be a Mother Theresa; individuals in richer countries would have to donate most of their income to help poor people in less developed nations; and so forth. Such implications, it is said, are

---

than by virtue. (And, indeed, most of the moral rules that are addressed in this literature seem to be those that are enforced by guilt.) Clearly, if one is induced to sacrifice act-utility to avoid a guilt feelings of a greater magnitude, one would wish that the situation had never arisen, but this hardly suggests that one cannot use a broad notion of utility to incorporate moral sentiments in a manner that explains individuals' tendency to conform their behavior to moral rules.

inconsistent with our moral intuitions, which in turn is taken to demonstrate that consequentialist moral philosophy is fundamentally defective.[62] Implicit or explicit in such arguments is the claim that our moral intuitions constitute (or at least indicate some of the contours of) an ideal moral system.[63]

Our analysis, as we have already suggested, offers an answer to this criticism, that is, an explanation for how a consequentialist view can be reconciled with our seemingly inconsistent moral intuitions. According to the posited criterion, it would be a good thing if individuals behaved as stated. But, given human nature, it is unrealistic to expect this. Thus, even if one attempted to inculcate a high degree of guilt for failing to make substantial sacrifices to help others, it would probably be insufficient to induce most individuals to so behave. The consequence would be that more individuals would frequently suffer guilt; moreover, the crowd-out of guilt in other realms would impose serious social costs. Accordingly, it is not optimal to use guilt to encourage such behavior. In other words, a proper consequentialist analysis would seem to oppose, not favor, deeming the failure to engage in such highly altruistic behavior as a moral wrong.

It may, however, be advantageous to use virtue: Such use may encourage some highly desirable acts and, because most will not act accordingly, there will be little depletion of the limited capacity for virtue and thus little social cost with regard to regulating other behavior. Hence, on instrumental grounds, one would indeed not use guilt but instead use virtue — consistent with moral philosophers' categorization of such acts as ones that individuals are not morally obligated to perform, but that it would nevertheless be morally worthy to do.

By contrast, one can consider acts of helping others in distress when there would be little disutility suffered by the actor, for example, calling for help or rescuing someone when only slight inconvenience would be involved. Here, many moral philosophers seem to agree that, when we consult our moral intuitions, it seems that there is a moral duty to act. This view is also suggested by our framework: One would probably use guilt to induce such behavior because little guilt would be required and because, since most individuals would be induced to behave properly, there would little use of the scarce reservoir of guilt.

Combining these two cases, what has appeared to be a puzzle to some moral philosophers — who have trouble rationalizing our moral intuition that tells us that there is no duty in the former case with our intuition that there is a duty in the latter — can be solved by a more explicit welfare-based analysis of the instruments of guilt and virtue (rather than focusing exclusively on the characteristics of the acts themselves, whether on their consequences or on the intrinsic properties that some believe them

---

[62]See, for example, Williams (1973). See Heyd (1982) for a broad exploration of supererogation under a range of moral and religious theories (including utilitarianism).

[63]Our previous discussion of two-level moral theories, and their relationship to problems of second-best optimization, suggests a direct objection to this form of reasoning, even if the argument to follow were invalid.

to have).[64]

The act/omission distinction. — Another set of debates among moral philosophers concerns the act/omission distinction.[65] The problem, as often put, is that there is another important type of conflict between our moral intuitions and the implications of a consequentialist approach. Namely, one can consider two situations in which an act in the first has precisely the same consequences as inaction (an omission) in the second. Moreover, in some such cases, our moral intuitions would distinguish the two situations, forbidding the act in one situation but permitting (or requiring) the corresponding omission in the other, such as when the act and omission both raise welfare — perhaps five lives would be saved at the expense of one. It is suggested that our moral intuition is such that it is impermissible to sacrifice the one life to save five through an act, whereas it would be permissible (or even mandatory) to do so as a result of a failure to act (that is, one should not act to save one if doing so would kill five).

Our framework offers a number of possible ways to reconcile this seemingly inconsistent character of common morality with a consequentialist moral system. One explanation, as discussed earlier in this subsection and in subsection 5.5, involves the grouping of acts: The groupings that naturally arise are based upon characteristics of different types of behavior that relate to how we perceive the world and organize it in our minds, and these need not correspond to the groupings that would be ideal from the perspective of formulating moral rules. Thus, a particular act may be condemned not because it is itself socially undesirable, but because it shares characteristics with other acts that together form a cluster for which it makes sense to inculcate guilt. For example, the rare situation in which killing one individual will save five may be grouped in the general category of killing, which does not ordinarily raise social welfare.

