ELSEVIER

Journal of Economic Behavior and Organization
23 (1994) 177–194

JOURNAL OF
Economic Behavior
& Organization

Cognitive dissonance and social change

Matthew Rabin

Department of Economics, University of California at Berkeley, Berkeley, CA 94720, USA

Received October 1991, final version received October 1992

Abstract

When people behave immorally according to their own standards, they feel bad. Rational people may therefore engage in less of an immoral activity than would be in their material self-interest. Despite this fact, I show that increasing people's distaste for being immoral can *increase* the level of immoral activities. This can happen because of the psychological phenomenon of *cognitive dissonance*: people will feel pressure to convince themselves that immoral activities are in fact moral; if each person's beliefs affect the beliefs of others, then increasing the pain from being immoral may cause members of society to convince each other that immoral activities are morally okay, and society will engage in more of such activities.

Key words: Cognitive dissonance; Social issues

JEL classification: A12; A13

1. Introduction

All of us prefer to think of ourselves as moral people, and when we do something we know hurts others, we feel bad because such behavior conflicts with our view of ourselves as moral people. If we believe that wearing fur is cruel to animals, we will feel bad if we wear fur.

This paper examines the effects of such moral concerns when members of a society interactively decide how much of a morally dubious activity to engage in. I feel that the issues of this paper apply most closely to social issues, where members of society debate what constitutes proper individual behavior in a certain realm, with many people urging others to change their

*I thank Eddie Dekel-Tabak, Bill Dickens, Sean Ennis, Vai-Lam Mui, and two anonymous referees for helpful comments.

SSDI 0167-2681(94)E0045-2

behavior.[1] While clearly this process of shifts in beliefs and behavior is dynamic, for simplicity I present a static model, with behavior and attitudes shifting immediately in response to shifts in the relevant parameters.

My main conclusion is that increasing the propensity of people to feel bad when they engage in immoral activities might actually *increase* the level of these immoral activities. This perverse effect could *not* occur with an isolated individual, but rather occurs only when members of society learn about and care about each other's beliefs about morality.

I use the standard economic tool of a formal rational-choice model, but with a psychological twist: I incorporate *cognitive dissonance* into people's preferences. Cognitive dissonance occurs when a person does something that is inconsistent with his beliefs. This dissonance between behavior and beliefs is unpleasant to most people, and psychologists have long studied how cognitive dissonance affects people's feelings and behavior [see, for example, Aronson (1980)]. This paper focuses on 'moral dissonance': engaging in immoral activities conflicts with our notion of ourselves as moral people. This broad class of cognitive dissonance is a frequently invoked example by psychologists.

Because it is unpleasant, people prefer to reduce cognitive dissonance. There are two ways to do so. As economists generally assume, people can change their behavior. Or – much less familiar to an economist – people can change their beliefs. Of course, the ability to change our beliefs is limited; to some degree, our beliefs reflect our 'true', disinterested consciences. I model a person's difficulty of maintaining 'false' beliefs with a cost function such that a utility-maximizing person will trade off his preference for feeling good about himself with the cost of maintaining false beliefs. While modeling belief-formation in this way is somewhat contrived, it is behaviorally accurate to assume people tend to change their views more when the cognitive benefits of doing so are greater.[2]

A natural conjecture is that the greater the distaste for doing something immoral – the more our consciences tend to bother us – the less of an immoral activity we will engage in. In section 2, I formalize this conjecture, and show that it holds when our notion of 'greater' pain from cognitive dissonance is that of an increase in both absolute and marginal disutility of cognitive dissonance. I also show that a systematic result holds for beliefs as well: the greater the disutility from cognitive dissonance, the more people will

[1] Throughout the paper, I illustrate my arguments with animal-rights issues. While I do so because I feel humankind's current treatment of animals is an important moral issue, the points of the paper should be applicable and interpretable to those who do not share my views.

[2] Moreover, this modeling approach is standard in the mini-literature of rational-choice, cognitive-dissonance models. Akerlof and Dickens (1982) introduced the general issue of cognitive dissonance into formal rational-choice models, and modeled beliefs directly into the utility function as a choice variable.

change their beliefs to convince themselves that what they are doing is right. That is, a greater propensity for cognitive dissonance is likely to *both* make people behave more in line with their underlying moral beliefs, and to alter their eventual, conscious beliefs to be less in line with these underlying moral beliefs. The more unpleasant it is to wear fur if you believe it is immoral, the less fur you are likely to wear, and the more tempted you will be to convince yourself that fur does not cause animal suffering.

In section 3, I add an aspect of belief-formation that is especially important when considering social issues and social change. Namely, *changes in beliefs by some individuals towards the 'true' morality are likely to increase the cost to other people of having beliefs far away from the 'true' morality.* People find it harder to convince themselves that an activity is ethical if nobody else believes it is ethical. If everybody else decides that torturing animals for fur is wrong, then it becomes harder for an individual to convince himself that there is nothing wrong with wearing fur. Conversely, if everybody ignores the suffering caused to the animals, then it becomes easier for each person to ignore the true nature of his actions.

These social effects in belief-formation have an important implication: a greater distaste for cognitive dissonance may lead not to less of an immoral activity, but rather to more of it. While a greater distaste for cognitive dissonance has the direct effect of decreasing the level of immoral activities, an indirect effect works in the opposite direction. Stronger cognitive dissonance will cause each person to believe that such activities are more acceptable; this in turns leads others to believe that the activity is more acceptable. Cognitive dissonance can lead to a conspiracy of silence, with everybody convincing themselves and each other that the activity that they are engaging in is okay, and this leads to more of the activity.

