UNIVERSITY OF CALIFORNIA AT BERKELEY

Department of Economics

Berkeley, California 94720-3880

Working Paper No. 95-241

# Moral Preferences, Moral Constraints, and Self-Serving Biases

Matthew Rabin

Department of Economics
University of California, Berkeley

August 1995

Key words: altruism, fairness, morality, non-expected utility, reciprocal altruism, selective exposure, self-serving biases, social influence

JEL Classification: A12, A13, B49, C91, D63, D81, D83

---

I thank Colin Camerer, Gary Charness, Tzachi Gilboa, Marcus Heng, Michael Kremer, David Laibson, Jeff Zweibel, and seminar participants at Berkeley, Stanford GSB, Cal Tech, Northwestern, and Ohio State, for helpful comments. I am also grateful to Gail Brennan, Jimmy Chan, Paul Ellickson, Marcus Heng, and Jin Woo Jung for valuable research assistance, and to the Russell Sage Foundation and Alfred P. Sloan Foundation for financial assistance on related projects.
Any comments very appreciated. Correspondence: Postal mail -- University of California, Department of Economics, 549 Evans Hall #3880, Berkeley, CA 94720-3880. Electronic mail -- rabin@econ. berkeley.edu. CB handle -- Game Boy.

**Abstract**

Economists have formally modeled moral dispositions by directly incorporating into utility analysis concern for the well-being of others. But sometimes moral dispositions are not *preferences*, as connoted by utility analysis, but rather are ingrained as (internal) *constraints*. I present a model fleshing out this distinction: If moral dispositions are internal constraints on a person's *real* goal of pursuing her self-interest, she will be keen to self-servingly gather, avoid, and interpret relevant evidence, for the purpose of relaxing this constraint and pursuing her self interest. This gives rise to self-serving biases in moral reasoning. I show that this alternative model has some implications different from a standard utility model. Specifically, because a person seeks to avoid information that interferes with her self interest, the scope for social influence in moral conduct is greater than it is in the conventional model. Outside parties can improve a person's moral conduct by a) forcing her to receive certain information, b) discouraging her from (selectively) thinking about other information, or c) encouraging her to think through moral principles before she knows where her self interest lies.

## 1. Introduction

We often feel a moral obligation both to refrain from enjoyable activities that hurt others and to pursue unpleasant activities that benefit others. Economists have begun to formally model such moral dispositions, by incorporating into a person's utility function concern for the well-being of others. For example, Becker (1981) and Margolis (1982) model altruism, and Rabin (1993) models reciprocal altruism. Some social scientists have argued, however, that integrating concern for others into the standard framework of utility maximization is either impossible or unproductive (see, e.g., Etzioni (1988) and Prelec (1991)). Such arguments vary, but one common theme is that many moral dispositions are not *preferences*, as connoted by utility analysis, but rather are ingrained as a set of (internalized) *rules* or *constraints*. To economists, this distinction may seem irrelevant or ill-conceived. If a consumer boycotts grapes because she believes their production exploits farm labor, such purposive, goal-directed behavior would seem readily conceptualized as a preference. Foregoing grapes out of moral qualms is just like foregoing them because they taste bad.

In this paper, I present a model fleshing out one possible distinction between moral preferences and moral constraints, and relate the distinction to self-serving biases in moral reasoning. I then show that this alternative model has some implications--especially regarding the scope for social influence in moral conduct--that differ from the predictions and prescriptions made by the standard preferences model.

People pursue preferences; they seek to circumvent constraints. In the realm of moral cognition, this distinction has interesting implications. If morality were direct preferences, a person would always have a taste for more information. But if morality is an internal constraint on her *real* goal of pursuing her self interest, a person will be keen to selectively and self-servingly gather, avoid, and interpret evidence that will tell her whether it is morally okay to pursue her self interest. Before purchasing food, we are always eager for more accurate information on whether it will taste good. We may be far less eager for accurate information about potential social harm of purchasing the food, because what we learn may interfere with pursuit of our self interest.

Formally, I develop a model of an agent maximizing her self interest subject to the (internal) constraint that she abide by certain moral obligations. Whether these internal constraints are binding depends on the information an agent acquires, so that her goal in information acquisition will be to relax these internal constraints. I believe the formal model parallels the psychological logic connecting moral obligations to self-serving biases. We search our minds for excuses to avoid our obligations to others; we (most of us) don't search for excuses to pursue moral obligations at the expense of our self interest. The psychologists Messick and Sentis (1983, pp. 89-90) hypothesize the following process, similar to the model I am proposing:

> ... Here, we will suggest that, in allocation-type tasks, subjects try out various ways of conceiving of fairness (equality with respect to what?) to determine the implications of the various options and select those conceptions of fairness that place them highest on their preference orderings. ... The result of this process is a type of constrained optimization of preference in which various ways of interpreting "fair" or "equal" determine the subset of outcomes that are feasible. ... Thus, when one maximizes one's preference within this constrained set, one is guaranteed of selecting an outcome that will be fair. ... We suspect that, at least in simple decisions, people are quickly aware of what they want and that most of the cognitive work that is expended is devoted to generating arguments and principles to support the preference.

Such a preference for information that supports our self interest only makes sense if self interest and "other interest" are *qualitatively* distinct components of our motivations, not merely arguments in a unified utility function.[1] This form of strategic belief manipulation, therefore, never arises in the standard expected-utility model.[2] The most significant behavioral

---

[1] Especially since economists have only begun to systematically explore departures from pure self interest, of course, probably most often the "preferences" approach is very adequate. This paper is merely an attempt to conceptualize some important limits to this approach.

[2] By assuming that we ourselves impose constraints in conflict with our own preferences, my model falls within the category of "multiple-self" models of human agency. Schelling (1978) calls such an approach "egonomics" and Ainslie (1992) calls it "picoeconomics." See also Elster (1985) and relevant citations in Section 7. My model also formally falls under the rubric of non-expected utility theory, though it differs substantially from any of the literature with which I am familiar. The fact that people may prefer less information to more has been discussed in this literature; see, for example, Kreps and

departure from expected-utility theory, and the main source of the model's predictions, is that an agent with moral constraints sometimes strictly prefers less information to more: When her beliefs tell her it is morally okay to engage in an enjoyable activity, an agent will avoid gathering further information that might jeapordize her moral green light.

The dangers of self-serving self-deception have often been identified when discussing great moral issues.[3] But the issues addressed in this paper arise as well in day-to-day economic, social, and legal interactions, and sometimes with important implications. One implication of the model is that people will behave less morally than their own moral principles would seem (if interpreted as "true preferences") to imply, because people will gravitate toward beliefs that allow them to pursue their self-interest. But evidence does not suggest that self-serving biases are strong enough to render moral constraints powerless, and I will not in any event focus in this paper on whether morality is a strong or weak influence on our behavior; rather, I shall emphasize implications of moral constraints that are *qualitatively* different than either straightforward moral preferences or pure self interest. For instance, some researchers have argued that a major source of legal and

---

Porteus (1979) and especially Wakker (1988). Finally, because it assumes preferences over beliefs that are not purely instrumental in the conventional sense, it is related to attempts by economists to incorporate cognitive dissonance into formal economic analysis; see Akerlof and Dickens (1982), Dickens (1986), Akerlof (1989), Montgomery (1993), and Rabin (1994). The current model differs from this literature by explicitly modeling both the motivation for changing our beliefs, and the mechanism by which we do so.

[3] For instance, Mousavizadeh (1995) quotes Richard von Weizsacker, former President of Federal Republic of Germany, in his speech to the German Parliament on the fortieth anniversary of the end of World War II:

> ... Whoever opened his eyes and ears and sought information could not fail to notice that Jews were being deported. There were many ways of not burdening one's conscience, of shunning responsibility, looking away, keeping mum. When the unspeakable truth of the Holocaust then became known at the end of the war, all too many of us claimed that they had not known anything about it or even suspected it...

(This quote's intermingling of the hypothesis that people *claim* to have been unaware of morally challenging facts versus the hypothesis that they were *truly* unaware is typical of many examples of such arguments, and sure to frustrate economists. The psychology discussed below in Section 6 focuses true unawareness.)

economic disputes are disputants' self-serving beliefs about what is fair.[4] Such research identifies effects that arise *only* because people care about fairness *and* are biased in their perceptions of fairness. Similarly, my model's predictions shall differ from the predictions of a model of moral dispositions as preferences, no matter how large or small a component of the utility function such dispositions were hypothesized to be.

In Section 2, I present two contrasting models of an agent's choice in deciding whether to engage in an activity that is personally beneficial but possibly causes harm to others. The preferences model assumes that moral dispositions enter directly into the utility function an agent maximizes; the constraints model assumes an agent maximizes her self interest subject to the constraint that she not engage in immoral behavior. For *given* beliefs, there isn't a big behavioral distinction between these two models of morality: For any utility function incorporating moral preferences, we can specify a *fixed-belief behaviorally equivalent* (*fbbe*) moral-constraint utility function that will generate exactly the same behavior for all beliefs.

In Section 3, however, I show that the two models of morality can lead to very different behavior when people can manipulate their beliefs: The likelihood that an agent will cause social harm is higher if she abides by moral constraints than if she were to abide by fbbe moral preferences.[5] Despite the evidence that self-serving biases violate Bayesian rationality, the model of Section 3 assumes Bayesian information processing. I use the Bayesian model because no alternative yet matches it for coherence, tractability, robustness, and marketability to economists, and to highlight

---

[4] See e.g., Loewenstein *et al* (1993) for a discussion of the role of self-serving biases in legal disputes; and Babcock and Olson (1992), Thompson and Loewenstein (1992), and Babcock *et al* (1995) on labor disputes.

[5] A legal analogy may be apt: Consider a company owned by somebody who, in addition to pursuing profits, directly prefers to minimize the probability that the goods she sells will cause birth defects. Contrast this with the case where the firm's owner does not directly care about birth defects, but is subject to a law stating that firms must not knowingly sell hazardous goods. Then, similarly to the results of Section 3 below, even if the threshold for tolerable level of risk were identical for the two firms, the "constraints" firm will be more likely to sell a dangerous product than the "preference" firm. This is because when the firm happens to verifiably anticipate a low probability of danger, the preference firm might still gather further information, but the constraints firm won't.

the role that self-serving biases can play *even if* people were Bayesians. The "bias" in moral cognition is not a statistical bias, but rather a selection bias: An agent will choose the informational signal that maximizes her probability of ending up with beliefs that allow her to pursue her self interest, rather than pursue a "disinterested" strategy of acquiring all the (low-cost) information available.

In Section 4, I discuss an array of non-coercive "informational interventions" by outside parties that the model predicts can increase the morality of conduct. In standard theory, if people avoid cheap information, it won't affect their behavior if they get it. This is not so if people operate under moral constraints--unwelcome information can alter behavior. This suggests many forms of social influence that standard theory says won't affect people's behavior. The simplest of these is what I call *salience injection*--society may seek to force people to think about issues and receive information that they don't want to think about. A second form of influence is more subtle. While *forcing* people to receive more information always encourages better calibrated moral conduct, merely making more information *available* to people can worsen things, because the information will likely be acquired self-servingly and selectively. There is, therefore, a case to be made for *moral dogmatism*--because discretionary use of information invites self-serving biases, encouraging too much free thought may worsen moral conduct. Finally, the model implies that society has an interest in getting people to develop their moral judgments *before* they are motivated to distort their thinking. I call this *moral priming*--society improves social welfare simply by encouraging thought by people about moral principles before they anticipate they are likely to apply these principles. This can work because people won't always, when their self interest demands it, be able to undo their earlier disinterested judgments. All the forms of social influence would not be necessary nor have the desired effect if morality were merely a matter of preferences in the conventional sense.