---

[64]Relatedly, many philosophers suppose that there must be a qualitative distinction between the acts — since there is moral obligation in one case and not the other — whereas there seems to be only differences in degree, namely, the cost of self-sacrifice. (Sometimes these differences are nevertheless described in qualitative terms, such as mere inconvenience versus disturbances to the integrity of one's self-defined mission in life.) In our model, however, differences in degree can translate into differences in kind. In particular, as the number of individuals who will not behave in a first-best manner increases, at some point it is no longer optimal to employ guilt. (And, as our discussion in note 17 implies, this change can even be discontinuous: One might move from an interior optimum for g to an optimum at which g = 0.)

[65]Debate over the act/omission distinction is closely related to that concerned with the doctrine of double effect, the doctrine of doing and allowing, and other principles that draw similar distinctions. For differing views, see, for example, Bennett (1995) and Williams (1973). For prior analysis of the act/omission distinction from a psychological perspective, with the suggestion that it involves overgeneralization of an otherwise useful decisionmaking heuristic, see, for example, Spranca, Minsk, and Baron (1991).

We note that the act/omission distinction relates to the foregoing subject of unlimited individual obligations under consequentialism because many moral theorists attribute that alleged problem to consequentialism's failure (inability) to distinguish acts and omissions, whereas if duties are largely limited to affirmative acts, individuals' obligations can more readily be limited. See Bennett (1995), who discusses the failure of attempts by the classical utilitarians to deal with the problem. Subsequently, as discussed in Heyd (1982), a number of moral theorists have considered whether "negative utilitarianism" (under which some variant of the act/omission distinction is embraced) can be a plausible or appealing moral theory.

Another possibility, which we have not raised previously, concerns the stimuli necessary to trigger feelings of guilt and virtue.[66] Because these emotions need to be reasonably automatic to function, it must be that they are experienced not as a result of careful contemplation and reflection, but rather due to particular patterns being identified. Perhaps it is the case that acts are more naturally capable of triggering these emotions than are omissions, for committing an act may result in a more identifiable stimulus than an omission. A related problem is that many omissions (failure to devote more resources to helping the poor) are ongoing — every instant in which we could have acted but did not is an omission — whereas we do not have the capacity to experience constant flows of guilt or virtue that will register in a meaningful way.[67] Hence, the underlying psychology of moral emotions may impose important constraints on their use that help to explain the system of moral rules and the use of moral sentiments that we observe, but in a manner that does not imply that behaviors are benign (from an ideal perspective) simply because they are not associated with guilt or virtue.

5.9. On the use of morality and the legal system to control behavior. — Throughout, we have assumed that the moral system was the only mechanism available to control behavior. But there are others, notably, the legal system, including regulation, taxes and subsidies, the criminal law, and so forth.[68] Therefore, it is natural to ask when it is optimal to use morality alone, just the legal system, or some combination of the two.

We briefly sketch some of the relevant considerations. The moral system has the advantage that enforcement is automatic for internal sanctions and nearly so for external sanctions in some settings, the administrative costs are low, and those who apply the sanctions (both internal and external) often have the pertinent information already. The formal legal system, however, usually can impose higher sanctions, can influence individuals not greatly restrained by the moral system, can adjust more quickly in response to changing conditions, and can employ more complex and fine-tuned rules than those possible with a moral code that must be inculcated in children and applied with little deliberation. We explore these factors further in current research and suggest that the actual use of common morality and

---

[66]A third, related point is that, to the extent that the labeling of behaviors as involving acts or omissions (is failing to hold open a door for the next person an omission, or the commission of the act of releasing the door so as to impose the risk of injury on another?) may be inculcated at the same time guilt and virtue are being inculcated, it may be that we encourage people to perceive as distinct acts only those behaviors for which we are going to inculcate guilt or virtue. Thus, whenever our analysis suggests that it is not optimal to employ guilt, for example, there is little point in teaching individuals to recognize the corresponding undesirable behavior as a distinctive act.