The results of this paper can be instructive for how, collectively and individually, we can best promote desirable social change. I show that, under plausible conditions, two straightforward results hold: decreasing the direct, material utility of an immoral activity will always lead to less of an immoral activity, and increasing the difficulty of maintaining immoral beliefs will also lead to less of the activity. Thus, if we tax fur more heavily, people will buy less of it. If we better educate people about the cruelty of fur production, we will presumably make it more difficult for people to convince themselves that fur is moral, and people will buy less of it. But the 'perverse' result noted above shows that if we try harder to convince people that they should not be cruel, we might increase the amount of fur people buy. Thus, trying to increase the unpleasantness of cognitive dissonance through social pressures – through, for instance, the indoctrination of general religious principles – might backfire. To actively inculcate people with the principle that they must do nothing that they know hurts animals might lead people to hurt animals more than they otherwise would have.

To obtain all of these results, I assume that people's beliefs about the morality of an activity spread automatically. However, how much people convey their true beliefs is subject to strategic and psychological constraints. Spreading of beliefs is therefore not automatic, and complicated issues arise. In section 4, I discuss how extensions of the model in this paper might capture such complications, and conjecture how the results would be affected.

2. Cognitive dissonance, beliefs, and behavior

Suppose that there is some activity that a person would enjoy a great deal if all moral considerations were ignored. Let $X \in [0, \infty)$ be the amount of the activity the person engages in, and let the 'material utility' from the activity be $U(X)$, where $U'(X) > 0$ for all X. Thus, in the absence of moral considerations, the person's preferred level of the activity would be $X = \infty$.

There might, however, be a moral-dissonance cost to choosing a high X, because the person may realize that he is hurting others by engaging in this activity. I shall assume that the person believes that there is some morally legitimate level of the activity, Y, such that the person suffers from cognitive dissonance if he chooses level X greater than Y. I represent this cognitive dissonance function by $D(X-Y)$, where $D' > 0$ for all values of $X - Y > 0$. For simplicity, I assume that $D(X-Y) = 0$ if $X \leqq Y$. That is, the person experiences no cognitive dissonance if he does not consume above the morally legitimate level.[3]

I assume that people set not only X, but Y – people may alter their beliefs to feel better about engaging in questionable activities. If changing beliefs were costless, people would simply choose to believe that there is nothing wrong with an activity, so that without any qualms they could engage in as much of the activity as they wished. That is, they would choose $X = Y = \infty$. When considering whether to wear fur, people would simply convince themselves that wearing fur causes no suffering to animals.

But developing certain beliefs is likely to be costly. In general, there is likely to be a natural, intellectually honest set of beliefs about the morality of an activity. Developing beliefs that differ from this level is costly because it may intrinsically conflict with other parts of a person's belief system, and reintegrating it can involve laborious intellectual activity. It can also involve forgoing pleasant activities to avoid receiving new information (not going to a rock concert that promotes animal rights). To capture the difficulty of

[3] Under the specifications of the model, a person will never choose to engage in the activity to a less than morally legitimate degree. Yet, in principle, people may experience dissonance from consuming below the morally legitimate level, because this makes them feel like 'suckers'.

developing certain beliefs, I let the function $C(Y)$ represent the psychic cost of holding beliefs Y, where $C(0)=0$ and $C'(Y)>0$ for all Y. This function assumes that it is harder to believe that Y is high; this reflects my assumption that if a person were honest with himself, and not trying to relieve his conscience, he would set $Y=0$. Thus, Y represents the distance between a person's beliefs and the true moral level of activity. It is hard to convince yourself that hurting animals is okay, and the more you hurt them, the harder it is to defend your actions.[4] I shall furthermore assume throughout the paper that the 'honest' beliefs corresponds to the 'true' morality. That is, I assume throughout the paper that the socially optimal level of the activity is $X=0$.

To summarize, the two variables that a person chooses are X, the level of an activity, and Y, the beliefs about the morally legitimate level of that activity. He maximizes his utility by balancing his material utility with both the disutility from cognitive dissonance and the disutility from maintaining 'false' beliefs. To represent formally this objective, I will assume that the person's overall utility is simply his material utility minus each of the two types of disutility. That is, in choosing his behavior and beliefs, each person implicitly solves the following:

$$\underset{X,Y}{\text{Max}}\, L(X,Y) \equiv U(X) - D(X-Y) - C(Y);$$

where $X, Y \geqq 0; U'(X), D'(X-Y), C'(Y) > 0;$

and $U''(X) < 0, D''(X-Y), C''(Y) > 0.$

The second-order conditions on these functions are standard means of guaranteeing that the first-order conditions are sufficient for obtaining maxima. I have neither proven that the results of this paper are true without these convexity assumptions, nor have I found any counter-examples.

There are three natural conjectures about how changing the three components of a person's utility function will affect his behavior and beliefs:

Conjecture A

If a person receives less material utility from engaging in an activity, he is likely to engage in less of that activity. And because this means he is under

[4] In the formal model, I discuss X as if it represents the level of some homogenous activity. But as the animal-rights example makes clear, it is often better interpreted as a reduced form of some set of activities of a similar nature, but of varying severity. If X is a composite activity of using animal products, eating veal or wearing fur might increase X by a lot, whereas eating fish or conducting animal experiments may increase it by a little. Also note that assuming $X=0$ as the socially optimal level is merely for simplicity; the thrust of this paper would hold even if we believed that, say, torturing animals for medical research were often justified, so that some $X>0$ would be the socially optimal level.

less pressure to convince himself that the activity is morally defensible, he will change his beliefs towards thinking it is less moral.

For instance, the less a person likes the taste of veal, or the more expensive it is, the less likely he is to eat veal, and the more repugnant he will find eating veal. People who detest the taste of veal would let their consciences act freely, and view veal production as completely immoral.

Conjecture B

The more costly it is to maintain dishonest beliefs, the less a person will convince himself that an immoral activity is okay, and in turn, the less of the activity he is likely to engage in.

If an animal-rights concert features good rather than mediocre bands, it is more costly to avoid the concert, and thus more costly to avoid thinking about animal rights. This leads a person to have more qualms about wearing fur, and to wearing less of it.

Conjecture C

The greater the distaste for cognitive dissonance, the less of an immoral activity a person will engage in, and the more moral will he think the activity.