The paper's model assumes that an agent manipulates her beliefs in a way that either doesn't bother her, or about which she is unaware. Given sufficient awareness of how she is manipulating her beliefs, an agent may demand of herself that not only her actions but also her information

processing be morally optimal.[6] In Section 5, I show that a simple model of full moral accountability in information processing leads behavior along the lines of the standard preferences model. Only by positing some plausible limits on the degree to which we monitor our belief manipulation do the qualitative results of Section 3 still hold.

To argue that such limits, and the other the assumptions of my model, are plausible, I briefly discuss related psychological research in Section 6. Though no one line of psychological research has conceptualized the issues in precisely the way I have, many different research programs have investigated related phenomena. In Section 7, I briefly discuss some of the model's shortcomings. I then conclude by considering how the model might apply to the study of self control and social contagion in moral judgment.

## 2. Moral Preferences and Moral Rules

Consider an agent choosing whether or not to engage in a pleasurable activity. She chooses $x \in \{0,1\}$, where $x = 1$ means she engages in the activity and $x = 0$ means she does not. The activity brings enjoyment $V(x)$ to her. Assume $V(0) \equiv 0$ and $V(1) \equiv v \in (0,1)$, where $v$ is a parameter measuring how enjoyable the agent finds this activity. An agent can eat grapes or not eat grapes, and eating grapes is enjoyable.

But this person may not want to engage in the activity if she believes it hurts others.[7] If she believes that farm workers are exploited in the production of grapes, then she doesn't want to buy grapes. Let $W(x)$ be a measure of the social harm caused by the activity. For simplicity, I assume

---

[6] The legal analogy is again apt: Negligence laws do *not* focus solely on actual beliefs by injurious parties at the time they make decisions. They also hold parties responsible for reasonable effort to learn relevant information before taking actions. A company will be deemed guilty of negligence if it can be proven that it purposely avoided available information that its product was harmful. (But closer to the main model, Mukerjee (1995, p. 22) notes that companies are not required to perform tests on all potentially harmful chemicals they expose their workers to, but they *are* required to report to OSHA any knowledge they have about potential harm.)

[7] I focus on the case where the activity is potentially socially harmful; nothing would be conceptually different if we considered the alternative case of an unpleasant activity that might be socially beneficial.

that the activity either causes harm--W(1) = 1--or does not cause harm--W(1) = 0. I assume that not engaging in the activity never causes harm; W(0) = 0. There is uncertainty about whether the activity causes harm.[8] Assume the agent believes that the activity causes harm with probability $q$; thus, W(1) = 1 with probability $q$ and W(1) = 0 with probability $1-q$. Throughout this section, I assume that $q$ is fixed.[9]

If an agent maximizes her expected utility that includes both her own enjoyment and potential social harm, then she will choose $x \in \{0, 1\}$ that maximizes $V(x) - q \cdot W(x)$.[10] I shall call such an agent a *preferences* agent. Given previous assumptions, this utility function can be re-written:

Definition 1:

An agent has *Moral Preferences* (she is a *P* agent) if she maximizes the Von Neumann-Morgenstern utility function $U_p(v,q)$, where:

$$U_p(v,q) = \begin{cases} v - q & \text{if } x = 1 \\ 0 & \text{if } x = 0 \end{cases}$$

---

[8] Though there is uncertainty regarding W(·), I assume no uncertainty regarding the value of V(·). This is inconsequential--V(·) can readily be interpreted as expected pleasure and yield the same predictions. (Of course, focusing only on "expected pleasure" means I am treating uncertainty over personal consequences differently than social consequences. But the hypothesis underlying my model is that people *do* treat uncertainty over social consequences differently than uncertainty over self-interested consequences.)

[9] This probabilistic model may seem an awkward fit for most moral dilemmas. Many times our moral decision-making does not really concern probabilistic assessments regarding the effects of an action, but rather the process of evoking whether it is "right" or "wrong" to do the action. If you have found $100 on the street, and are trying to decide whether to turn it in, part of your dilemma is figuring out the likelihood that the owner will be found, that the local police officials are corrupt, etc.; such cases are captured by this model. But part of the moral decision-making may simply be thinking through whether or not you have a moral obligation to turn it over. While the language throughout suggests physical uncertainty, much can be interpreted in terms of uncertainty about the appropriate moral criteria. Also, while the entire paper deals with the probabilistic model and is framed in terms of preferences, analogous results could readily be derived from a (psychologically more realistic) non-probabilistic model that departs more radically from standard economic formulations.

[10] $W(x)$ represents the agent's own weight placed on social harm done, and need not correspond to the "true," utilitarian-compatible social harm.

For notational convenience, I do not include the variable $x$ as an argument in the utility function. For all utility functions I consider in this paper, $U(0) = 0$, so little is lost by suppressing $x$ as an explicit argument, and I shall mean the $x = 1$ case when referring to the utility function. The above specification implies that the agent will engage in the activity if and only if $q \leq v$.[11] That is, she will engage in the activity if and only if the probability of social harm is less than or equal to the personal benefit she gets from engaging in the activity.

This specification of the utility function treats a person's concerns about the possible social harm within the standard, expected-utility framework. I turn now to an alternative specification of how the perceived social consequences of an action might influence an agent's utility. Instead of maximizing social surplus, suppose an agent maximizes her pleasure subject to the constraint that she not engage in an action that is too likely (by some standard) to harm others. I call an agent thusly motivated a *rules* agent, formally represented as follows:

Definition 2:

An agent is following a *Moral Rule* (she is a $R$ agent) if there exists $y > 0$ such that she maximizes the utility function $U_R(v,q)$, where:

$$U_R(v,q) = \begin{cases} v - g(q) & \text{if } x = 1 \\ 0 & \text{if } x = 0, \end{cases}$$

where $g(q) = 1$ if $q > y$ and $g(q) = 0$ if $q \leq y$.

That is, if the probability that her activity causes social harm is high enough, the agent does not engage in the activity.[12] If the probability of social harm is not deemed so high, she gets full pleasure from the activity.

---

[11] For simplicity, I will assume here and elsewhere that an agent chooses $x = 1$ rather than $x = 0$ whenever indifferent. This guarantees that the optimal choice of $x$ will always be determinate. When considering information acquisition, I shall also assume that, when indifferent, an agent acquires the level of information that maximizes her probability of engaging in the activity.

[12] The assumption that the utility drops below zero if the agent attributes a high enough probability of social harm merely guarantees that she will not engage in the activity in such games. Specifying any $g(q) > v$ would therefore be equivalent to this definition.

This reflects the difference between an agent guided by moral rules rather than moral preferences. But despite this difference, there is nothing in the alternative specification to guarantee that *behavior* will be different. Indeed, the two different specifications of the utility function can yield identical behavior in the following sense:

Definition 3:

The utility functions $U_P(v,q)$ and $U_R(v,q)$ are *fixed-belief behaviorally equivalent* (*fbbe*) if, for all $q$ and $v$, $\text{argmax}_x\ U_P(v,q) = \text{argmax}_x\ U_R(v,q)$.

That is, if we knew an agent's beliefs, but did not know whether her preferences were $U_P(\cdot,\cdot)$ or $U_R(\cdot,\cdot)$, we would still always make the same prediction about behavior so long as these utility functions were fbbe. The condition that makes a P agent fbbe to a R agent is clear:

Lemma 1:

$U_P(v,q)$ and $U_R(y,q)$ are fbbe iff $y = v$.

The lemma says that a R agent will behave like a P agent for all possible beliefs if and only if her moral rule is "don't do anything that isn't defensible from the perspective of social welfare." The behavioral equivalence is therefore obvious, guaranteed by assumption. Comparing the implications of moral rules to their fbbe moral preferences will be the focus of much of the paper, and from here on in I will assume that the moral-rule utility function is based on $y = v$. With this assumption, these two different types of agent will choose to take the action under exactly the same conditions: when $q \leq v$.

## 3. Belief Manipulation

The previous section specifies two models of preferences, "moral preferences" and "moral rules", that yield the same behavior for all beliefs. I now analyze what these different utility functions imply about how the utility-maximizing agent will manipulate her beliefs. To model the information structure, let $\mathcal{F}$ be the set of all possible probability distributions over $q$, where $q$ is the agent's probabilistic belief that her activity will cause social harm. Let $f(q)$ denote the probability assigned by distribution $f$ to

beliefs $q$.[13] Let $c:\mathcal{F} \rightarrow \mathbb{R}$ be a function determining the agent's costs of belief manipulation, where $c(f) \geq 0$ is the cost of generating the probability distribution $f \in \mathcal{F}$. The function $c$ fully characterizes the agent's pre-decision information structure.[14] The agent's initial beliefs, $q_0$, are part of the information structure $c$. This is captured by assuming that $c(f_0) = 0$, where $f_0$ is the probability distribution putting full weight on the beliefs $q_0$. I shall denote the fully informative signal by $f^*$, where $f^*(0) = 1-q_0$ and $f^*(1) = q_0$. Denote by $\mu(f)$ the mean of distribution $f$.

The agent first chooses the signal $f \in \mathcal{F}$ she wishes to collect. The signal she chooses then probabilistically generates beliefs, after which the agent chooses the optimal action given the realized beliefs. I denote the agent's optimal choice of whether to engage in the activity, given her beliefs $q$, by $x(q) \in \{0,1\}$. Recall that, for both types of agents, $x(q) = 1$ iff $q \leq v$.

What is an agent's overall utility function given such cost functions? I assume that these costs are additively separable components of the utility function, and are measured relative to the benefit of the activity. For a P agent, this means that her overall utility function will be as follows:

$$U_P(f;c) \equiv \int_q [x(q)\cdot(v-q)]f(q)dq - c(f).$$

Because $x(q)$ is zero if the agent does not engage in the activity, the term $x(q)\cdot(v-q)$ is a gimmick to take into account both cases where the activity is engaged in and when it is not. A P agent will then choose $f$ to maximize this utility function. What will be the utility function for a R agent? Her utility function will be:

$$U_R(f;c) \equiv \int_q [x(q)\cdot v]\cdot f(q)dq - c(f).$$

---

[13] We can also make $f(q)$ the density function; nothing depends on assuming that all signals have finite support set. I will often be sloppy with notation; there should be no ambiguity.

[14] Costs of acquiring signals can be interpreted as costs of cognition, rather than literal costs of acquiring "external" information. For example, the agent's costs of changing her beliefs may reflect her decision about how hard to think about an ethical decision, rather than phone bills or payments for books.

Note that in extending the utility functions characterized in the previous section, I am incorporating comparisons of utility *across* different belief states. Because many of my results are precisely about differentiating agents' taste for information acquisition, assumptions about how an agent feels about different beliefs are clearly crucial.

The agent's overall decision is to first choose a signal $f$ and then, after getting information based on this signal, to decide whether to engage in the activity.[15] To guarantee that the agent's "belief manipulation" is Bayesian, I assume $c(f) < \infty$ only if $\mu(f) = \mu(f_0)$. Because $f_0$ puts probability 1 on beliefs $q_0$, this means that the agent's expected beliefs (no matter her belief-manipulation strategy) at the time of decision are $q_0$.