[67]Of course, some omissions are distinctive, so the failure to call for help when one sees a person drowning is different from the ongoing failure to donate half of one's income to charity. Bennett (1995) offers further examples in which what would seem to be omissions, as ordinarily defined, are viewed in moral discussions as if they were acts. And, as noted in the preceding discussion, it seems that some such omissions — notably, failing to aid when the sacrifice to oneself is trivial and the benefit to the third party is great — are indeed associated with feelings of guilt.

[68]In addition to common morality and the legal system, considerations of self-interest (such as concerns for reputation) also influence externality-causing behavior. See note 45.

the legal system is roughly in accord with what seems optimal. See Shavell (2002); see also Ellickson (1991), Posner and Rasmusen (1999) and Sidgwick (1897).

6. Conclusion

Our analysis offers a theory of how moral sanctions and rewards — feelings of guilt and virtue — would be assigned to particular acts or to natural groups of acts if the purpose was to maximize social welfare. In applying standard economic techniques to the subject of moral rules and the moral sentiments, we hope to illuminate a set of problems that has occupied a range of scholars over the years. Some writers, such as Hume (1739, 1751), Mill (1861), and Sidgwick (1907), have adopted the view that the function of moral rules and, relatedly, the moral sentiments (feelings of guilt and virtue) and social behavior (expressions of disapprobation and approval) that accompany them is to promote well-being. Others, including Kant (1785) and many modern followers, have taken a different view. From a purely descriptive perspective, it is helpful in considering this question to determine what would be the contours of a system of common morality if indeed its purpose was to advance welfare.

Our inquiry looks beyond the most apparent features of moral rules — that the behavior that is condemned tends to be socially undesirable (on account of negative externalities) and that the behavior that is deemed worthy tends to be socially desirable (due to positive externalities) — to ask whether the manner of enforcement is consistent with welfare maximization. We take into account that the use of guilt and virtue affects social welfare not only by affecting behavior, but also because the actual experiencing of guilt or virtue affects individuals' utility. Moreover, we consider how inculcation costs and constraints on the use of guilt and virtue affect how they should be employed in the enforcement of moral rules.

As we discuss, our analysis has implications concerning whether moral rules will be supported by guilt or virtue, a subject that to our knowledge has not been systematically examined previously. In addition, our results (and, it would seem, many plausible extensions thereof) help to explain a number of well-known features of moral rules, such as their tendency to be overinclusive and to overlap and conflict. Moreover, these explanations help to illuminate certain longstanding debates among moral philosophers.

A more refined understanding of how to formulate moral rules to advance welfare also has normative significance. Most directly, much attention is devoted to how individuals should rear their children, to what should be the content of education (especially as it pertains to values and behavior toward others), and, for still significant groups, to what should be understood as proper religious precepts.[69] Moreover, in more conventional realms of economic policy analysis, there is increasing

[69]Prior explorations of the general idea that welfare may be raised by changing individuals' utility functions include Harsanyi (1953-1954) and Weisbrod (1977). The potential normative implication of our analysis assumes, of course, that maximizing aggregate well-being should indeed be the social objective, a subject that we do not consider here but pursue as some length in Kaplow and Shavell (2002).

attention by some scholars (for example, legal scholars) to how government policy may be used to reinforce or modify common morality (often described as social norms).[70] Thus, it might be urged that certain laws should be enacted despite their inefficacy, if they would support norms against certain undesirable behavior. The idea seems to be that government policy plays a part in the process of inculcating guilt and virtue. Viewed in this manner, our analysis suggests that it is an oversimplification to assume that it is always a good idea to inculcate more guilt for undesirable acts and virtue for desirable acts. In addition to the direct costs of inculcation, there are other possible costs, notably that if one inculcates guilt but it is not very successful in controlling behavior, the primary effect will be to reduce the utility of most individuals due to their experiencing guilt feelings, and that due to constraints on individuals' capacity to experience moral emotions, one may be eroding the effectiveness of guilt and virtue in areas where they are more successful in controlling behavior. In all, the optimization problem is more complex and subtle than generally seems to be appreciated.

At the present stage of our investigation, however, it is obviously premature either 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. First, we need a better understanding of the actual workings of our moral emotions along a number of dimensions that we identified. Second, further analysis is necessary in order to confirm or refute various conjectures that we offer and to address important considerations that, in this preliminary inquiry, we have overlooked. Finally, both positive and normative applications of such analysis requires a far more sophisticated understanding of the actual system of morality that we have and how various actions might influence it. 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 enforce moral rules.