That is, the more unpleasant a person finds behaving in a way that violates his moral principles, the more likely he is to engage in less of that activity. Yet greater discomfort from cognitive dissonance also creates a greater incentive for him to convince himself that the activity is okay. Taken together, this means that the greater the discomfort from cognitive dissonance, the closer a person will be to behaving in accordance to his morals.

Note how Conjectures B and C differ, relating respectively to the functions $C(\cdot)$ and $D(\cdot)$. If we convince people not to hurt others (thus increasing $D(\cdot)$) we still leave them to decide which activities do in fact hurt others. Convincing a person that he must not cause animals to suffer will stop him from wearing fur only if he believes that fur production causes suffering. This distinction will be important for the results of the next section.

One formal interpretation of these conjectures is that they refer to comparisons in behavior of two different functions. For instance, we might say that $\hat{U}(\cdot)$ represents lower utility than $U(\cdot)$ if for all X, $U(X) > \hat{U}(X)$. However, as is frequently the case in economics, it turns out that the marginal values of each of the functions are important for determining the behavior and beliefs of different people. For instance, the effects of increasing the material utility of engaging in an activity is ambiguous; the implications can depend on whether the marginal utility of the activity is also increased.

Thus, we need a definition of when one utility function is greater than another that compares both absolute and marginal levels. The following definitions suffice: $U(\cdot)$ represents greater utility than $\hat{U}(\cdot)$ if, for all X, $U(X) > \hat{U}(X)$ and $U'(X) > \hat{U}'(\cdot)$; $\hat{D}(\cdot)$ represents a greater propensity for cognitive dissonance than $D(\cdot)$ if, for all $X > Y > 0$, $\hat{D}(X-Y) > D(X-Y)$ and $\hat{D}'(X-Y) > D'(X-Y)$; and $\hat{C}(Y)$ represents a greater cost of changing beliefs than $C(Y)$ if, for all Y, $\hat{C}(Y) > C(Y)$ and $\hat{C}'(Y) > C'(Y)$.

With the above definitions, we can consider formally whether each of the three conjectures are true by comparative statics exercises that separately examine changes in each of the three component functions. Thus formalized, Propositions 1A–C demonstrate that the three conjectures are true:[5]

Proposition 1. Consider the functions $U(\cdot)$, $D(\cdot)$, $C(\cdot)$, $\hat{U}(\cdot)$, $\hat{D}(\cdot)$, *and* $\hat{C}(\cdot)$, *where* $U(\cdot)$ *is greater than* $\hat{U}(\cdot)$, $\hat{D}(\cdot)$ *is greater than* $D(\cdot)$, *and* $\hat{C}(\cdot)$ *is greater than* $C(\cdot)$, *and where* $U(\cdot)$ *and* $\hat{U}(\cdot)$ *are concave, and* $D(\cdot)$, $C(\cdot)$, $\hat{D}(\cdot)$, *and* $\hat{C}(\cdot)$ *are convex. Then the following results hold:*

(A) If (X^*, Y^*) *solves* $\text{Max}_{X,Y} U(X) - D(X-Y) - C(Y)$ *and* $(\hat{X}, \hat{Y})$ *solves* $\text{Max}_{X,Y} \hat{U}(\cdot) - D(X-Y) - C(Y)$, *then* $\hat{X} < X^*$ *and* $\hat{Y} < Y^*$.

(B) If (X^*, Y^*) *solves* $\text{Max}_{X,Y} U(X) - D(X-Y) - C(Y)$ *and* $(\hat{X}, \hat{Y})$ *solves* $\text{Max}_{X,Y} U(\cdot) - D(X-Y) - \hat{C}(Y)$, *then* $\hat{X} < X^*$ *and* $\hat{Y} < Y^*$.

(C) If (X^*, Y^*) *solves* $\text{Max}_{X,Y} U(X) - D(X-Y) - C(Y)$ *and* $(\hat{X}, \hat{Y})$ *solves* $\text{Max}_{X,Y} U(\cdot) - \hat{D}(X-Y) - C(Y)$, *then* $\hat{X} < X^*$ *and* $\hat{Y} < Y^*$.

Proof. See appendix.

Proposition 1 has some moral implications. Suppose that we want to reduce the level of the activity X, because we find it immoral. How might we achieve this goal? One option – as formalized by Proposition 1A – is to simply lower people's utility functions from engaging in the activity X. We could tax fur; we could splash paint on people who wear fur in public; or we could simply make fur illegal. All of these are simply coercive methods to stop an activity by lowering the welfare of those who engage in it.[6]

[5] Implicit in the Proof of Proposition 1 is the fact that each of the following results hold even without the convexity assumptions.
(A) If . . . then $\hat{X} \leqq X^*$.
(B) If . . . then $\hat{X} - \hat{Y} \leqq X^* - Y^*$.
(C) If . . . then $\hat{Y} < Y^*$.

[6] In reality, each of these options will not only induce changes in the material utility from an activity, but in the beliefs as well. In the next section, I will discuss how society's overall beliefs about the morality of an activity can affect the costs to an individual of maintaining given beliefs. Introducing this issue, it becomes clear that people might infer other people's beliefs from all of the actions listed above. If fur is made illegal, this is a signal that those in power believe it is immoral. It may thus become harder to convince yourself that wearing fur is okay, so that the cost function $C(\cdot)$ is likely to increase.

When considering social issues we may not be able to use such coercive methods, especially when a majority do not share our beliefs, and when culprits are well protected from harrassment. Another means of lowering the level of activity – as formalized in Proposition 1B – is to provide information about how bad the activity is. Presumably this information would raise the cost of convincing oneself that the activity is okay. The more we publicize the means by which fur is produced – with, say, graphic illustrations of how traps work – the harder it is for people to convince themselves that wearing fur is moral. A one-minute explanation of how veal calves are raised makes it substantially harder to convince oneself that eating veal does not contribute to torture. In terms of the model, educating people would increase the absolute and marginal values of the function $C(\cdot)$.