Before providing precise formulas for the beliefs and behavior of the two types of agents, I define convenient notation. Let $x_P(v,c)$ and $x_R(v,c)$ be the probabilities of engaging in the activity for, respectively, a P and fbbe R agent who behaves optimally. (Note that a given information structure $c$ implies some initial beliefs $q_0$; I shall omit $q_0$ from notation whenever a specific $c$ is specified.) Let $z_P(v,c)$ and $z_R(v,c)$ be the probabilities that social harm will be realized. Let $x(f)$ be the probability of engaging in the activity generated by the signal $f$, and let $z(f)$ be the probability of causing social harm generated by the signal $f$. Note that it makes sense to define the notation $x(f)$ and $z(f)$ independent of whether we are considering a P agent or R agent, precisely because we are examining fbbe preferences: Once we know the distribution of beliefs that an agent might have, a P and R agent will engage in the activity and cause harm with the exactly the same probabilities. When it causes no confusion, I shall denote by $U_P(f)$ and $U_R(f)$ the P and R agents' expected utility generated by the signal $f$; this suppresses the information structure $c$.

Using this notation we can provide formulas for the agents' expected utilities as a function of the signal they acquire, which will help clarify the logic of the results below:

---

[15] For current purposes, little of conceptual interest is lost by assuming that the person makes a one-time choice among elements in $\mathcal{F}$, rather than obtaining information sequentially. Whatever her stopping rule in a sequential process, this will generate an element in $\mathcal{F}$, and an *expected* cost of search.

$$U_P(f) \equiv x(f)\cdot v - z(f) - c(f)$$
$$U_R(f) \equiv x(f)\cdot v - c(f)$$

These formulas make the basic difference between P and R agents clear: At the stage of belief manipulation, a P agent will take into account the social harm she will cause by choosing a signal; the R agent won't.

Within the class of Bayesian information structures, I consider first the case where the agent can fully manipulate her beliefs. In particular, suppose that an agent can generate, at no cost, any probability distribution over beliefs that involves the same expected beliefs as her initial beliefs:

Definition 4:

Initial beliefs $q_0$ are *maximally manipulable* if, for all $f$ such that $\mu(f) = q_0$, $c(f) = 0$. Denote this information structure as $c^{mm}(q_0)$.

This is the richest Bayesian information structure imaginable in this context, and will generate the sharpest contrast between moral preferences and moral rules. Note in particular that a maximally-manipulable information structure is stronger than merely assuming that the person has full information available for free. Of interest in many cases is the decision by the agent to *not* fully inform herself; a maximally manipulable information structure means the agent can cheaply gather precisely as much *or as little* information as she wishes. That this definition is worth making reflects the main lesson of the paper. For classical expected-utility preferences, when full information is available for free, the availability of less-than-full information is redundant. It is the fact that the R agent may strictly prefer less information to more information that both drives the main results of this paper and necessitates a more complete specification of the information structure.

Given a maximally-manipulable information structure, what signal will an agent choose? Clearly, a P agent would choose full information, generating beliefs $q = 0$ with probability $1-q_0$ and $q = 1$ with probability $q_0$. This result is of course a specific case of the more general fact that expected-utility maximizers always prefer more information to less. Notably, the P agent will never cause social harm.

By contrast, a R agent who could readily manipulate her beliefs will typically *not* obtain full information. For contingencies where she is allowed

to engage in the activity (i.e., when $q_0 \leq v$), she will gather no information of behavioral consequence. If $q_0 > v$, she will choose the signal that maximizes the probability that her beliefs will be exactly $v$. The R person will therefore engage in the activity with probability 1 if $q_0 \leq v$ and with probability $(1-q_0)/(1-v)$ if $q_0 > v$. This can be summarized as follows:

Proposition 1:

Suppose that the information structure is $c^{mm}(q_0)$. Then:

1) for all $q_0$, $x_P(v,c) = 1-q_0$ and $z_P(v,c) = 0$;

2) if $q_0 \leq v$, $x_R(v,c) = 1$ and $z_R(v,c) = q_0$; and
if $q_0 > v$, $x_R(v,c) = (1-q_0)/(1-v)$ and $z_R(v,c) = v\cdot(1-q_0)/(1-v)$.

Proof:

Proof of Part 1: A P agent will choose $f \in \mathcal{F}(q_0)$ that maximizes $x(f)\cdot v - z(f)$. Note first that the agent will never choose an $f$ that puts positive weight on beliefs $q \in (0,v]$, because such an $f$ is dominated by $f'$, where $f'$ replaces weight $f(q)$ with the distribution $f(q)\cdot(1-q)$ on belief = 0, and $f(q)\cdot q$ on belief = 1. This increases utility by $f(q)[((1-q)\cdot v - 0) - (v - 1)] = f(q)[1 - v] > 0$.

The agent will never choose an $f$ that puts positive weight on $q \in (v,1)$, because such an $f$ is dominated by $f'$, where $f'$ replaces weight $f(q)$ on $q$ with $f(q)\cdot(1-q)$ on belief = 0, and $f(q)\cdot q$ on belief = 1. This raises $x(f)$ without changing $z(f)$, and thus increases utility. Thus, the only signal possibly chosen is the fully revealing one.

Proof of Part 2: A R agent simply chooses the $f$ the maximizes the probability that $x(f) = 1$, which means she chooses an $f$ that maximizes $F(v)$, where $F(\cdot)$ is the cumulative distribution corresponding to $f$. If $q_0 \leq v$, she chooses a signal with the property $F(v) = 1$. These all yield $x(f) = 1$ and $z(f) = q_0$. If $q_0 > v$, then $F(v)$ is maximized by $f$ such that $f(v) = (1-q_0)/(1-v)$ and $f(1) = (q_0-v)/(1-v)$.

This proves the proposition. Q.E.D.

Thus, if beliefs are maximally manipulable, a R agent will be more likely to engage in the activity than a P agent, and will be more likely to cause harm. What can one say about the differences in behavior of R and P agents for more general information structures? It turns out that we cannot fully replicate Proposition 1--there exist information structures where a P agent is

*more* likely to engage in the activity than a R agent. Consider the following situation. An agent's initial beliefs make her (just barely) unwilling to engage in an activity. She could at low cost acquire a little bit of information that would with some probability make her (just barely) willing to engage the activity. Or, at much greater expense, she could learn for sure whether the activity will cause harm. To consider a specific formal example, let $q_0 = .5 + \varepsilon$, where $\varepsilon > 0$ is very small. There are two possible signals available at finite cost: $f^*$ yielding beliefs $q = 0$ with probability $.5 - \varepsilon$ and $q = 1$ with probability $.5 + \varepsilon$, and $f'$ yielding beliefs $q = .5 - 2\varepsilon$ with probability 1/4 and beliefs $q = .5 + 2\varepsilon$ with probability 3/4. Assume $v = .5$. Let $c(f') = 0$ and $c(f^*) = .2$.

To figure out what each of the two agents would do, note first that her utility is 0 if she does not acquire either signal. If the P agent acquires the signal $f^*$, then her expected utility will (ignoring the $\varepsilon$'s) be $.5 \cdot v - c(f^*) = .25 - .2 = .05$. If she acquires $f'$, her expected utility will be $.25 \cdot (v - .5) = 0$. A P agent will acquire signal $f^*$. The R agent's expected utility from acquiring $f^*$ is .05, the same as the P agent. By contrast, the expected utility for the R agent of acquiring the signal $f'$ is $.25 \cdot v = .125$. The R agent will therefore acquire signal $f'$, and thus *less* likely than the P agent to engage in the activity.

Why does the P agent prefer the signal $f^*$ to $f'$ while the R agent prefers $f'$ to $f^*$? The difference is that the P agent has a direct taste for decreasing the likelihood of social harm done by her activity; because $f^*$ eliminates the possibility that she will do harm, it is attractive to her above and beyond the benefits of increasing the probability of engaging in the enjoyable activity. Because the R agent only cares about increasing the probability that she will be morally comfortable with engaging in the activity, she will pay for extra information only to get the probability of harm low enough to justify her behavior. Thus, the additional cost of $f^*$ beyond $f'$ is worth it to the P agent but not the R agent.[16] Proposition 1 and the example explore both the likelihood that the agent engages in a morally dubious activity and the

---

[16] Note, however, that if costs for the two signals were the same, then the R would also acquire signal $f^*$. This fact will be true in general, and suggests a generalization of Proposition 1 to the class of simple information structures where all obtainable signals cost the same. This result was formally presented in a previous draft of the paper.

likelihood that her activity causes harm. While the likelihood of engaging in the activity is of concern for certain issues, for other issues the likelihood of *harm* done is of more interest. In the example provided, while the P agent is more likely to engage in the activity than the R agent, she is *less* likely to cause harm--she is, in fact, certain to cause no harm. Intuitively, the P is more likely to engage in activity only because she is also willing to pay more for information that will reduce the likelihood of causing harm. Indeed, it turns out that for *any* cost structure, the probability that the P agent causes harm is lower than the probability that the R agent causes harm:

Proposition 2:

For all $v$ and $c$, $z_P(v,c) \leq z_R(v,c)$. Moreover, there exist cases where the inequality is strict.

Proof:

I show that, given any $v$ and $c$, if a P agent prefers $f_1$ to $f_2$, but a fbbe R agent prefers $f_2$ to $f_1$, then it must be that $z(f_1) < z(f_2)$. This will prove the result, because it holds for the case where $f_1$ is the P agent's optimal signal and $f_2 \neq f_1$ is the R agent's optimal signal, in which case $z(f_1) = z_P(v,c)$ and $z(f_2) = z_R(v,c)$. But the proof of the first sentence is straightforward:

$$\begin{aligned} U_P(f_1) - U_P(f_2) &\equiv [x(f_1)\cdot v - z(f_1) - c(f_1)] - [x(f_2)\cdot v - z(f_2) - c(f_2)] \\ &\equiv [x(f_1)\cdot v - c(f_1)] - [x(f_2)\cdot v - c(f_2)] + [z(f_2) - z(f_1)] \\ &\equiv U_R(f_1) - U_R(f_2) + [z(f_2) - z(f_1)] \end{aligned}$$

This means that $U_P(f_1) - U_P(f_2) \geq 0$ and $U_R(f_1) - U_R(f_2) < 0$ can simultaneously hold only if $z(f_2) - z(f_1) > 0$. Q.E.D.

Along with Proposition 4 below, I feel Proposition 2 is the crux of this paper. Because the P agent is a conventional expected-utility maximizer who cares directly about the social harm she might be causing, she is always eager for information that can help her avoid causing social harm. By contrast, a R agent is averse to information about possible social harm whenever it risks denying her the moral green light to engage in an enjoyable activity.

The next proposition shows that both types of agents accord to the straightforward intuition that the more an agent enjoys an activity, the more likely she is to engage in it. An examination of Proposition 1 shows that this is the case for the maximally-manipulable information structure. Proposition 3

shows that it is true generally:

Proposition 3:

For all $c$ and all $v_1 > v_2$,

1) $x_P(v_1,c) \geq x_P(v_2,c)$.

2) $x_R(v_1,c) \geq x_R(v_2,c)$.