[70]See, for example, Sunstein (1996), Weisbrod (1977).

Appendix

Proof of Proposition 1. In proving this proposition, we will maximize (3.1) without regard to the constraint (3.2). As will be seen, the left side of (3.2) is zero at the optimum we will obtain — that is, guilt is never experienced (as point (b) of the proposition claims); hence, the constraint will be satisfied. Furthermore, to choose the function g(u,h) to maximize (3.1), the optimum, g*(u,h), is found by maximizing the difference between the integrands of (3.1) for each (u,h) because the choice of g for any particular (u,h) does not affect how g should be chosen for any other (u,h).

There are two cases to consider. First, if u > g, the act is committed. In this case, the difference between the integrands of (3.1) is (u - h - g)f(u,h) - "(g), which is maximized at g = 0, allowing us to write this value as (u - h)f(u,h). Second, if u # g, the act is not committed. In this case, the difference between the integrands is -"(g), which is maximized (subject to the constraint for this case that u # g) at g = u, giving a value of -"(u).

As a consequence, when u > 0, g* = u if -"(u) > (u - h)f(u,h), which is to say, if "(u) < (h - u)f(u,h) — that is, if expression (3.3) holds. Otherwise, g* = 0. This establishes proposition 1(d). For proposition 1(a), the claim that g* > 0 implies that u < h follows because g* > 0 implies that the left side of (3.3) is positive, so the right side of (3.3) must be positive as well, which in turn requires that u < h. That g* > 0 implies that g* = u has already been shown. Proposition 1(b) also follows from the above analysis because, whenever g* > 0, the act is not committed. Proposition 1(c) requires that desirable acts — acts for which u > h — never be deterred, which also follows from (3.3) because the right side would be negative in such situations.

Proof of Proposition 2. To demonstrate proposition 2(a), we first observe that, from expression (3.8), g_i must maximize $W_i(g_i) - 8y_i(g_i)$ because the $W_j(g_j)$ and $y_j(g_j)$, j ... i, do not depend on g_i. Thus, $g_i > 0$ cannot be optimal if the following expression is positive for all $g_i > 0$.[71]

[71]In writing (A.1), we find it convenient, with respect to using expression (3.5) for W_i, to state g_i separately, taking advantage of the fact that g_i is a constant when integrating with respect to u and h.

$$(A.1)\quad \left(W_i(0) - \lambda y_i(0)\right) - \left(W_i(g_i) - \lambda y_i(g_i)\right)$$

$$= p_i \int_0^{\infty} \int_0^{g_i} (u-h) f_i(u,h)\,du\,dh + \alpha_i(g_i) + p_i(1+\lambda) g_i \left(1 - F_i(g_i)\right).$$

If $u \geq h$ for all (u,h) in S_i, then (A.1) is clearly positive for any $g_i > 0$, meaning that $g_i^* = 0$ if acting is first best for all acts in S_i.

To prove 2(b) and 2(c), it suffices to construct an example. For simplicity, let the example be such that constraint (3.2) 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 set S_i because the only effect of raising g above 1 would be to increase inculcation costs.) Now, as long as G is taken to exceed 1, (3.2) will never be binding. Hence, $\theta = 0$. For the remainder of the example, confine attention to a particular set S_i. Assume that the distributions of u and of h on that set are independent. For u, assume a triangular distribution on [-1,1], such that $f(-1) = f(1) = 0$ and $f(0) = 1$. 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 η be constant and equal to 0.0375. Now, using the first-order condition (3.9) 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, $g_i^* = 0.5$.[72] Proposition 2(b), that guilt may sometimes be experienced, is true because, whenever $u > 0.5$, the act is committed and guilt is therefore experienced. Proposition 2(c), that undesirable acts may be committed and desirable acts deterred is also immediate. For the former, 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). For the latter, all acts for which $u \leq 0.5$ are deterred, but, for all such $u > 0$, there will be situations in which $h < u$.

Finally, 2(d) follows because expression (3.9), the first-order condition, is a necessary condition for an interior optimum.