Finally, we could simply inculcate people with the general moral principle that they should do less of an activity that they know hurts others. In terms of the model, we can try to increase the function $D(\cdot)$. We could employ such moral pressure specifically – making people feel very bad about cruelty to animals. But increasing pressure to be moral is likely to be employed at a more general level: We can convince others that, in all of life's activities, they should be good to others. We can indoctrinate children with the golden rule without providing them with specific moral prescriptions.

Proposition 1 yields fairly straightforward relatively unsurprising results about how affecting individuals utility functions will affect behavior. In the next section, I illustrate a somewhat more surprising result. When each person's beliefs are influenced by the prevalent beliefs of society as a whole, one of the methods listed above for reducing the activity may backfire: making people feel worse for engaging in immoral activities may increase the level of immoral activities.

3. Social effects

In this section, I consider the possibility that the cost of maintaining certain beliefs can be affected by the prevailing beliefs in society. Formally, I assume that the average (arithmetic mean) of people's beliefs, $\bar{Y}$, enters into the function $C(\cdot)$, so that the cost of dissonance can now be written as $C(Y, \bar{Y})$. For simplicity, I assume that everybody has the same utility function, and that the society is 'large' enough such that no individual's beliefs substantially affects society's average beliefs.

By modeling each person's utility as depending on society's average beliefs, I am implicitly assuming that people's beliefs are directly revealed to each other. I discuss in the next section the idea that people's utilities may instead depend on the behavior of others.

Just as I assumed in section 2 that $C' > 0$ and $C'' > 0$, here $C_Y > 0$ and

$C_{YY}>0$. Furthermore, I shall assume that the less moral society believes an activity to be, the more difficult it is for each individual to convince himself that the activity is moral. In particular, I shall assume that $C_{\bar{Y}}<0$ and $C_{Y\bar{Y}}<0$; the higher $\bar{Y}$, the lower both the absolute and marginal cost of convincing yourself that an activity is more moral. If everybody around you believes that eating veal is moral, then it is easier to convince yourself of this.

We can formalize a person's beliefs and behavior as solving the following optimization problem:

$$\underset{X,Y}{\text{Max}}\, L(X, Y, \bar{Y}) \equiv U(X) - D(X - Y) - C(Y, \bar{Y});$$

where $X, Y \geqq 0; U'(X), D'(X-Y), C_Y(Y, \bar{Y})>0;$

and $U''(X)<0, D''(X-Y), C_{YY}(Y, \bar{Y})>0;$

and $C_{\bar{Y}}(Y, \bar{Y})<0$ and $C_{Y\bar{Y}}(Y, \bar{Y})<0.$

In the previous section the individual's behavior and beliefs were found directly by solving such an optimization problem; but because beliefs are now assumed to be dependent on society's overall beliefs, which in turn are derived from each individual's optimization problem, we must now solve the model by assuming the equilibrium condition that $Y=\bar{Y}$.

As before, if we change one of the component functions of a person's utility function, then that person changes his behavior and beliefs. Will adding 'social effects' change the conclusions from section 2? Will Conjectures A–C still be true? The main point of this paper is that Conjecture C may be false when social effects are incorporated: if we increase people's distaste for behaving in a way they feel is immoral, then they may behave more immorally, not less.

If everybody's propensity to experience cognitive dissonance increases, then the direct effect will be (as discussed in section 2) to decrease their level of immoral activities and to convince themselves that the activity is less immoral. But precisely because they each start to believe the activity is more morally defensible, the indirect effect will be to increase the level of the immoral activity. If this indirect effect outweighs the direct effect, then *an increase in the unpleasantness of moral dissonance can increase the level of an immoral activity.*

This result has some resonance for broader moral debates, and it illustrates a situation in which purist ethical standards may in the end lead to undesirable results. If as ethicists we are concerned that people behave morally, it may not always be best to constantly insist to them to do so.

The possibility that increasing the distaste for immorality may lead to more immorality is illustrated in the appendix. For a given numerical example, I show that a small increase in the pain caused by cognitive

dissonance will lead to an increase in the level of the immoral activity. This shows that adding the 'social effects' of beliefs can reverse the conclusions of Proposition 1C.

I show, however, that in the example, Propositions 1A and 1B still hold. In fact, Propositions 2A and 2B – presented formally in the appendix – show that, if we invoke a certain stability property, small changes in the utility function or the cost of changing beliefs will always cause the equilibrium levels of X and Y to shift in accordance with Proposition 1. The 'stability' of an equilibrium is motivated by the following consideration. Suppose that, for whatever reason, each person thought that everybody shared slightly different beliefs than him. How would that person change his beliefs? An equilibrium is unstable if such a small change in the perceived societal beliefs leads each individual to change his beliefs by more than that change. This is unstable because slight misperceptions about the prevalent beliefs will lead beliefs away from the equilibrium beliefs by more than the original misperceptions. Conversely, an equilibrium is stable if a small change in perceived societal beliefs would lead to a yet smaller change in actual societal beliefs.

Definition. An equilibrium pair (X, Y) is *stable* if and only if a small exogenous shift in each individual's perception of society's average beliefs leads to a smaller change in each individual's beliefs in the same direction.

Propositions 2A and 2B. In any stable equilibrium

(A) a small increase in the material utility of an activity will increase the level of the activity, and will increase the level of the activity that people think is moral; and

(B) a small increase in the cost of maintaining dishonest beliefs will decrease the level of the activity, and will decrease the level of the activity that people think is moral.

Propositions 2A and 2B parallel Propositions 1A and 1B. Because the equilibrium in the Example in the appendix is stable, I have already established that the analog of Proposition 1C does not hold.

Proposition 2C. There exist stable equilibria in which a small increase in the disutility from cognitive dissonance will increase the level of activity people engage in.

In fact, the formal version of Proposition 2C defined in the appendix shows that the conditions for an increase in the level of activity are easily characterized: the determining factor is whether the level of the activity that society believes to be moral will decrease the marginal cost of changing beliefs, when the social effects are taken into consideration. If this is the case, then a small increase in the disutility from cognitive dissonance will increase the level of an activity.