Moreover, there exists cases where both of the inequalities are strict.

Proof:

I prove part 1; the proof of part 2 is nearly identical. I show that, given any $c$, if a P agent with $v_1$ prefers $f_1$ to $f_2$, but a P agent with $v_2$ prefers $f_2$ to $f_1$, then it must be that $x(f_1) > x(f_2)$. This will prove the result, because it holds for the case where $f_1$ is the $v_1$ agent's optimal signal and $f_2 \neq f_1$ is the $v_2$ agent's optimal signal, in which case $x(f_1) = x_P(v_1,c)$ and $x(f_2) = x_P(v_2,c)$. But the proof of the first sentence is straightforward. For any $v$, consider $\Delta(v)$, where

$$\Delta(v) \equiv U(f_1) - U(f_2) = [x(f_1)\cdot v - z(f_1) - c(f_1)] - [x(f_2)\cdot v - z(f_2) - c(f_2)]$$
$$= v\cdot[x(f_1) - x(f_2)] + [z(f_2) - z(f_1)] + [c(f_2) - c(f_1)]$$

In order for $\Delta(v_1) \geq 0$ but $\Delta(v_2) < 0$, clearly we need $\Delta(v_1) - \Delta(v_2) > 0$. But $\Delta(v_1) - \Delta(v_2) = [v_1 - v_2]\cdot[x(f_1) - x(f_2)]$. Since $v_1 > v_2$, this means that $x(f_1) - x(f_2) > 0$. Q.E.D.

I have found no general characterizations of how the likelihood that either type of agent will engage in the activity depends on the costs of gathering information or on the initial beliefs $q_0$. But some comparative statics of interest for maximally-manipulable information structures follow from Proposition 1. For a P agent, increasing $q_0$ decreases the probability that the agent will engage in the activity, and does not affect the probability of harm (because the P agent will never cause harm). For the R agent, if $q_0 < v$, the agent will for sure engage in the activity, so that slightly increasing $q_0$ will have no effect on her behavior and will directly increase the probability of harm done. If $q_0 \geq v$, however, increasing $q_0$ decreases both the probability of the activity and the probability of harm done. This is because making $q_0$ higher decreases the likelihood that a R agent can generate beliefs $q \leq v$, thus decreasing the likelihood that she will engage in the activity.

Finally, the model helps identify potential "behavioral

misidentifications." Suppose that economists have wrongly been assuming that people are P agents, when they are really R agents. Now suppose that we have evidence on people's belief-contingent behavior, because we have observed them in situations where beliefs are obvious. By Proposition 3, we know that a R agent will behave less morally than the fbbe P agent. Therefore, if we wrongly perceive that morality is governed by preferences, observing belief-contingent behavior leads us to exaggerate the degree to which moral considerations influence behavior.

Such misidentification might apply to economists' inference from laboratory experiments. Many experiments test whether subjects' are influenced by a concern for others. The best of these experiments strive for setting up clear and blatant tradeoffs, where standards of fairness are obvious; this allows clearer hypothesis testing and verifies that people do not solely seek to maximize their own payoffs. But most real-world situations are much more ambiguous, allowing people immense flexibility in interpreting the moral demands on them. Morality may play less of a role in the real world than in the laboratory. This effect of ambiguity would not be predicted by the standard model, because there is no reason to suppose that on average uncertainty makes pure expected-utility maximizers either more moral or less. In the rules model, it has a predictable effect.

A second, contrary type of misinterpretation can occur when we observe people's behavior in settings where *a*) we are pretty sure that good information is available, but *b*) do not directly observe people's actual beliefs. Because economists suppose that people always gather cheap information that might influence their behavior, we assume that intentionally ignored information will have little influence on behavior if people are forced to learn it. But it may be that if information about harm done were more salient or unavoidable, the agent would not engage in the activity.[17]

---

[17] In principle, misinterpreting laboratory evidence may reflect this form of misidentification, and lead us to underestimate the role of moral concerns in real-world settings. It could be that laboratory settings are perceived by subjects as being so artificial that they do not know how to apply any familiar norm of fairness. But in the type of allocation decisions typically tested, I think the first type of misidentification is more likely. Most economics experiments have subjects participate in choices that have solely monetary consequences. Money is relatively unambiguous, and may not provide the scope for self-serving biases that many real-world decisions allow.

Indeed, some of the issues explored in the next section are very related to this fact that people may behave very differently when information is unavoidable than they do when it is merely available.

## 4. Moral Constraints and Social Influence

In this section, I consider "informational interventions" that can influence the moral conduct of people operating under moral constraints, focusing on interventions that function very differently than if people had moral preferences. The reader is invited to read this section for its positive rather than normative implications: All results pertain to ways third parties can influence behavior, whether or not we perceive the changes in behavior as good or bad. But I shall discuss the examples as if reducing the probability of social harm is a good thing. Reducing the probability of the enjoyable activity is a bad thing, of course. Overall the first best may be to get the agent to behave as she would if she were fully aware of the consequences of her actions. But even fully informed people are surely too self-interested, so it may be ideal to reduce an activity even further for classical externalities reasons. When it makes a difference, I will consider social welfare both with and without putting weight on the person's personal enjoyment. I will also focus only on interventions by outside parties who are no better informed than each agent about the social consequences of her actions. In this sense, any potential welfare improvements show that zero-influence outcomes are not "constrained efficient"--they can be improved upon by outside parties with no superior knowledge.

The first and simplest informational intervention is what I call *salience injection*: Society may usefully seek to make as salient as possible the social consequences of choices people make. To distinguish this from "merely" providing information, I focus on the case where people could readily obtain the relevant information, but, because their current beliefs allow them to pursue their self interest, they choose not to. Outside parties may then serve a role in making already *available* information *unavoidable*. For instance, society might pass a law providing ample, hard-to-miss space on packages of all consumer products for various groups to make (truthful) claims about the consequences of buying those products. When people are motivated to avoid awareness of something, such in-your-face sharing of information may have an

effect even if it is merely pushing information that was already readily available. On a smaller scale, you as a concerned bystander may try to call attention to relevant issues when observing morally relevant decision making by those around you.

Again I emphasize that the role of salience injection derives from the model of moral constraints. If people are guided by moral preferences, then so long as information is available, we would not need to worry about forcing them to process this information. If people are simply amoral, forcing information on them doesn't help with anything they will pursue their self interest in any event. But if people are guided by moral constraints, then being very forceful in imposing the truth on them can be productive, even if we do not force their behavior in any way and even if the information was already available to them.[18]

To formalize salience injection, I extend the framework of the previous

---

[18] The special role of salience in morally relevant decision-making has been posited by others. In his famous study of obedience to authority, Milgram (1974, p. 6-7) argues (more extremely than I have) that socially-mediated salience plays a huge role in determining the extent to which people follow their moral ideals:

> The force exerted by the moral sense of the individual is less effective than social myth would have us believe. Though such prescriptions as "Thou shalt not kill" occupy a pre-eminent place in the moral order, they do not occupy a correspondingly intractable position in human psychic structure. A few changes in newspaper headlines, a call from the draft board, orders from a man with epaulets, and men are led to kill with little difficulty. Even the forces mustered in a psychology experiment will go a long way toward removing the individual from moral controls. Moral factors can be shunted aside with relative ease by a calculated restructuring of the informational and social field.

Milgram's argument that the "calculated restructuring of the informational and social field" can hurt moral conduct emphasizes how outside parties might be in the business of manipulating moral cognition. Indeed, he argues later (p. 122) that nefarious parties are well aware of how to manipulate salience for evil purposes:

> Any competent manager of a destructive bureaucratic system can arrange his personnel so that only the most callous and obtuse are directly involved in violence. The greater part of the personnel can consist of men and women who, by virtue of their distance from the actual acts of brutality, will feel little strain in their performance of supportive functions.

section. Consider an agent who starts with initial beliefs $q_0$. Recall that $f_0$ is the signal where the agent does not update her beliefs from $q_0$, and $f^*$ is the acquisition of full information, and $c(f^*) \geq 0$ is the cost to the agent of acquiring full information. In Section 3, I assumed that $c(f_0) = 0$, so that the cost to the agent of maintaining her initial beliefs $q_0$ was always 0. But if information can be made unavoidable, this raises the possibility that $c(f_0) > c(f^*)$. A agent guided purely by preferences would always choose $f^*$ if $c(f_0) \geq c(f^*)$, and even if $c(f_0)$ were a little less than $c(f^*)$. But a constraints agent will sometimes choose to remain ignorant even if $c(f^*) < c(f_0)$. Salience injection would then involve making the information as unavoidable as possibly by raising $c(f_0) - c(f^*)$ as high as possible. Extreme in-your-face unavoidability is when $c(f_0) - c(f^*) = \infty$.

The merits of salience is folk knowledge among those trying to raise money for charity. Making the value of donations especially salient--by showing the picture of the child you would save--may have especially pronounced effects, even when the pictures do not show any information that a contributor could reasonably not have available to her. At any moment we choose, we could all make available to ourselves the knowledge that, without fundamentally sacrificing our standard of living, we could save the life of a child in poverty. There is a wedge between our actual day-to-day, self-versus-other allocation choices and what we would choose if confronted with the tradeoffs more starkly. And it is not mere forgetfulness that prevents us from more often having an awareness of how our donations can help. We could even, if we chose to, conjure up our own powerful image of a child in need. We don't often do so on our own. There are many reasons why we dislike thinking about suffering, and I do not surmise that the specific mechanism presented in this paper predominantly drives this phenomenon. But something of great social consequence seems to be going on: I contend that the market for good deeds simply does not seem to conform well to a simple model of moral preferences, because even participants who would "trade" actively if participating in the market are very reluctant to enter the market.[19]

Of course, the same argument that says that making information

---

[19] Some speculations along these lines are presented in Freeman's (1993) conference draft for discussion, though he conjectures that social pressure may play a dominant role.

unavoidable is useful says that merely making information available likely *won't* be that useful. If people without information feel free to pursue their self interest, they will likely avoid additional information made available to them. Thus, low-key attempts by concerned parties to educate people about the consequences of their conduct will be less effective than if people were governed by moral preferences. To use a topical example, if people don't want to be aware that their conduct is likely to spread AIDS to other people, simply making educational materials available may not have much of an effect. And this can be true even if we correctly believe that few people would *knowingly* expose others to AIDS.[20,21]

---

[20] See Kremer (1995) for a hypothesis along these lines.

[21] All the above arguments ignore the more general importance of salience, even when people are not the least bit reluctant to know more. Psychological evidence indicates that people are remarkably influenced by how salient information is relative to how informative it is; we are often more likely to be influenced by a graphic story of bad luck with a particular make of car than with much more informative statistics. See Nisbett and Ross (1980) for a good summary of this phenomenon. My hypothesis here is that salience is *especially* important in the context where somebody is motivated to avoid the relevant information. The difference between making information available and making it unavoidable is greater in (say) inducing people to donate to charities than in getting them to buy consumer products. This is merely a hypothesis--I am not familiar with research making such comparisons.