Proof of Proposition 3. Proposition 3(b): Let us first show that, at the optimum, positive guilt is never experienced. If guilt $g > 0$ is experienced, then by definition it must be that the person commits the act, that is, $u - g > v$. Hence, social welfare in situation (u,h) is $(u - h - g)f - \eta(g) - \theta(v)$, where f stands for f(u,h). Clearly, if g is lowered to 0, then the person still commits the act (for $u > v$ must hold)

[72]This must be a maximum because 0.5 is the only solution to (3.9) and the derivative of (3.8) with respect to g_i is positive at $g_i = 0$ (it is $0.1(1-\frac{1}{2})-0.0375 = 0.0125$).

and social welfare rises. Also, as g has been reduced, the constraint (4.2) will still be satisfied; and since v is not realized under the present supposition, the constraint (4.3) is not affected. In summary, lowering g to 0 raises social welfare and does not affect the constraints, so that g > 0 cannot have been optimal.

Let us next show that v*(u,h) > 0 implies that virtue is experienced in situation (u,h). If v > 0 is not experienced, then it must by definition be that the person commits the act, that is, u - g > v, so that social welfare is (u - h - g)f - "(g) - $(v). If v is lowered to 0, then the person still commits the act (for u - g > 0 must hold) and social welfare rises due to the elimination of inculcation costs for v. Also, since v is reduced, the constraint (4.3) will still hold. Hence, v > 0 with v not being experienced cannot characterize the optimum.

To prove the other results, we will maximize (4.4) over functions g(u,h) and v(u,h), taking advantage of the preceding result, which implies that the Lagrange multiplier 8 (corresponding to the constraint on realized guilt) is zero. As in our proof of proposition 1, the optimal g*(u,h) and v*(u,h) are found by maximizing pointwise, for each (u,h), the combination of the integrands in (4.4) — but, as just noted, ignoring the fourth term. That is, for each (u,h), we choose g $ 0 and v $ 0 to maximize

$$(A.2)\quad (u - g - h)f - \alpha(g) - \beta(v) \qquad \text{if } u - g > v, \text{ and}$$

$$(A.3)\quad (1 - \mu)vf - \alpha(g) - \beta(v) \qquad \text{if } u - g \leq v,$$

where, as before, f stands for f(u,h). Expression (A.2) corresponds to the case where the act is committed and (A.3) to the case where the act is not committed.

Proposition 3(a): Let us first show that, if g*(u,h) > 0 or v*(u,h) > 0, then u < h, which is to say that not acting is first best. To demonstrate this, we will show that, if u $ h, then g* = v* = 0. If the act is committed, then 3(b) implies that g* = v* = 0. Hence, from (A.2) the maximand is (u - h)f > 0. If the act is not committed, then (A.3) applies, and it must be negative: the assumption that $ℕ(0) > (1 - :)f implies that (1 - :)vf < $ℕ(0)v; $O(v) $ 0 and $(0) = 0 imply that $ℕ(0)v # $(v); hence (1 - :)vf < $(v), so the value of (A.3) is negative. But the value of (A.2) is positive, so it is optimal to select g and v such that the act is committed. And, as already shown, in that case g* = v* = 0.

To complete the proof of 3(a), we need to show that, if g*(u,h) > 0 or v*(u,h) > 0, then g* + v* = u. Now we know from 3(b) that, if g or v is positive, the act is not committed. Hence, g* + v* $ u must hold and (A.3) applies. We now show that g* + v* > u cannot be optimal. Suppose that it is optimal and that g* > 0. If g is lowered slightly, g + v > u will still hold, so the act will still not be committed, and the value of (A.3) will rise, contradicting the assumption that the posited g* and v*

are optimal. Now suppose that g* + v* > u and that v* > 0. If v is lowered slightly, g + v > u will still hold, so the act will still not be committed. Moreover, (A.3) is falling in v: the derivative of (A.3) is (1 - :)f - $N(v), which is less than or equal to (1 - :)f - $N(0) since $O(v) $ 0, and (1 - :)f - $N(0) < 0 by hypothesis. It follows that (A.3) rises when v falls, contradicting the assumption that the posited g* and v* are optimal. We conclude that g* + v* = u must hold.

Proposition 3(c): To show that the only possible deviation from first-best behavior is the commission of undesirable acts, we need to show that when u > h, the act is committed. Proposition 1(a) implies that g* = v* = 0 in this case, which, when combined with the fact that u > h $ 0, implies that the act is committed.