4. Extensions

In the previous section, I assumed that each individual's beliefs about the morality of an activity are influenced by beliefs of other people in society. In order for society's beliefs to be influential, however, these beliefs must be observable. Is this assumption realistic?

Trivers (1985) and Frank (1988) argue that people's beliefs are often revealed – even when they wish to hide those beliefs – through unconscious signals. Even if they control their words carefully, people's tone of voice, eyes, facial expressions, and body language give them away. As Frank quotes Nietzche, 'One can lie with the mouth, but with the accompanying grimace one nevertheless tells the truth.' If you witness an argument at a cocktail party about whether wearing fur is moral, you may often be able to tell whether the participants in the debate sincerely believe their arguments.

Indeed, Trivers (1985) makes clear that the type of moral self-justification considered in this paper may be a natural evolutionary response to exactly such difficulties in hiding one's true beliefs. If your true beliefs are observable to others, and if it is generally advantageous to convince others that you are moral, then it follows that that there is advantage in deceiving yourself; you are more likely to convince others that you are moral if you believe it yourself.

But there are reasons to believe that true beliefs will not in general be perfectly observable. And there is every reason to suppose that people won't voluntarily reveal their true beliefs; revelation of beliefs depends on the communicational structure of a society, as well as on psychological incentives people have to communicate their beliefs about morality in a misleading way.[7] Because behavior is more likely than beliefs to be observable, a natural alternative to the model of this paper would be one in which the cost to individuals of holding given beliefs depends on the average behavior of society.

We could formalize the idea that both beliefs and behavior might influence the cost of maintaining beliefs by modifying the cost function to be $C(Y, \alpha\bar{Y}+(1-\alpha)\bar{X})$, where $\bar{X}$ is society's average level of activity, and $\alpha \in [0,1]$ is a parameter capturing the relative importance of beliefs and behavior. The model of section 3 corresponds to setting $\alpha = 1$.

What would be the likely results if α were close to 0, so that people are more directly influenced by the observed behavior of others? The main result of this paper would not hold in such a model. The main result is derived because making people feel worse when being immoral induces people to convince themselves that a dubious activity is more moral, and this affects

[7] Kuran (1990) discusses the related issue of why, in social and political contexts, individuals are unlikely in general to fully reveal their preferences.

the ease with which people convince themselves that an activity is morally upright. When $\alpha=0$, however, the social effects from individuals modifying their beliefs is eliminated. Thus, increasing the cost to people of perceived immorality would always reduce the level of immoral activities.

I believe, however, that considering this modification of the model exaggerates the problems with the results of this paper. This is because the fact that society's beliefs do not directly affect people's behavior does not mean that they do not do so indirectly. In particular, in addition to observable behavior, most people choose their morally relevant activities within the context of societal debate; at cocktail parties and family get-togethers, people are often required in one way or another to 'announce' their moral stances. If you encounter friends debating about animal rights, you will often feel compelled to state your own beliefs. (And, even if you don't, your friends may infer much from your silence.)

Importantly, a person experiences cognitive dissonance if he argues a moral case different than his true beliefs. This is similar to the cognitive dissonance he experiences if he behaves immorally, but now it represents the dissonance he feels if he does not speak truthfully, given his perception that he is an honest person. This seems a plausible emotion: people often behave immorally, but feel bad if they do not acknowledge that their behavior is immoral. There were famous slaveholders in American history (such as Thomas Jefferson) who could not bring themselves to free their slaves, but felt compelled by their intellectual and moral needs to speak about the evils of slavery. If the need to speak honestly is powerful enough, the proper (reduced-form) model would be set $\alpha=1$, even if beliefs are not directly observable.

There is also, however, an incentive for people to be dishonest in discussing their feelings about social issues: they feel like hypocrites if they announce beliefs about what is moral without behaving according to those beliefs. We feel like hypocrites if we say that wearing fur is immoral, yet we wear fur. Moreover, the cost of hypocrisy may not be totally internal. If behavior is observable and if society punishes hypocrisy, then the cost of hypocrisy may be greater than the internal cost.[8]

In announcing his beliefs, therefore, a person will weigh 'announcement dissonance' against the disutility of hypocrisy. A person will choose an

[8] Frank (1988) is instructive on why people are generally disposed to convince others that they are moral. A disposition to be moral will signal to others that you tend not to behave opportunistically in relationships. This also helps explain the development of emotional needs to be honest – it commits you to not exploit others by lying to them when convenient. Of course, there are times when appearing nasty may be advantageous, but being observably pre-disposed in behaving for others' benefits will generally help a person to involve himself in advantageous cooperation.

announcement in between X and Y to balance these two factors. If the cost of a person convincing himself that something is moral now depends on society's average announced beliefs, then the results of this paper may hold when 'announcement dissonance' is strong enough to induce announcements to be close to $\bar{Y}$.

When considering the cost of hypocrisy, another interesting possibility arises that is much like the result in section 3. In particular, punishing people severely for being hypocritical – for not practicing what they preach – might lead to more rather than less of an immoral activity. This is because the direct effect of decreasing X – people will engage in less of the activity so as to reduce hypocrisy given what they are announcing – may be outweighed by the indirect effect – people will preach less forcefully against the immoral activity, and this will make it easier for the rest of society to convince itself that the activity is moral. If we condemn as hypocritical those who declare themselves vegetarians for ethical reasons while continuing to wear leather shoes, we may end up with fewer vegetarians and more leather-shoe wearers.

Condemning hypocrisy can also be damaging in political debate. Consider Ted Kennedy, who advocates redistribution of income from the wealth to the poor. He must reconcile this view – obviously sincere – with his behavior: he has not unilaterally redistributed his money to the poor. While he and other wealthy liberals may rationalize this seeming hypocrisy in various ways (e.g., that their money cannot do any good in the absence of other people's money, or that 'horizontal equity' is a meaningful and important social good), it seems to me that the simple explanation is that they truly believe what they are advocating, but are boundedly generous in sacrificing their own material well-being. The model indicates that, if you care about promoting positive social change, it might be counterproductive to punish such hypocrisy too severely. While convincing Ted Kennedy to give more of his money before letting him speak out against poverty might increase total assistance to the poor, more likely it will backfire. If we told Ted Kennedy to put up or shut up, he would probably shut up, and there would be fewer voices around to convince us that we should redistribute more of our money.