Milgram's book cited above is laden with arguments that suggest he thinks the power of salience in the moral realm comes largely from people's desires to avoid moral responsibility. For instance, his choice of the word "denial" in the following passage (p. 38) indicates that he feels that *if people were so motivated*, they could see the true moral implications of their behavior:

> *Denial and narrowing of the cognitive field.* The Remote condition allows a narrowing of the cognitive field so that the victim is put out of mind. When the victim is close it is more difficult to exclude him from thought. He necessarily intrudes on the subject's awareness, since he is continuously visible. ... In the Proximity conditions his inclusion in the immediate visual field renders him a continuously salient element for the subject. The mechanism of denial can no longer be brought into play.

Milgram's perspective that people crawl through even very small windows of ambiguity to absolve themselves of moral responsibility is not based on the view that people are simply unconcerned about the morality of their conduct (p. 41):

> Subjects have learned from childhood that it is a fundamental breach of moral conduct to hurt another person against his will. ... It is clear from the remarks and behavior of many participants that in punishing the

Indeed, encouraging people to think about their moral conduct may backfire, by facilitating self-serving biases. Discouraging free thought may be best. I call this *moral dogmatism*: Indoctrinating people with socially-imposed moral dictums could produce more moral behavior than allowing people to think through for themselves the moral consequences of their actions, even if the dictums prevent people from accessing better information about the consequences of their actions.

Before considering this possibility further, I present a final general proposition that clarifies a relevant and stark distinction between moral preferences and moral constraints. Proposition 4 shows that while "greater flexibility of thought" decreases the likelihood of social harm when people have moral preferences, it *increases* the likelihood of social harm when they are governed by moral constraints. While I have found no general definition for "greater flexibility", Proposition 4 formalizes this intuition when comparing the *most* flexible belief structure--maximally manipulable beliefs--to any other information structure:

Proposition 4:

For all $v$ and $c$ with initial beliefs $q_0$:

1) $z_P(v, c^{mm}(q_0)) \leq z_P(v, c)$; and

2) $x_R(v, c^{mm}(q_0)) \geq x_R(v, c)$ and $z_R(v, c^{mm}(q_0)) \geq z_R(v, c)$.

Moreover, there exists cases where all of the inequalities are strict.

Proof:

The result follows from the proof of Proposition 1. In particular, we know that $z_P(v, c^{mm}(q_0)) = 0$, so that part 1 holds automatically. A R agent who begins with $q_0 \leq v$ will engage in the activity for sure, no matter what the cost structure. If $q_0 > v$, then a maximally-manipulable information structure will lead a R agent to seek out the signal that is both most likely to permit her to engage in the activity and to leave her with the highest probability of

---

victim they were often acting against their own values. Subjects often expressed disapproval of shocking a man in the face of his objections, and others denounced it as stupid and senseless. Yet many followed the experimental commands.

See also his discussion of "avoidance" in Chapter 12.

causing harm by engaging in the activity. For any $q_0 > v$. and any cost structure that makes such a signal very expensive, all the inequalities are strict.

Q.E.D.

Proposition 4 directly implies that the morality of conduct can be undermined if people are too open-minded and flexible in their moral reasoning. Of course, open-minded contemplation of moral issues is a good thing if it helps people sacrifice for others only when it is wise to do so. But the self-serving bias may outweigh such benefits. To illustrate this issue, I develop a model comparing a "liberal" religion to a "dogmatic" one.

Suppose a person choosing whether to give a dollar to a pauper can be guided in her decision by one of two religions, A or B. By the standards of *both* religions, one half of all paupers encountered are morally deserving of getting a dollar; one half don't deserve assistance. Giving to the first type yields a positive social benefit of 4 utils; giving to the second type yields a social *cost* of 1 util. The person with the money values the dollar at 1 util. As a reference point, we can calculate the expected social benefits (as viewed by both religions) of giving to a pauper when an agent has no information as .5(4) + .5(-1) = 1.5. Since this is greater than the benefits of keeping the dollar (1 util), people will always donate when uninformed.

Now consider the "information structure" the two religions arm people with in entering each new interaction with a pauper. Religion A arms people with a "dogmatic" information structure: "Half the people are needy, half are not. You are to judge a person *solely* by some designated, easily observed signal (e.g., is the pauper male or female) that is known to be correlated with the deservingness." The signal says either "deserving" or "undeserving", each signal is accurate proportion $\gamma \in [.5,1]$ of the time, and each signal appears half the time.[22] If the exogenous parameter $\gamma = 1$, the signal is perfect; if $\gamma = .5$, it provides no information. If $\gamma < .6$, adherents to Religion A will always refrain from engaging in the activity. Of greater interest is if $\gamma \geq .6$; here, if an agent gets the signal "deserving", she will

---

[22] It does not matter in this example whether adherents to Religion A are forced to accept the signal--they (weakly) prefer to do so, because their initial beliefs don't permit pursuit of self interest. The dogmatism here is that adherents are not permitted to gather any different information instead.

give to the pauper, but will refrain if she gets the signal "undeserving". Expected social welfare will be $.5(1) + .5(\gamma \cdot 4 + (1-\gamma)(-1)) = 2.5 \cdot \gamma$.

Religion B is more "liberal". It preaches: "Use your judgment. Look for any relevant cues as to deservingness. For instance, is there anything in the person's manner or appearance (does s/he appear to be drunk) that would suggest lack of true need? You be the judge; take into account as many relevant factors as possible." I will assume that "using judgment" here generates a maximally manipulable information structure. That is, liberated to think for themselves, adherents to Religion B can always figure out all the relevant information--but they are also free not to notice the signal that adherents to Religion A automatically gather. An adherent to Religion B will try to maximize the probability that her beliefs will be below .4, so that she won't feel obligated to donate. This means that with probability .2 she will donate the money, having figured out the pauper was deserving. With probability .8, she will keep the money, having determined that donating yields exactly the same expected social benefits as keeping the dollar. Overall expected social benefits will be $.8(1) + .2(4) = 1.6$.

By comparing $2.5\gamma$ to 1.6, we see that Religion B will yield a better social outcome than Religion A if $\gamma < .64$.[23] Essentially, the "dogmatic" Religion A isn't providing adherents with enough information to make a wise choice. But in the opposite case--when $\gamma > .64$--Religion A is more effective than Religion B. This is because now Religion A's information is not so superior, and the social harm allowed by the flexibility undermines the improvement. Recall, the two religions have exactly the same moral aspirations, but happen to be employing different "dogma strategies". *By Religions A's standards*, the issues that Religion B adherents are encouraged to think about are morally relevant. But the leaders of Religion A are, in their way, more pragmatic than the leaders of Religion B, because they recognize that encouraging taking into account relevant factors invites adherents to do so self-servingly.

---

[23] Proposition 4 quickly tells us that for all $\gamma$ the probability of social harm is greater in Religion B than in Religion A. In this sense, the dogmatic religion is unambiguously better. The liberal thinking is in this case only better insofar as adherents' enjoyment from keeping the dollar is a concern; Religion B has the merits of encouraging a better-calibrated decision before giving the dollar away.

I now turn to an informational intervention that may play a role in society's large investment in moral indoctrination and education. I call it *moral priming*: Society has an interest in encouraging people to think through appropriate moral criteria before they know whether and how they have to apply those criteria. If you get people to think through moral issues when they do not percieve it as likely that the issue will affect their self interest, they are likely to think them through in a relatively disinterested way. Then if the situation *does* arise, a person who has reached certain moral judgments may not readily be able to reverse her opinion to fit her self interest.

Teaching people to think about situations that might not arise can, of course, be wasteful. When disinterestedness is not an issue, we don't generally encourage such thought. The lower the probability that a situation will arise, the less sense it makes to expend energy and resources on getting people to figure out their behavior beforehand.[24] But when it comes to moral considerations, getting people to think about their behavior in low-probability contingencies may be *ideal*--it may make *more* sense than getting them to think about high-probability scenarios. With unlikeliness, we get disinterestedness; with disinterestedness, we get better moral judgment.

This perspective can be formalized by again considering the case where $c(f^*) < c(f_0)$, but interpreting things a bit differently than when considering salience injection. Whereas for salience injection interventions would need to crank up $c(f_0) - c(f^*)$ very high, here I want to consider the case where $c(f_0) - c(f^*)$ is relatively small. This is the merit of disinterest: No matter how small is $c(f_0) - c(f^*)$, if the probability that the decision will ever be implemented is sufficiently small (or far enough in the future), then an agent will decide what she thinks is right.[25] Thus, mere curiosity, or a little bit

---

[24] Of course, sometimes last-minute or high-pressure thinking and information gathering is hard; so we may want people to think about how they will deal with a contingency before that contingency arises.

[25] Besides the event being low probability, disinterestedness is helped when people don't know which side of a conflict they will be on. If you are asked to think how two people who stumble upon money should split it, but don't know whether you are likely to be the person who found it or merely her companion, your assessment will likely be more neutral. This is slightly different from the low-probability case, however, because you may more strongly prefer not to think about the issue now if you think it is likely, waiting until you find out where your interest lies. (Such a sophisticated and cynical strategy won't be likely if your moral preferences are time-inconsistent in the way mentioned

of intellectual integrity, can generate honest moral cognition when self interest is not salient.[26] It bears emphasizing that the model predicts that even very hands-off, "non-judgmental" moral education can improve moral conduct; merely encouraging people to think through moral issues for themselves can improve moral conduct, even if no guidance whatsoever is given in what the right answers are to any moral dilemmas. This says we can improve behavior of the next generation of citizens even if we do not claim to have better information about what is right and wrong; we can rely on their own insights into what is right and wrong, so long as we induce those insights in a disinterested setting.

When it comes to getting people to make moral judgments, our motto should be "Get 'em while they're disinterested." This suggests a comparison to Rawls's (1971) initial-position/veil-of-ignorance thought experiment. In the popularized version of Rawls's argument of which I am familiar, we can help determine ideals of distributive justice by hypothesizing what people would choose before they knew their lot in life. The Rawlsian thought experiment has people making the decision before they find out where their interests lie. The moral-priming hypothesis is a weaker (but less hypothetical) analog to such an argument: I am arguing that even if people don't make their decision until they find out where their self interest lies, we can reduce self-interested behavior by having them think before they find out. Of course, there are limits on how irreversible opinions are--the same self-serving biases that cause problems to begin with may overcome earlier moral education regarding what is right and wrong. But having decided some general principle, people may not always have the cleverness to talk themselves out of that principle.[27]

---

in the next footnote.)

[26] In the concluding section, I discuss the role of moral rules when a person has time-inconsistent preferences. In this context, moral priming may be especially useful, and may be self-imposed. My today self may want my tomorrow self to do the morally proper thing, and so today I may gather information in a non-self-serving way so as to mitigate self-serving biases tomorrow.

[27] Besides 'temporal disinterest', a more subtle form of moral priming seems frequently to be an element of moral debate. When we are argue with others about their moral conduct, we often try to get them to pursue some line of reasoning, catching them off guard about how their conclusions in this line of reasoning will reflect on their self interest.

While I have emphasized Sunday School and general moral socialization as examples of moral priming, the mechanism can be used in narrower economic contexts as well. A company that regularly observes some form of intra-organizational conflict, and suspects that self-serving biases will exacerbate such conflicts, will have an incentive to induce disinterested analysis. The moral-priming hypothesis says that the company might encourage new employees to decide what is right or wrong in that situation. This will backfire if new employees form overly strong judgments based on too little information; but it will improve things if it gets them to make a disinterested judgment that is sufficiently intelligent.