Proposition 3(d): Suppose that g*(u,h) > 0 or v*(u,h) > 0. We want to show that (4.5) holds for the g and v that maximize (4.4) subject to g + v = u. But (4.4) is equivalent to (A.2) and (A.3). Now, by 3(a), g* + v* = u, so that g* and v* must maximize (A.3) subject to g* + v* = u. Moreover, optimality requires that (A.3) be greater than (A.2), the maximand when the act is committed, and we know from 3(b) that g = v = 0 in that case. Combining these expressions yields (4.5).

Conversely, suppose that u > 0 and (4.5) holds for the g and v that maximize the Lagrangian (4.4) subject to g + v = u. We want to show that g*(u,h) > 0 or v*(u,h) > 0. Now, if the Lagrangian is maximized subject to the constraint g + v = u, then, as explained above, (A.3) is maximized subject to this constraint at the same g and v. Hence, the fact that (4.5) holds implies that not committing the act is optimal. And, for the act not to be committed, it is necessary that g + v $ u. Since it is assumed that u > 0, this implies that it is optimal to instill guilt or virtue.

Proof of Proposition 4. That proposition 3(b) continues to hold is clear, for the above proof of 3(b) does not make use of the assumption that $N(0) > (1 - :)f. Also, from the proof of proposition 3, we know that solving (A.2) and (A.3) determines the optimal g and v.

Proposition 3(a) need not hold: Consider the following example. First, assume that G = 0, so that guilt is never employed. Next, suppose that, for all acts, 0.1 < u < 0.2 and h < 0.1, that $N = 0.1, and that V > 1. At v = 0, inculcation costs are zero, all acts are committed (because u > 0 in every situation), and social welfare must be less than 0.2 (as 0.2 exceeds the utility of any act and therefore exceeds the average utility of acts minus the average harm caused by acts). Now, suppose that, for all (u,h), we set v = 1. (This is feasible since V > 1.) All acts are deterred. Welfare as a consequence of experiencing virtue is 1. Inculcation costs are 0.1. Hence, total welfare is 0.9. Because welfare is higher when v = 1 for all acts than when v = 0 for all acts, it is not optimal to set v = 0 for all acts even though u $ h for all acts. This contradicts the first claim of proposition 3(a). With regard to the second claim of proposition 3(a), in the present example it requires that, if v* > 0 for any act, then v* = u for that act. This claim is contradicted because welfare is higher at v = 1 for all acts (it is 0.9) than if we set v = u for all acts for which v* > 0, because welfare in the latter case must be less than 0.2 (acts for which v = u produce utility of v, which is less than 0.2; those for which v = 0 produce utility of u, which is less than 0.2, and also cause harm, and there are inculcation costs).

Proposition 3(c) need not hold: This requires that it might be optimal for desirable acts to be deterred, which is demonstrated by the foregoing example.

Proposition 3(d) need not hold: We will show that it can be optimal to instill guilt or virtue — virtue in particular — even if (4.5) does not hold at the g and v that maximize the Lagrangian (4.4) subject to g + v = u. We use the foregoing example, adding the assumption that f(u,h) = 1 on S^o, where S^o = {(u,h)*h 0 [0.001, 0.002],u 0 [0.11, 0.12]}. Now, since G = 0, we know that g = 0 in all situations. Therefore, 3(d) states that, if v*(u,h) > 0, then, (4.5) must hold at v = u. For any (u,h) 0 S^o, the left side of (4.5) is greater than or equal to 0.011 because the lowest u 0 S^o is 0.11, v = u, and $(0.11) = 0.011. For any (u,h) 0 S^o, the right side is less than or equal to 0.002 because the greatest h 0 S^o is 0.002, v = u, and f = 1. (If : > 0, the right hand side is even lower.) Therefore, (4.5) does not hold for any (u,h) 0 S^o. Nevertheless, it is optimal to instill virtue on S^o: for any (u,h) 0 S^o, at v = 0, the act is committed and welfare cannot exceed 0.12 (the highest value of u on S^o), but at v = 1, the act is deterred, virtue of 1 is experienced, and the inculcation cost is 0.1, so welfare equals 0.9.

(Observe that the converse part of 3(d) — that if (4.5) holds, it is optimal to instill guilt or virtue — is still valid, for the above proof of that part of 3(d) did not make use of the assumption that $N(0) > (1 - :)f or of any results that relied on it.)