5. Conclusion

This paper uses the rational-choice approach traditional among economists to model a largely non-economic issue. Since Becker's (1981) seminal work, rational-choice analysis has become more popular as a means of studying phenomena not traditionally studied by economists. But while the traditional economists' approach to analyzing social phenomena is a productive avenue of research, its usefulness is negated when we do not expand our conception of human motivation. Overly materialistic models miss out

on important aspects of important social decision-making mechanisms.[9] This paper follows several other recent papers in adding the important psychological issue of cognitive dissonance, and uses this assumption to help us understand a realm where assuming people pursued solely their material self-interest would be profoundly misleading.

Appendix

Proof of Proposition 1. By the convexity assumptions, we know that there is a unique optimal pair (X, Y) that solves each maximization problem involved in the proposition.

Proof of part A. Because $(X^*, Y^*) = \text{argmax}_{X,Y}(U(X) - D(X-Y) - C(Y)$, $U(X^*) - D((X^*, Y^*) - C(Y^*) > U(\hat{X}) - D(\hat{X} - \hat{Y}) - C(\hat{Y})$. Because $(\hat{X}, \hat{Y}) = \text{argmax}_{X,Y}\hat{U}(X) - D(X-Y) - C(Y)$, $\hat{U}(\hat{X}) - D(\hat{X} - \hat{Y}) - C(\hat{Y}) > \hat{U}(X^*) - D(X^* - Y^*) - C(Y^*)$. Combining these two facts, we get that $U(X^*) - U(\hat{X}) - \hat{U}(X^*) - \hat{U}(\hat{X})$. But because $U'(\cdot) > \hat{U}'(\cdot)$ everywhere, and both functions are increasing, this can only be true if $X^* > \hat{X}$.

Because $(X^*, Y^*) = \text{argmax}_{X,Y}U(X) - D(X-Y) - C(Y), U(X^*) - D(X^* - Y^*) - C(Y^*) > U(X^*) - D(X^* - \hat{Y}) - C(\hat{Y})$. Because $(\hat{X}, \hat{Y}) = \text{argmax}_{X,Y}\hat{U}(X) - D(X-Y) - C(Y)$, $\hat{U}(\hat{X}) - D(\hat{X} - \hat{Y}) - C(\hat{Y}) > \hat{U}(\hat{X}) - D(\hat{X} - Y^*) - C(Y^*)$. These in turn imply that $C(\hat{Y}) - C(Y^*) > D(X^* - Y^*) - D(X^*, \hat{Y})$ and $C(Y^*) - C(\hat{Y}) > D(\hat{X} - \hat{Y}) - D(\hat{X} - Y^*)$. Multiplying the second of these by negative 1, we get that $D(\hat{X} - Y^*) - D(\hat{X}, \hat{Y}) > D(X^*, Y^*) - D(X^*, \hat{Y})$. But because D is increasing and convex, and because we have already proven that $X^* > \hat{X}$, this implies that $Y^* > \hat{Y}$.

Proof of part B. Because $(X^*, Y^*) = \text{argmax}_{X,Y}U(X) - D(X-Y) - C(Y)$, $U(X^*) - D(X^* - Y^*) - C(Y^*) > U(\hat{X}) - D(\hat{X} - \hat{Y}) - C(\hat{Y})$. Because $(\hat{X}, \hat{Y}) = \text{argmax}_{X,Y}U(X) - \hat{D}(X-Y) - C(Y)$, $U(\hat{X}) - \hat{D}(\hat{X} - \hat{Y}) - C(\hat{Y}) > U(X^*) - \hat{D}(X^* - Y^*) - C(Y^*)$. Combining these two facts, we get $D(\hat{X} - \hat{Y}) - D(X^*, Y^*) > \hat{D}(\hat{X} - \hat{Y}) - \hat{D}(X^*, Y^*)$. Because D and $\hat{D}$ are each increasing, and because D derivative dominates $\hat{D}$, this implies that $\hat{X} - \hat{Y} < X^* - Y^*$. (This

[9] This literature began with Akerlof and Dickens (1982), which considers the effects of cognitive dissonance in a labour market where job safety is a consideration. Dickens (1986) studies the effects of adding cognitive dissonance to the economic approach to crime and punishment. Montgomery (1989) combines psychology, sociology, and economics to consider the economic effects on the underclass of the moral imperitive to work. Akerlof (1989) discusses the role of cognitive dissonance in political choices over the level of public goods to provide. Work by Kuran (1989, 1990, 1991) has considered variously psychological and social effects on group phenomena such as revolutions.

step is not necessary for the following proof, but because it utilized no convexity assumption, it proves the proposition of footnote 5.

Because $(X^*, Y^*) = \text{argmax}_{X,Y} U(X) - D(X-Y) - C(Y), U(X^*) - D(X^* - Y^*) - C(Y^*) > U(\hat{X}) - D(\hat{X} - Y^*) - C(Y^*)$. Because $(\hat{X}, \hat{Y}) = \text{argmax}_{X,Y} U(X) - \hat{D}(X-Y) - C(Y)$, $U(\hat{X}) - \hat{D}(\hat{X} - \hat{Y}) - C(\hat{Y}) > U(X^*) - \hat{D}(X^* - \hat{Y}) - C(\hat{Y})$. These in turn imply that $D(\hat{X} - Y^*) - D(X^* - Y^*) > U(\hat{X}) - U(X^*)$ and $D(X^* - \hat{Y}) - \hat{D}(\hat{X} - \hat{Y}) > U(X^*) - U(\hat{X})$. Combining these two facts, we get $D(\hat{X} - Y^*) - D(X^* - Y^*) > \hat{D}(\hat{X} - \hat{Y}) - \hat{D}(X^* - \hat{Y})$. Suppose that $\hat{Y} > Y^*$. Then because D and $\hat{D}$ are each increasing, and because $\hat{D}$ derivative dominates D, this implies that $\hat{X} < X^*$. Because we know that $\hat{X} < X^*$ if $\hat{Y} \geqq Y^*$, this proves that $\hat{X} < X^*$.