Moral priming is a mechanism to get individuals to behave morally in their own life choices. But society also wants its members to be disinterested in various contexts where they are called upon to determine the rightness or wrongness of *other people's* behavior. Mostly we care about direct self interest--we don't want to a judge or congressperson to make a decision that affects one of her investments, because she might sacrifice the social good for her own benefit. But we may also want such decision makers to be disinterested in a stronger sense--we want not only that the outcome she chooses has no self-interested implications, but also that the *reasoning* she uses in reaching a decision has no self-interested implications.

Suppose you own one of two small shops in town, and are asked to sit in judgment about the other shop keeper, who didn't pay her taxes to the King. You are asked to judge whether that is an act of treason, punishable by death. You yourself have not always paid your taxes, and have a taste for continuing to shirk on this front. What do you decide? Obviously there are many factors, but I wish to isolate two particular motives: Your direct self interest may lie in deeming her guilty of treason--you will have ridded yourself of a competitor. But the reasoning that lets you deem her guilty may unavoidably guarantee that you will hold yourself responsible for paying taxes. Operating under moral constraints, you may show moral solidarity with a person who is in a similar situation as you, even if your material self interest lies in condemning that person. The King's court may quite sensibly find that, as a general rule, people whose life choices are too similar to those of the accused will be too "biased" to be appropriate judges.

## 5. Moral Manipulating

The results in Sections 3 and 4 rely on the assumption that the agent is either (a) not aware of or (b) not bothered by her belief manipulation. Under interpretation (b), the model does not violate any principle of rationality or rational expectations. But I am uncomfortable with the assumption that people with full awareness manipulate their beliefs to reduce compliance with their own moral standards. We do not consciously and openly think such thoughts as "I merely need to turn over this paper to find out the social consequences of an action I plan, but I choose to remain ignorant so that I feel justified in taking the action." Rather, I believe (based on psychological evidence of the type discussed in the next section) the model is legitimate because agents manage not to be fully aware of how their information processing is biased.

But it bears emphasizing that this assumption of limited awareness is central to all the results of this paper; an agent who fully recognizes and disapproves of belief manipulation just as much as she disapproves of immoral actions will behave just like a P agent. To model this most simply, I model it tautologically: I assume the agent abides by a moral rule that requires her to gather as much information as she would if she had fbbe moral preferences.

Definition 5:

An agent is following the *Golden Moral Rule* (she is a *G* agent) if her utility function is as follows:

$$U_G(c,v) = \begin{cases} v - g(q,f) - c(f) & \text{if } x = 1 \\ - c(f) & \text{if } x = 0, \end{cases}$$

where $g(q,f) = 1$ if $q > v$ or $f \notin \text{argmax}_{f' \in \mathcal{F}} U_P(f')$, and
$g(q,f) = 0$ else.

A G agent gets severe disutility from engaging in an activity when *either* she believes it is likely to cause social harm *or* she is aware that she did not try sufficiently hard to gather relevant information about its social harm.[28] This formalization is a brute force way of capturing the idea that an

---

[28] This definition assumes that the agent derives no disutility from belief manipulation if she does not engage in the activity. But the behavioral

agent feels a moral obligation to pursue the social ideal both in choosing her activities *and* in gathering information and thinking through the issues involved. As such, we know that a G agent will gather the same information and choose the same actions as the P agent.[29] Let $x_G(c,v)$ be the probability that a G agent engages in the activity, and $z_G(c,v)$ be the probability that she does harm. Then:

Lemma 2:

For all $c$ and $v$, $x_G(c,v) = x_P(c,v)$ and $z_G(c,v) = z_P(c,v)$.

Proof:

Let $f_P$ be the signal the P agent would choose. Suppose that $f_P = f_0$. (Recall that $f_0$ is the "signal" that involves no updating from $q_0$.) Then the G agent will clearly choose $f_0$; her utility will be negative otherwise. Suppose that $f_P \neq f_0$. Then, since the P gets positive utility from getting this signal, so will the G agent (her utility is the same in this case as the P's agent). Since she gets non-positive utility from choosing any other signal, she will choose the signal. Since the G agent is fbbe to the P agent, she will choose to engage in the activity given the exact same set of realizations as does the P agent. Q.E.D.

An immediate corollary to Lemma 2 is, of course, that a G agent will obtain full information whenever it is readily available. Indeed, returning to issues of Section 5, we can use this model to consider a form of social

---

implications would be identical if the agent experienced disutility from suboptimal information-gathering whether or not she engaged in the activity.

[29] In a previous draft of the paper, I developed a more complicated information structure than in Section 3 that allows us to consider assumptions in between a R agent and G agent, and which assumes a kind of 'bounded awareness' of, or bounded compunction regarding, belief manipulation that seems consonant with psychological evidence. Imagine, for instance, that an agent feels an obligation to gather manifestly available information about whether the activity causes harm, but that she feels no moral obligation to gather information about whether such information is available, nor even to find out how to gather the information when she knows that it is probably available. That is, an agent holds herself "first-order" morally responsible, but not "second-order" responsible for finding out the consequences of her action.

influence that would clearly be useful: We should encourage people to recognize and disapprove of their self-serving biases, and to develop skills at unbiasing their thought process. Under one interpretation, getting people to disparage of belief manipulation might be considered an attempt to change their preferences. But it can also be interpreted as an informational intervention, where society wants to encourage an awareness that we tend to perceive things self-servingly. Consider the Messick and Sentis quote in the introduction, for instance. What would happen if we pointed out to agents that they were going through a list of different norms of fairness, and selecting *only* the ones that best match their self-interest? This may reduce the bias.

## 6. Related Psychological Research

I have developed a stylized, formal model of moral cognition, not based tightly on any one line of psychological research, but rather attempting to capture a general pattern of behavior observed in various strands of research. I now briefly review some of this literature.

Perhaps closest to the literal model of belief manipulation of Section 3 is the selective-exposure literature. Largely inspired by cognitive dissonance theory, this literature has looked at ways that people tend to avoid information that they find unappealing.[30] The totality of this literature well demonstrates that people sometimes prefer less information to more information, generally because they prefer to have more comfortable beliefs in some dimension. This is in contrast with the standard economic presumption that people care about information only for its instrumental value, and thus always prefer more of it to less. But cognitive dissonance, as both a buzz word and a unifying framework, has fallen out of favor among psychologists in recent years, and many strands of research with similar themes are no longer presented under the rubric of cognitive dissonance. Moreover, the emphasis in much of the cognitive dissonance literature is somewhat different from my concerns here. For instance, while the theory is broad enough (and vague enough) to accommodate belief manipulation at any stage, the emphasis in the literature has been how people alter their beliefs after making decisions,

---

[30] See Cotton (1985) for a review.

rather than before decisions. (Many of the economic applications, however, posit the ex ante belief manipulation.)[31]

Research on selective exposure, and more generally on mechanisms of self deception, eventually gave rise to the question of whether people are aware of their own information-processing strategies and biases. Such a question is particularly pertinent to the issues raised in Section 5 about the extent of people's moral scrutiny regarding their belief manipulation.[32] Experimental evidence supports the hypothesis that we can be very unaware of how we process information. Gur and Sackeim (1979, pp. 148-149) discuss this issue as follows:

> In the psychological literature, a ... paradox was raised by critics of research on subliminal perception and perceptual defense. The point was made that in order for a perceiver to avoid perceiving a stimulus, the stimulus must first be perceived. When it is assumed that perception implies awareness of the percept, notions like subliminal perception and perceptual defense are paradoxical. Proponents of the existence of these phenomena have argued that it is erroneous to assume that cognition must be subject to awareness. In support of this position, Nisbett and Wilson (1977) have summarized a sizable body of literature that indicates that

---

[31] Also, although presentations of cognitive dissonance theory often invoke moral issues (see, e.g., Aronson (1980)), little of the experimental evidence directly tests "moral dissonance." One obvious reason that moral issues cannot as readily be tested in the laboratory is that it is harder to interpret self-reported beliefs that are morally relevant. Those beliefs may be "impression management"--attempts by subjects not to appear to experimenters to be creeps. (See, e.g., Tetlock and Manstead (1985) for a discussion of this problem.) Some of the evidence on selective exposure outside the moral realm is less susceptible to being confounded by impression management.

[32] Though Gur and Sackiem (1979) and others insightfully address the issue, I have found much of the selective-exposure literature to be remarkably unencumbered by the law of iterated expectations. Bayesian information processing demands that if an agent sees enough of a signal to form expectations about the implication of the full signal, she must update her beliefs in expected terms even if she does not choose to get the signal. Much research explores how people avoid signals they don't like (e.g., that people avoid reading a favorable article about a politician they don't like) without discussing how this issues jibes with the law of iterated expectations. My model shows that selective exposure can matter even if it *doesn't* violate the law of iterated expectations. But it seemingly *does*. In addition to the type of automatic screening out explored by Gur and Sackiem (1979), one way by which it does is probably a sort of salience or memory-based mechanism--we know that if we read a positive magazine article about somebody we don't like, the positive images will stick with us; the momentary decision to not read an article will not stick with us.

> people may lack awareness for both the contents and processes involved in cognition. [Several citations have been eliminated from this quote.]

Gur and Sackeim (1979) develop an experiment that very cleverly establishes that people self-deceive themselves in a strong sense. Using physiological measurements, they establish that subjects who have received an unpleasant signal will, when motivated to do so, screen this signal out from their consciousness. Closer to the context of moral constraints, and in support of "Section 3 model" over the "Section 5 model", Ainslie (1992, pp. 175-176) discusses attempts to escape an internal rule--proposing something in between the "don't care" and "unaware" interpretations of belief manipulation presented in Section 5:

> To some extent, attention control can conceal behavior that violates a rule, or conceal the applicability of the rule to that behavior. That is, a person may be aware that certain information is apt to make a rule relevant to his current choice before he has actually collected that information. Although such a course evades his rule, it does not violate it; it is as legitimate as the congressional practice of 'stopping' the official clock when the members do not want to observe a previously imposed time limit on their deliberations. It permits the behavior to be performed without regard to the rule, rather than in violation of it. A person skilled in using attention control may examine an impulsive behavior fairly thoroughly, but still not ask the questions that would lead him to declare it either 'deliberate' or a lapse. The easiest targets will be behaviors that can be completed in short periods of time and those that are already ambiguous as to whether or not they follow the rule.

Separately from the research lines discussed above, there is overwhelming psychological evidence that people have self-serving biases in a range of matters. We tend to believe what is comfortable for us to believe. We are over-optimistic regarding our health and other aspects of our life (see Weinstein (1980, 1984) and Kunda (1987)), and think that we are superior to others in all sorts of ways (see Klein and Kunda (1993)). This includes believing we are fairer than other people. Moreover, in case-by-case applications of general norms of fairness, we tend to see as fair that which serves our own self-interest (see Kelley and Thibaut (1978) and Messick and Sentis (1983)).

Is this evidence in favor of the type of model I adopt in this paper? In fact, the source of self-serving biases has been the subject of much debate among social psychologists. With the cognitive revolution in social

psychology, "motivational" interpretetations of self-serving biases have been persistently challenged. Biases are motivated if we have them because our thinking is affected by what we want to be true; biases are *unmotivated* if they are solely artifacts of heuristic errors, unaffected by our motivations. Self-serving biases may be unmotivated. We may all believe we are better-than-average drivers not because we find this an attractive belief, but perhaps because the precautions we take are more salient to us than are the precautions other people take.[33] The case against motivated cognition is captured by the following quote from a prominent book on social cognition (Nisbett and Ross (1980, p. 12)):

> With so many errors on the cognitive side, it is often redundant and unparsimonious to look also for motivational errors. We argue that many phenomena generally regarded as motivational (for example, self-serving perceptions and attributions, ethnocentric beliefs, and many types of human conflict), can be understood better as products of relatively passionless information-processing errors than of deep-seated motivational forces.