Proof of Proposition 5. To establish proposition 5(a), we first observe that, from expression (4.12), g_i and v_i must maximize $W_i(g_i,v_i) - 8y_i(g_i,v_i) - :z_i(g_i,v_i)$. Thus, neither $g_i > 0$ nor $v_i > 0$ can be optimal if the following expression is positive for all $g_i > 0$ and $v_i > 0$.[73]

[73]As in note 71, 73 concerning (A.1), when writing (A.4) we find it convenient, with respect to using expression (4.7) 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.

$$(A.4)\ \left(W_i(0,0) - \lambda y_i(0,0) - \mu z_i(0,0)\right) - \left(W_i(g_i,v_i) - \lambda y_i(g_i,v_i) - \mu z_i(g_i,v_i)\right)$$

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

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

If u $ h for all (u,h) in S_i, then (A.4) 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, all terms are obviously strictly positive or equal to zero except possibly for the last one. But we now show that it, combined with the $\$_i\mathbb{N}(v_i)$ term, is positive. Given that $\$_i\mathbb{N}(0) > (1-\kappa)p_i$, $\$_i\mathbb{N}(0)v_i > p_i(1-\kappa)v_i$ $ $p_i(1-\kappa)v_i F_i(g_i+v_i)$. Moreover, $\$_i(v_i) > \$_i\mathbb{N}(0)v_i$ because $\$_i(0) = 0$ and $\$_i\mathbb{O}(0)$ $ 0. Therefore, $\$_i(v_i) > p_i(1-\kappa)v_i F_i(g_i+v_i)$, so the two terms combined are positive.

To prove 5(b) and 5(c), it suffices to construct an example for each claim. For all of the claims except the latter claim of 5(b), that virtue may not always be experienced, a modification of the example used to prove 2(b) and 2(c) will suffice: Use that same example, combined with the assumption that V = 0, so that virtue cannot be used.

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 set (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, $(0.1) < 0.2, $(1) > 1, and V > 1.[74] 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 < 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

[74]These assumptions are consistent with the assumption of the proposition that $\$_i\mathbb{N}(0) > (1-\kappa)p_i$: as will be seen, v* < 1, so the constraint is not binding, which implies that $\kappa = 0$; thus, the right side equals 1; finally, the assumption that $\$\mathbb{N}(0) > 1$ is consistent with the assumption in this example that $(0.1) < 0.2.

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.

Finally, 5(d) follows because expressions (4.13) and (4.14), the first-order conditions, are necessary conditions for an interior solution.

Proof of Proposition 6. It suffices to offer an example in which $v_i^* > 0$ even though u $ h for all (u,h) in S_i. Suppose that there is only a single subset (with p = 1), that for all acts u < 0.1 and h < u, that $(1) < 0.5, and V > 1. Furthermore, assume that G = 0, so that guilt cannot be used. Now, at v = 0, expected social welfare is less than 0.1, for every act is committed, results in act-utility of less than 0.1 and possibly some harm (no inculcation costs are incurred). If, however, v = 1, total welfare exceeds 0.5, for all acts are deterred, in each situation virtue equal to 1 is experienced, and inculcation costs are less than 0.5. Moreover, because V > 1, v = 1 is feasible. Hence, the claim is established.

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.

Anderson, Elizabeth. 2000. Beyond Homo Economicus: New Developments in Theories of Social Norms. *Philosophy and Public Affairs* 29: 170-200.

Austin, John. 1832. *The Province of Jurisprudence Determined*. Wilfrid E. Rumble, ed. Cambridge: Cambridge University Press (1995).

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

Barash, David P. 1982. *Sociobiology and Behavior*. Elsevier: New York, 2nd ed.

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.

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

Bennett, Jonathan. 1995. *The Act Itself*. Oxford: Oxford University Press.

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

Berridge, Kent C. 1996. Food Reward: Brain Substrates of Wanting and Liking. *Neuroscience and Biobehavioral Reviews* 20: 1-25.

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

Brandt, Richard B. 1979. *A Theory of the Good and the Right*. Oxford: Oxford University 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.

Ellickson, Robert. 1991. *Order without Law: How Neighbors Settle Disputes*. Cambridge: Harvard University Press.

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

Elster, Jon. 1999. *Alchemies of the Mind: Rationality and the Emotions*. Cambridge: Cambridge University Press.