Because $(X^*, Y^*) = \text{argmax}_{X,Y} U(X) - D(X-Y) - C(Y), U(X^*) - D(X^* - Y^*) - C(Y^*) > U(X^*) - D(X^* - \hat{Y}) - C(\hat{Y})$. Because $(\hat{X}, \hat{Y}) = \text{argmax}_{X,Y} U(X) - \hat{D}(X-Y) - C(Y)$, $U(\hat{X}) - \hat{D}(\hat{X} - \hat{Y}) - C(\hat{Y}) > U(\hat{X}) - \hat{D}(\hat{X} - Y^*) - C(Y^*)$. These in turn imply that $C(\hat{Y}) - C(Y^*) > D(X^*, Y^*) - D(X^* - \hat{Y})$ and $C(Y^*) - C(\hat{Y}) > \hat{D}(\hat{X} - \hat{Y}) - \hat{D}(\hat{X} - Y^*)$. Combining these two facts, we get $D(X^* - \hat{Y}) - D(X^* - Y^*) > \hat{D}(\hat{X} - \hat{Y}) - \hat{D}(\hat{X} - Y^*)$. Because D and $\hat{D}$ are each increasing, and because $\hat{D}$ derivative dominates D, and because we know that $\hat{X} < X^*$, this implies that $\hat{Y} > Y^*$.

Proof of part C. Because $(X^*, Y^*) = \text{argmax}_{X,Y} U(X) - D(X-Y) - C(Y)$, $U(X^*) - D(X^* - Y^*) - C(Y^*) > U(\hat{X}) - D(\hat{X} - \hat{Y}) - C(\hat{Y})$. Because $(\hat{X}, \hat{Y}) = \text{argmax}_{X,Y} U(X) - D(X-Y) - \hat{C}(Y)$, $U(\hat{X}) - \hat{D}(\hat{X} - \hat{Y}) - \hat{C}(\hat{Y}) > U(X^*) - D(X^* - Y^*) - \hat{C}(Y^*)$. Combining these two facts, we get $\hat{C}(Y^*) - \hat{C}(\hat{Y}) > C(Y^*) - C(\hat{Y})$. Because C and $\hat{C}$ are each increasing, and because $\hat{C}$ derivative dominates C, this implies that $Y^* > \hat{Y}$.

Because $(X^*, Y^*) = \text{argmax}_{X,Y} U(X) - D(X-Y) - C(Y), U(X^*) - D(X^* - Y^*) - C(Y^*) > U(\hat{X}) - D(\hat{X} - Y^*) - C(Y^*)$. Because $(\hat{X}, \hat{Y}) = \text{argmax}_{X,Y} U(X) - D(X-Y) - \hat{C}(Y)$, $U(\hat{X}) - \hat{D}(\hat{X} - \hat{Y}) - \hat{C}(\hat{Y}) > U(\hat{X}) - D(\hat{X} - Y^*) - \hat{C}(Y^*)$. These in turn imply that $D(\hat{X} - Y^*) - D(X^* - Y^*) > U(\hat{X}) - U(X^*)$ and $D(X^*, \hat{Y}) - D(\hat{X} - \hat{Y}) > U(X^*) - U(\hat{X})$. Combining these two facts, we get $D(\hat{X} - Y^*) - D(X^* - Y^*) > D(\hat{X} - \hat{Y}) - D(X^* - \hat{Y})$. Because D is increasing and convex, and because we know that $Y^* > \hat{Y}$, this implies that $X^* > \hat{X}$. Q.E.D.

Example

$$\text{Max}_{X,Y} L(X, Y, \bar{Y}) \equiv U(X) - D(X-Y) - C(Y, \bar{Y}),$$

where $U(X) = 3\ln(X)$,

$D(X-Y) = 2(X-Y)^2$,

and $C(Y, \bar{Y}) = 2Y^2/(\bar{Y}^2 + 5\bar{Y}/4)$.

In order to know what each individual will do, we must know what, on average, society believes. But because $\bar{Y}$ is the average of the Y chosen by all individuals, and all individuals are identical, each individual must be choosing $\bar{Y}$. Therefore, we must then find a pair (X^*, Y^*) that maximizes $L(X, Y, \bar{Y})$ given society believes $\bar{Y} = Y^*$. Only such pairs will form equilibria, in the usual sense that no individual wishes to change his behavior or beliefs given the behavior and beliefs of everybody else.

Using the two first-order conditions – $U'(X) = D'(X - Y)$ and $D'(X - Y) = C_Y(Y, Y)$ – it can be shown that the only pair that forms an equilibrium in this example is $X^* = 3$ and $Y^* = 11/4$.[10]

What happens if we increase the propensity for cognitive dissonance in Example 1? One way of considering this is to rewrite the maximization equation as $\text{Max}_{X,Y} L(X, Y, \bar{Y}) \equiv U(X) - \delta D(X - Y) - C(Y, \bar{Y})$, where $\delta = 1$. We can then ask: what happens when we slightly increase δ? If the resulting changes in the equilibrium values of X increases, then this indicates the small increases in the unpleasantness of cognitive dissonance will increase, rather than decrease, the level of the immoral activity.

We can calculate how X^* and Y^* change as a function of δ by totally differentiating the first-order conditions. Note that, by definition of an equilibrium, $\bar{Y}$ will change exactly as much as Y.