The model of this paper clearly adopts the perspective that self-serving biases are motivated. But while psychologists may debate which type of explanation is "better", I am here more concerned merely with whether motivated cognition is a substantial part of the explanation. The skepticism expressed by critics of glib motivational interpretations of biases has induced efforts to find stronger evidence for the motivational basis of self-serving biases. And I'd like to think that recently the tide has turned among psychologists, and there now exists more careful evidence that self-serving biases are often motivated. As an example of such research, Kunda (1987, pp. 636-637) develops several related experiments that convincingly

---

[33] For a good example of this perspective on self-serving biases, see Miller and Ross (1975), and especially Nisbett and Ross (1980). Occasionally, samples of subjects and issues have been found with a bias that can be explained on unmotivated grounds but not motivational grounds. For instance, the often-cited paper of Ross and Sicoly (1979) found that a sample of subjects tended to believe that they started more fights than did their spouses; this belief is a bias, but goes against the motivation to believe that we are better than others are (and the bias was not due to worshipful attitudes towards spouses). But nobody on either side of the debate argues that roughly half the time our bias is self-serving and half the time is "self-damaging". Mostly our biases are self-serving. The debate is over why this pattern holds.

suggest motivational biases:

> The present approach takes into account both motivational and inferential factors. In this view, the cognitive apparatus is harnessed in the service of motivational ends. People use cognitive inferential mechanisms and processes to arrive at their desired conclusions, but motivational forces determine which processes will be used in a given instance and which evidence will be considered. The conclusions therefore appear to be rationally supported by evidence, but in fact the evidence itself is tainted by motivation: its production and evaluation are guided by motivation.

Ditto and Lopez (1992, pp. 568-569) nicely summarize the history of this research question:

> The intuition that hopes, wishes, apprehensions, and fears affect judgments is compelling and persistent. Turning this intuition into a viable empirical and theoretical fact, however, has proved to be one of the most recalcitrant problems in the history of of experimental psychology. Yet, after the reaching its nadir with the publication of Tetlock and Levi's (1982) essay on the intractability of distinguishing "cognitive" and motivational explanations for self-serving attributional biases, progress in conceptualizing the role of motivational factors in judgment processes has resurged. ... Within the motivated-judgment literature, the empirical finding that has received the greatest amount of attention and controversy in recent years is the robust tendency of individuals to perceive information that is consistent with a preferred judgment conclusion ... as more valid than information that is inconsistent with that conclusion. [Several citations have been eliminated from this quote.]

I believe the hypothesis that our information processing is often motivated is well established. But experimental studies where the motivated beliefs are specifically about moral issues are harder to find. There has, however, been some research focused more specifically on moral issues. The research on fairness biases, discussed by Messick and Sentis (1983), is probably the most economically relevant example. Research by Bandura (1990) on moral disengagement has a very similar flavor to the issues I've discussed, but does not carefully calibrate the type of belief manipulation central to my model. Another relevant area of research is the related literatures on the just-world hypothesis, blaming the victim, and victim derogation. We tend to believe that the world is a more just place than it really is, and that people tend to get what they deserve. This tendency is at its nastiest when it leads us to blame and derogate victims that might more reasonably be considered undeserving of blame. One hypothesis for why we blame victims is that it

allows us to alleviate our guilt for either having caused a person's problems, or to let ourselves off the hook in helping them. Such an explanation fits well with the model I have presented.[34]

Finally, I turn to an example of experimental evidence on one of the informational interventions that I discussed in Section 4. (Footnotes in Section 4 already discussed Milgram's (1974) arguments on the role of "salience".) In their research on self-serving biases in the legal context, Loewenstein *et al* (1993) and Babcock, Loewenstein *et al* (1995) provided evidence on the moral-priming hypothesis. In a general experiment on the role of self-serving biases in legal contexts, they demonstrated that subjects given roles (and real stakes) in a hypothetical legal dispute were apt to interpret evidence self-servingly. Subjects manifestly cared about the fairness of the outcome, but, when reading a description of the case, their assessments of fairness depended on which role (plaintiff or defendant) they knew they were going to play in the ensuing negotiations. The researchers show that the costliness of disputes increased due to the biased reading.

But the researchers conducted a variant of the experiment that strongly supports the moral-priming hypothesis. They informed some subjects before reading the case whether they were defendants or plaintiffs. But other subjects were told of their role *after* they read the case, but *before* they began negotiating. In this latter group, dispute costs were far lower. Without knowing where their self-interest lied, subjects were reading the evidence disinterestedly, and could not quickly re-bias themselves.

---

[34] My sense of the research, however, is that demonstrably other factors play a role, and probably a more important role. For instance, we may prefer to believe in a just world because it is scarier for us to believe innocent people are victims--we might be next. This is certainly a motivated bias, but not the type of alleviation of moral responsibility explored in my model. See Lerner and Miller (1978) for an excellent discussion of research on the motivations behind the just-world hypothesis, and for arguments against interpreting the phenomenon in terms of us relaxing our own moral obligations.

## 7. Discussion and Conclusion

I have said nothing about where moral constraints come from. Perhaps the moral rules we live by are internalized remnants of the external rules our parents and society taught us as we grew up. Such socialization is much studied by developmental psychologists, sociologists, and others, and these lines of research will likely yield a better understanding of the type of moral constraints explored in this paper.

Hypothesizing that moral constraints come from socialization suggests that *innate* other-regarding behavior may operate as true moral *preferences*. As has been emphasized recently by evolutionary scientists, kin altruism may be as innate a human motivation as self interest, because genes shared among kin are more likely to survive if they induce the kin to help each other. If such innate predispositions enter as preferences, then they may not be well-captured by a model of morality as internal constraints. (I do not know if there is research on the intensity of self-serving biases among kin.)

Just as "true preferences" don't stop as we leave a person's bodily self, moral constraints don't start only upon departure. Some moral constraints that operate fully within the sphere of self interest: We are often in the position of thinking that we "should" do something that is good for ourselves, though we "want" to do something else. You want cheesecake for desert, but feel you should have an apple. Many such instances of internal tension appear to be usefully conceptualized as a self-control problem arising from time-inconsistent preferences; we prefer today that we not over-indulge tomorrow, but tomorrow prefer to over-indulge. A model of this phenomenon has been developed brilliantly by Ainslie (1992).[35]

Based on clinical experience, Ainslie emphasizes the strong role that internal moral rules play in overcoming self-control problems. Just as society imposes upon us moral rules to overcome the mistreatment of other people, so too we impose upon ourselves moral rules to overcome the mistreatment of our future selves. Our "true preference" is for immediate gratification, but we

---

[35] Formally, Ainslie models preferences as involving hyperbolic time-discounting rather than exponential. See also Ainslie (1991) and Laibson (1994) for further developing this model, and Schelling (1978) and Thaler and Sheffrin (1981) for earlier models of self-control issues.

feel (morally) obliged to take care of our future self.[36] Such a conceptualization of intertemporal choice invites the type of analysis I have developed in this paper, and the results of Sections 3, 4, and 5 have analogs in this realm.

At the end of Section 3, for instance, I discussed a way that economists might, by "misidentifying" moral behavior as coming from true moral preferences, exaggerate the degree to which we actually take into consideration others' well-being. Precisely the same logic indicates that economists are exaggerating the degree to which people have time-consistent "exponential" preferences. Even if belief-contingent behavior suggests (when interpreted in the standard framework) that people exponentially discount future payoffs, because of belief manipulation people may *de facto* be behaving more like the Ainsliean, hyperbolic discounters that they truly are. That is, we may over-eat not because we consciously sanction over-weighting current enjoyment well-being over future well-being, but because we systematically deceive ourselves in ways that support immediate gratification.

The types of informational interventions discussed in Section 4 are probably quite important in the intrapersonal realm. Arguably, efforts such as providing prominent warnings on cigarette cartons are an example of salience injection by outside parties. While this and other health-related warnings could simply be (good, old-fashioned) provision of information, it seems plausible that the effort to incessantly remind people of well-known facts may be an effort to aid people in their constant struggle for long-run well-being over immediate gratification.[37] Moreover, while government authorities may

---

[36] Ainslie (1992, p. 57) puts it, less extremely than have I, as follows:

> We are used to thinking of ourselves as consistent, and often we are right. People do not radically devalue the future for most purposes, but guard their interest over the years and save their money. ... It is just as supportable, however, to say that living mostly for the present moment is our natural mode of functioning, and that consistent behavior is sometimes acquired, to a greater or lesser extent, as a skill.

See also Prelec (1991) for an insightful discussion of the moral realm in intrapersonal choice, as well as the analogies between such issues and the interpersonal realm.

[37] One thought experiment is to ask whether we work harder to make salient those health risks that are not immediate than those that are immediate. For instance, it is my perception that diabetics do not need loud in-your-face

label packages for you, in the realm of intrapersonal morality the "outside party" may be *You*. Most of us who have self-control problems are aware of them, and attempt to intervene to overcome them.[38] In a non-existent companion paper, I consider all these issues further.

I have argued in this paper that describing moral dispositions straightforwardly as preferences may not always be adequate. But my model develops only one specific alternative, and it in turn is inadequate for capturing certain issues. One is the role that post-decision feelings, such as self-image and guilt, play in our moral conduct. Our motivation in behaving morally is often to achieve a positive self-image--we know we'll feel good about ourselves if we do the right thing, and feel guilty if we knowingly hurt people. If people had perfect and unbiased memories, these emotions would probably be adequately captured by my model. But since we can forget things, new issues arise that are likely to undermine some of the implications of the model. Of particular interest, limited memory raises self-inference and self-signaling issues. In determining whether to feel guilt for past actions, we must first recall the relevant decisions and reconstruct the information we had available at the time of decision. All sorts of issues arise; if we know we will forget having had to make a choice (e.g., if we pass a homeless person

---

labeling of dangerous goods.

[38] Again, Ainslie (1992, p. 199) captures all these issues well:

> [A dieter] must evaluate additional options, such as finding loopholes to permit consumption in the case at hand, or avoiding the perception of a lapse by not gathering or processing the necessary information--the equivalents, in intrapersonal negotiation, of a businessman hiring lawyers or keeping a false set of books, respectively. By such means the dieter may justify a contemplated excess, or direct his notice away from the properties of his food that are relevant to his diet.

But these very manipulations lead the same person, in her incarnation as long-term welfare maximizer, to "informationally intervene":

> Insofar as these evasions have worked, it will be in his long-range interest to study the need for more rules to protect the original rules from just such evasion, redefining them to close loopholes or requiring more systematic testing of whether or not they are being followed.