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.

Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. Psychological Games and Sequential Rationality. *Games and Economic Behavior* 1: 60-79.

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

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.

Hechter, Michael, and Karl-Dieter Opp, eds. 2001. *Social Norms*. New York: Russell Sage Foundation.

Heyd, David. 1982. *Supererogation: Its Status in Ethical Theory*. Cambridge: Cambridge University Press.

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).

Hutcheson, Francis. 1725-1755. *Philosophical Writings*. R.S. Downie, ed. London: J.M. Dent (1994).

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

Kagan, Jerome. 1984. *The Nature of the Child*. New York: Basic Books.

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.

Kant, Immanuel. 1785. *Groundwork of the Metaphysics of Morals*. Translated and edited by Mary Gregor, Cambridge: Cambridge University Press (1997).

Kaplow, Louis. 1990. A Note on the Optimal Use of Nonmonetary Sanctions. *Journal of Public Economics* 42: 245-247.

Kaplow, Louis, and Steven Shavell. 2002. *Fairness versus Welfare.* Cambridge: Harvard University Press (forthcoming).

Kosslyn, Stephen M., and Olivier Koenig. 1992. *Wet Mind: The New Cognitive Neuroscience*. New York: Free Press.

LeDoux, Joseph E. 1996. *The Emotional Brain*. New York: Simon & Schuster.

Lewontin, R.C., Steven Rose, and Leon J. Kamin. 1984. *Not in Our Genes: Biology, Ideology, and Human Nature*. New York: Pantheon Books.

Mackie, J.L. 1985. *Persons and Values: Selected Papers, Volume II*. Joan Mackie and Penelope Mackie, eds. Oxford: Oxford University Press.

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.

Ostrom, Elinor. 2000. Collective Action and the Evolution of Social Norms. *Journal of Economic Perspectives* 14: 137-58.

Pettit, Philip. 1990. *Virtus Normativa*: Rational Choice Perspectives. *Ethics* 100: 725-55.

Pinker, Steven. 1997. *How the Mind Works*. New York: W.W. Norton & Co.

Polinsky, A. Mitchell, and Steven Shavell. 1984. The Optimal Use of Fines and Imprisonment. *Journal of Public Economics* 24: 89-99.

Posner, Eric A. 2000. *Law and Social Norms*. Cambridge: Harvard University Press.

Posner, Richard A., and Eric B. Rasmusen. 1999. Creating and Enforcing Norms, with Special Reference to Sanctions. *International Review of Law and Economics* 19: 369-82.

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.

Romer, Paul. 1996. Preferences, Promises, and the Politics of Entitlement. In Victor R. Fuchs, eds., *Individual and Social Responsibility: Child Care, Education, Medical Care, and Long-Term Care in America*. Chicago: University of Chicago Press.

Ross, W.D. 1930. *The Right and the Good*. Oxford: Oxford University Press.

Sartorius, Rolf. 1972. Individual Conduct and Social Norms: A Utilitarian Account. *Ethics* 82: 200-18.

Scheffler, Samuel. 1992. *Human Morality*. New York: Oxford University Press.

Schelling, Thomas C. 1984. *Choice and Consequence*. Cambridge: Harvard University Press.

Sen, Amartya K. 1977. Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy and Public Affairs* 6: 317-44.

Shavell, Steven. 1987. The Optimal Use of Nonmonetary Sanctions as a Deterrent. *American Economic Review* 77: 584-92.

Shavell, Steven. 2002. Morality versus Law as Regulators of Conduct. *American Law and Economic Review* 4 (forthcoming).

Sidgwick, Henry. 1897. Law and Morality. In *The Elements of Politics*. Second Edition. Reprinted in Henry Sidgwick, *Essays on Ethics and Method*, Marcus G. Singer, ed. Oxford: Oxford University Press (2000).

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.

Sunstein, Cass. 1996. Social Norms and Social Roles. *Columbia Law Review* 96: 903-68.

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.

Tooby, John, and Leda Cosmides. 1990. On the Universality of Human Nature and the Uniqueness of the Individual: The Role of Genetics and Adaptation. *Journal of Personality* 58: 17-67.

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.

Williams, Bernard. 1973. A Critique of Utilitarianism. In J.J.C. Smart and Bernard Williams, eds., *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

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.