Totally differentiating at $\delta = 1$ yields

$$U''(X)\,dX = D'(X - Y)\,d\delta + D''(X - Y)\,dX - D''(X - Y)\,dY, \quad \text{and}$$

$$D'(X - Y)\,d\delta + D''(X - Y)\,dX - D''(X - Y)\,dY \equiv C_{YY}(Y, Y)\,dY - C_{Y\bar{Y}}(Y, Y)\,dY.$$

Solving these two equations at $X = 3$ and $Y = 11/4$ yields the result that $dX/d\delta = 1$. That is, in this equilibrium, a small increase in the pain caused by cognitive dissonance will lead to an increase in the level of the activity. This shows that adding the 'social effects' of beliefs can reverse the conclusions of Proposition 1C (because Proposition 1C implies that $dX/d\delta < 0$).

Can Propositions 1A and 1B also be reversed? In this example, they are not. Using the same approach as for cognitive dissonance, we can see how small changes in μ, where $U(X) = \mu U(X)$, or in χ, where $C(Y, \bar{Y}) = \chi C(Y, \bar{Y})$, affect the equilibrium levels of X. Solving for these, we get $dX/d\mu = 9/4$ and $dX/d\chi = -9$. Observe that each of these are consistent with the original conjectures and with Proposition 1: $dX/d\mu > 0$ means that if you increase the utility from engaging in the activity, a person will engage in more of that activity; $dX/d\chi < 0$ means that if you increase the cost to a person of

[10] The earlier convexity assumptions guarantee that the second order conditions for a maximum will hold whenever the first-order conditions hold.

convincing himself that an immoral activity is moral, he will engage in less of the activity.

Definition of stability. (X, Y) form a stable equilibrium if $0 \leqq dY/d\bar{Y} \leqq 1$, where dY and $d\bar{Y}$ are derived from the following equations:

$$U''(X)\,dX = D''(X-Y)\,dX - D''(X-Y)\,dY \quad \text{and}$$

$$D''(X-Y)\,dX - D''(X-Y)\,dY = C_{YY}(Y, Y)\,dY - C_{Y\bar{Y}}(Y, Y)\,d\bar{Y}.$$

Lemma. An equilibrium (X, Y) *is stable if and only if* $C_{YY}(Y, Y) + C_{Y\bar{Y}}(Y, Y) > U''(X)D''(X-Y)/[D''(X-Y) - U''(X)]$.

Proof. Algebra from the definition.

Using the Lemma, Propositions 2A, 2B, and 2C can be stated:

Proposition 2. In any stable equilibrium (X^*, Y^*),
(A) $dX/d\mu > 0$ *and* $dY/d\mu > 0$; *and*
(B) $dX/d\chi < 0$ *and* $dY/d\chi < 0$.

Proof. Algebra shows that

$$dX/d\mu = [U'(\cdot)(D''(\cdot) + C^*(\cdot))]/[C^*(\cdot)(D''(\cdot) - U''(\cdot)) - U''D''(\cdot)],$$

and

$$dX/d\mu = D''(\cdot)C^*(\cdot)/[U''(\cdot)D''(\cdot) - C^*(\cdot)(D''(\cdot) - U''(\cdot)],$$

where $C^*(\cdot) = C_{YY}(\cdot, \cdot) + C_{Y\bar{Y}}(\cdot, \cdot)$, and where all functions are calculated at their equilibrium values. Further algebra shows that, for any stable equilibrium, it must be that $dX/d\mu > 0$ and $dX/d\chi < 0$. Further algebra yet yields the results for $dY/d\mu$ and $dY/d\chi$. Q.E.D.

Propositions 2A and 2B parallel Propositions 1A and 1B.[11] Because one can readily verify that the equilibrium in the earlier example is stable, I have already established that the analog of Proposition 1C does not hold. In fact, the conditions for $dX/d\delta$ to be negative are easily characterized. Proposition

[11] Note that, while the Example had a unique equilibrium, there is no reason in general for uniqueness. The earlier convexity assumptions do not guarantee uniqueness, because of the 'feedback loop' from changes in Y leading to changes in $\bar{Y}$. As such, Proposition 2 does not fully correspond to Proposition 1, which held globally, not just for small changes in the component functions. (Also note that I have not proven the existence of a stable equilibrium; I do not think characterizing conditions for existence would be particularly useful in this context.)

2C states these conditions, and asserts that the conditions hold for Example 1.

Proposition 2C. In a stable equilibrium, $dX/d\delta<0$ *if and only if* $C^*(\cdot)>0$, *and there are cases in which* $C^*(\cdot)>0$ *in equilibrium.*

Proof. Algebra shows that

$$dX/d\delta = D'(\cdot)C^*(\cdot)/[C^*(\cdot)(U''(\cdot)-D''(\cdot))+U''(\cdot)D''(\cdot)].$$

If the equilibrium is stable, the denominator is negative. Because $D'(\cdot)$ is always positive, the numerator is positive if and only if $C^*(\cdot)$ is positive. This means that $dX/d\delta<0$ if and only if $C^*(\cdot)>0$.

Moreover, algebra shows that $C^*(\cdot)$ in the equilibrium of Example 1.

Q.E.D.

References

Akerlof, George A., 1989, The economics of illusion, Economics and Politics 1, 1–15.

Akerlof, George A. and William T. Dickens, 1982, The economic consequence of cognitive dissonance, American Economic Review 72, 307–319.

Aronson, Elliot, 1980, The social animal (W.H. Freeman, San Francisco).

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

Dickens, William T., 1986, Crime and punishment again: The economic approach with a psychological twist, Journal of Public Economics 30, 97–107.

Frank, Robert, 1988, Passions with reason: The strategic role of the emotions (W.W. Norton & Company, New York).

Kuran, Timur, 1989, Sparks and prairie fires: A theory of unanticipated political revolution, Public Choice 61, 41–74.

Kuran, Timur, 1990, Private and public preferences, Economics and Philosophy 6, 1–26.

Kuran, Timur, 1991, Cognitive limitations and preference evolution, Journal of Institutional and Theoretical Economics 147, June, 241–273.

Montgomery, James D., 1990, Revisiting Talley's Corner: Mainstream norms, cognitive dissonance, and underclass behavior, forthcoming, Rationality and Society.

Trivers, Robert, 1985, Social evolution (Benjamin/Cummings, Menlo Park, CA).