For related arguments, perceiving the long-run self as a "principal" who is trying to overcome the manipulations of a series of short-run "agents", see Laibson (1994).

with full knowledge that we won't remember this instance of having done so), we may successfully avoid moral obligations. On the other hand, especially if we observe the negative consequences of not having helped somebody, we may sometimes engage in a seemingly moral act that is not merited by our current information, because we are worried about *proving* to our future selves that we weren't being selfish. Guilt may deter not only impropriety, but the appearance (to ourselves) of impropriety.[39]

This paper has also concentrated solely on pleasant social dispositions. But many darker social motivations seem to be important, and economists have begun to investigate the implications of these emotions. Frank (1985, 1988), Thaler (1988), Rabin (1993), Mui (1995), Camerer and Thaler (1995), and Babcock *et al* (1995), for instance, explore such morally dubious preferences as envy, status-seeking, and revenge. I am unfamiliar with any research investigating the nature of biases in the realm of these darker motivations.[40]

Finally, I return to the choice of modeling belief manipulation as Bayesian. Besides matching the facts better, a non-Bayesian model may also be central for drawing out certain important implications of self-serving biases. In attempting to incorporate "moral dissonance" into formal models of social decision making, Rabin (1994) models agents as exhibiting cognitive dissonance that biases beliefs towards self interest. The crux of that model was that there is social contagion--if a person convinces herself that it is okay to wear fur, that makes it easier for her neighbors to convince themselves. Such "moral herding" can lead to strong and perverse effects when combined with cognitive dissonance. For instance, indoctrinating people more strongly that they should give to those who are needy may lead to less giving to the

---

[39] Because of such issues, I believe that my model is probably most applicable in those common situations where we never learn the consequences of our decisions. This may reduce the fear of guilt (as suggested by Milgram's arguments about the uses of bureaucracy in Section 4). For psychological research relating to self-perception, guilt, and self-signaling, see for instance Batson (1982) and Bem (1965, 1972).

[40] Insofar as one motivation for retaliation against unfair behavior may be to enforce social justice (an abusive person should be punished, lest she abuse the next person she interacts with), "revenge" is an altruistic act for the social good. Perhaps people will be be biased towards beliefs that allow them not to sacrifice their material well-being for the sake of socially-useful retaliation. (Indeed, this could be one interpretation for some of the findings of Croson (1993).)

homeless, because it will cause people to convince themselves, and each other, that homeless people aren't truly needy.

But the logic of such moral herding very much depends on a statistical bias *not* assumed in this paper; it is probably only reasonable to make a general conjecture about the social effects of self-serving biases if we know that *expected* beliefs are biased by self interest. As implied by Proposition 1, a Bayesian who self-servingly aims to figure out reasons not to give to a homeless person is *more* likely to develop strong views that a homeless person needs her help than to develop strong views that he doesn't; this is because once the Bayesian satisfies herself that giving *probably* isn't morally necessary, she is not going to try to think harder about it to decide *for sure* it isn't. Depending on precise parameters of the environment, the implications of self-justification among a society of Bayesians might go either way. Such indeterminacy may be correct in some ways; but more often than not we should suspect that society's beliefs will be biased towards the average self interest of its members. To capture that hypothesis, we need to model self-serving biases as non-Bayesian.[41]

---

[41] Another issue will arise if trying to incorporate self-serving biases into models of moral herding: We must make assumptions about whether people recognize self-serving biases in others. If we know that others' beliefs are often self-serving, we may not weight the beliefs of others as heavily when those beliefs are too consonant with the others' self interest. The degree of such skepticism is likely to matter importantly in determining the social contagion of self-serving beliefs.

## References


Ainslie, G. (1991), "Derivation of 'Rational' Economic Behavior from Hyperbolic Discount Curves," American Economic Review 81(2), May, 334-340.

Ainslie, G. (1992), Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person, Cambridge University Press.

Akerlof, G. (1989), "The Economics of Illusion," Economics and Politics 1, Spring, 1-15.

Akerlof, G. and Dickens, W. (1982), "The Economic Consequences of Cognitive Dissonance," American Economic Review 72, 307-319.

Aronson, Elliot (1980), The Social Animal, 3rd edition (W.H.Freeman, San Fransisco).

Babcock, Linda, Loewenstein, George, Issacharoff, Samuel, and Camerer, Colin (1995), "Do Bargainers Form Biased Expectations of Fairness in Bargaining?", American Economic Review, forthcoming.

Babcock, Linda and Olson, Craig (1992), "The Causes of Impasses in Bargaining," Industrial Relations 31, 348-360.

Babcock, Linda, Wang, Xianghong, and Loewenstein, George (1995), "Choosing the Wrong Pond: Social Comparisons in Negotiations that Reflect a Self-Serving Bias," manuscript, Carnegie Mellon University, June.

Bandura, Albert (1990), "Selective Activation and Disengagement of Moral Control," Journal of Social Issues 46(1), Spring, 27-46.

Batson, Daniel C. (1982), "Prosocial Motivation: Is it Ever Truly Altruistic?" Advances in Experimental Psychology 75, 73-98.

Becker, Gary S. (1981), A Treatise on the Family, Harvard University Press, Cambridge, MA.

Bem, D.J. (1965), "An Experimental Analysis of Self-Persuasion," Journal of Experimental Social Psychology 1, 199-218.

Bem, D.J. (1972), "Self-Perception Theory," in L. Berkowitz (ed.), Advances in Experiemental Social Psychology 6, New York: Academic Press.

Camerer, Colin and Thaler, Richard H. (1995), "Ultimatums, Dictators, and Manners," Journal of Economic Perspectives 9(2), Spring, 209-219.

Cotton, John L. (1985), "Cognitive Dissonance in Selective Exposure," in Dolf Zillman and Jennings Bryant (eds.), Selective Exposure in Communication, Lawrence Erlbaum Associates, 11-33.

Croson, Rachel T. (1993), "Information in Ultimatum Games: AN Experimental Study," manuscript, Harvard University, October.

Dickens, William T. (1986), "Crime and Punishment Again: The Economic Approach with a Psychological Twist," Journal of Public Economics 30, 97-107.

Ditto, Peter H. and Lopez, David F. (1992), "Motivated Skepticism: Use of Differential Decision Criteria for Preferred and Nonpreferred Conclusions," Journal of Personality and Social Psychology 63(4), pp. 568-584.

Elster, Jon (ed.) (1985), The Multiple Self, New York: Cambridge University Press.

Frank, Robert (1985), Choosing the Right Pond: Human Behavior and the Quest for Status, New York: Oxford University Press.

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

Freeman, Richard B. (1993), "Give To Charity -- Well, Since You Asked," discussion draft, LSE Conference on the Economics and PSychology of Happiness and Fairness, London, November 4-5.

Gur, Ruben C. and Sackeim, Harold A. (1979), "Self-Deception: A Concept in Search of a Phenomenon," Journal of Personality and Social Psychology 37(2), February, pp. 147-169.

Kelley, H.H. and Thibaut, J. (1978), Interpersonal Relations: A Theory of Interdependence, New York: Wiley.

Klein, William and Kunda, Ziva (1993), "Maintaining Self-Serving Social Comparisons: Biased Reconstruction of One's Past Behaviors," Personality and Social Psychology Bulletin 19(6), December, 732-739.

Kremer, Michael (1995), "Integrating Behavioral Choice into Epidemiological Models of AIDS," Preliminary Manuscript, Hoover Institution, April.

Kreps, David and Porteus, Evan (1979), "Temporal von Neumann-Morgenstern and Induced Preferences," Journal of Economic Theory 20, 81-109.

Kunda, Ziva (1987), "Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories," Journal of Personality and Social Psychology 53(4), 636-647.

Laibson, David (1994), Hyperbolic Discounting and Consumption, MIT Department of Economics Dissertation, May.

Lerner, Melvin J. and Miller, Dale (1978), "Just World Research and the Attribution Process: Looking Back and Ahead," Psychological Bulletin 85(5), 1030-1051.

Loewenstein, George, Issacharoff, Samuel, Camerer, Colin, and Babcock, Linda (1993), Journal of Legal Studies 22, January, pp. 135-159.

Margolis, Howard (1982), Selfishness, Altruism, and Rationality: A Theory of Social Choice, Chicago: University of Chicago Press.

Messick, David and Sentis, Keith (1983), "Fairness, Preferences, and Fairness Biases," in David M. Messick and Karen S. Cook (eds.), Equity Theory: Psychological and Sociological Perspectives, New York: Praeger Publishers, pp. 61-94.

Milgram, Stanley (1974), Obedience to Authority, New York: Harper and Row.

Miller, Dale T. and Ross, Michael (1975), "Self-Serving Biases in the Attribution of Causality: Fact or Fiction?" Psychological Bulletin 82(2), 213-225.

Montgomery, James (1993), "Revisiting Talley's Corner: Mainstream Norms, Cognitive Dissonance, and Underclass Behavior," Rationality and Society 6(4), October, 462-468.

Mousavisadeh, N. (1995), "States of Denial," The New Republic, June 19, pp. 40-43.

Mui, Vai-Lam (1995), "The Economics of Envy," Journal of Economic Behavior and Organization 26(3), May, pp 311-336.

Mukerjee, Madhusree (1995), "Toxins Abounding," Scientific American 273(1), July, pp. 22-23.

Nisbett, Richard and Ross, Lee (1980), Human Inference: Strategies and Shortcomings of Social Judgment, Englewood Cliffs, NJ: Prentice-Hall Inc.

Nisbett, R.E and Wilson, T.C. (1977), "Telling more than we can know: Verbal reports on mental processes," Psychological Review 84, 231-259.

Prelec, Drazen (1991), "Values and Principles: Some Limitations on Traditional Economic Analysis," in A. Etzioni and P. Lawrence (eds.), Socioeconomics: Toward a New Synthesis, New York: M. E. Sharpe, pp 131-145.

Rabin, M. (1993), "Incorporating Fairness into Game Theory and Economics," American Economic Review 83, 1281-1302, December 1993.

Rabin, M. (1994), "Cognitive Dissonance and Social Change," Journal of Economic Behavior and Organization 23, March, 177-194.

Rawls, John (1971), A Theory of Justice, Cambridge, MA: Harvard University Press.

Ross, M. and Sicoly, F. (1979), "Egocentric Biases in Availability and Attribution," Journal of Personality and Social Psychology 37, 322-336.

Schelling, Thomas C. (1978), "Egonomics, or the Art of Self-Management," American Economic Review 68(2), May, 290-294.

Tetlock, P.E. and Levi, A. (1982), "Attribution bias: On the inconclusiveness of the cognition-motivation debate," Journal of Experimental Social Psychology 18, 68-88.

Tetlock, P.E. and Manstead, A.S. (1985), "Impression Management versus Intrapsychic Explanations in Social Psychology: A Useful Dichotomy?" Psychological Review 92(1), 59-77.

Thaler, Richard H. and Sheffrin, Hersh (1981), "An Economic Theory of Self Control," Journal of Political Economy 89, April, 392-405.

Thompson, Leigh and Loewenstein, George (1992), "Egocentric Interpretations of Fairness and Interpersonal Conflict," Organizational Behavior and Human Decision Processes 51(2), March, 176-197.

Wakker, Peter (1988), "Nonexpected Utility as Aversion to Information," Journal of Behavioral Decision Making 1, 169-175.

Weinstein, N. D. (1980), "Unrealistic Optimism about Future Life Events," Journal of Personality and Social Psychology 39, 806-820.

Weinstein, N. D. (1984), "Why it Won't happen to me: Perceptions of Risk Factors and Susceptibility," Health Psychology 3, 431-457.