The Whip and the Bible: Punishment Versus Internalization[☆]

Rohan Dutta[1], David K. Levine[2], Salvatore Modica[3]

Abstract

A variety of experimental and empirical research indicate that prosocial behavior is important for economic success. There are two sources of prosocial behavior: incentives and preferences. The latter, the willingness of individuals to "do their bit" for the group, we refer to as internalization. We do not view internalization as a fixed constant but something that a group can invest in. This means that there is a trade-off between using incentives and internalization to encourage prosocial behavior. By examining this trade-off we shed light on the connection between social norms observed inside the laboratory and those observed outside in the field. For example, we show that increasing the benefit of cooperation outside the laboratory can lower the use of incentives inside the laboratory even as it increases their usage outside.

JEL Classification Numbers: A1, D7, D9

Keywords: endogenous social norms, ostracism, Lucas critique, experimental economics

[☆]First Version: October 14, 2017. We would like to thank Marco Casari, Andrea Ichino, Andrea Mattozzi, Rohini Somanathan and seminar audiences at WUSTL, Warwick, Queen's, the Paris Institutions Conference, the Zurich Political Economy Conference, Delhi School of Economics and the University of Trento. We gratefully acknowledge support from the EUI Research Council.

[*]Corresponding author David K. Levine

Email addresses: rohan.dutta@mcgill.ca (Rohan Dutta), david@dklevine.com (David K. Levine), salvatore.modica@unipa.it (Salvatore Modica)

[1]Department of Economics, McGill University

[2]Department of Economics, EUI and WUSTL

[3]Università di Palermo, Dipartimento SEAS

"it is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner" Adam Smith

"Teach self-denial and make its practice pleasure, and you can create for the world a destiny more sublime that ever issued from the brain of the wildest dreamer." Sir Walter Scott

1. Introduction

A variety of experimental and empirical research indicate that prosocial behavior is important for economic success. There are two sources of prosocial behavior: incentives and preferences. People may behave prosocially because failure to do so may result in punishment by others. But even in the absence of incentives people may behave prosocially out of ethical considerations that determine their preferences: we refer to this as internalization. There is evidence for both types of behavior. For example, people make altruistic choices in double-blind treatments in the laboratory where there is no possibility of punishment or reward. The pages of history are filled with tales of great individual sacrifices for the common good. On the other hand people are not always guided by societal needs, which is why rewards and punishments exist: we do have both murderers and prisons. Here we develop a theory of group behavior in which both sources of prosocial behavior coexist and are endogenously determined. The questions we address are when we are likely to see incentives rather than internalization, whether they are complements or substitutes, and what the implications are for economic problems and empirical research.

A key feature of our theory is that internalization is not possible on a case by case basis while punishment is. That is either we produce people who are prosocial or we do not, but putting them in a particular environment should not have any impact on this one way or the other. By contrast, incentives can be adapted to circumstances: there is no reason that the same incentive system should be used by a business firm as by a political party. Consequently the level of internalization is determined by the most important problems faced by a group. This poses issues for inferring behavior in the large from behavior in the small. Inside the laboratory, for example, we expect internalization to be exogenous but punishment endogenous. What does behavior observed in the laboratory then tell us about behavior outside the laboratory where both internalization and punishment are endogenous?

Our theory combines several standard elements. We follow the ethical voter literature[4] and the experimental literature on warm-glow giving[5] in assuming that there is a fraction of the group, the acolytes, who lose utility for failing to do their social duty. Second, we follow a long empirical literature - in particular Coase and Ostrom (1990) - that argues that groups are good at providing incentives to their members to achieve group objectives, effectively solving mechanism design problems. The type of incentives we study are punishments as in Ostrom (1990) and Fehr and Gachter (2000). These might be social punishments such as ostracism, or even monetary punishments such

[4]See particularly Feddersen and Sandroni (2006).

[5]See particularly Andreoni (1990) and Palfrey and Prisbrey (1997).

as fines. We model monitoring following Levine and Modica (2016): this model has been used by Levine and Modica (2017) to study lobbying groups and by Levine and Mattozzi (2017) to study political parties. In this model there is a noisy signal of individual behavior and the possibility of imposing punishments based on these signals.

Unlike existing models in which the fraction of acolytes is fixed, we allow groups to make costly investments in producing acolytes. We do not think there is any mystery in this. Prosocial behavior is learned and taught: by our parents, in school, and by our peers. Internalization in this view is an investment by society in changing preferences. Like social norms themselves we view this investment as endogenous and functional and ask how a group making collective decisions optimally invests in internalization.

In our setting we distinguish between the primary problem faced by a group, a stylized public good problem, and the secondary problem, which in our applications is a laboratory experiment. The latter is much less likely or much less frequent - hence plays little role in determining the investment in acolytes.

There are several takeaways from the theory. The first is that internalization can have large effects by complementing punishment. This is especially the case when it is difficult to provide incentives to monitors. Because acolytes are willing to accept small costs to engage in honest monitoring this can be leveraged to provide incentives through punishment. There are also important differences between the primary and the secondary problem. In the secondary problem there can be "excess" internalization - that is, it may be possible to achieve the first best without any monitoring cost simply by having acolytes engage in production. This cannot happen in the primary problem: there it would never pay to achieve the first best.

One of the key issues we examine is how changing the primary affects the solution of the secondary. Consider increasing the value of the public good in the primary. This will generally increase internalization and this will spill over into the secondary. Hence if we observe societies with different primaries and compare the same secondary, for example, in a laboratory experiment, we will observe different outcomes. In particular, as the value of the public good in the primary goes up and so does internalization, in the secondary we will first observe little punishment as there are few acolytes willing to punish, then punishment will go up as more acolytes are available, then decrease as the burden of production is born by acolytes. In contrast in the primary increasing the value can never lower the level of punishment. Hence if we measure punishment in the secondary it need not be related to punishment in the primary.

We consider specific applications to the classical secondary problem, the laboratory experiment. We engage in a quantitative calibration similar to that in the behavioral economics literature - see in particular Levine (1986) and Fehr and Schmidt (1999). We first consider the classical public good experiment with punishment analyzed by Fehr and Gachter (2000) as this is similar to the type of public goods game in our basic model. In that experiment it is observed that while the average contribution is very low when there is no punishment, roughly half of the group are willing to bear the cost of punishment and by doing so induce substantial contributions. We show that

this result is well explained by a simple calibration of our model.

Second, we examine dictator and ultimatum experiments. These are not ideal from our point of view since it is not entirely clear what the underlying mechanism design problem is, but the experiments are the only ones where substantial cross-cultural data is available. We observe that risk aversion will create a mechanism design problem in which there is demand for "fairness" and that several other considerations point to a social objective function of this type. Using that idea we find that the results of dictator giving are quantitatively consistent with the Fehr and Gachter (2000) public goods data and that we can give a reasonable quantitative explanation of ultimatum bargaining as well. In both Fehr and Gachter (2000) and the ultimatum data there is evidence both that the participants are trying to solve a mechanism design problem and that in the limited time available are not completely successful in doing so. In particular, in both cases we find evidence of inefficient "over" punishment. Without communication and trying by trial and error to establish a social norm, we do not find this surprising.

Our final application is the cross-cultural ultimatum data from Henrich et al (2001). Here we have substantial cross country variation in the value of the public good in the primary. Our theory predicts that when this is low we see very bad offers and few rejections. In the middle range we will see good offers and substantial rejections and this will be insensitive to variation in the value of the public good. Finally at the upper end offers will be very good and rejections very few again. This is indeed what we find in the data.

2. Economic Environment

We study an organized group with many members engaging in a representative producer-recipient-monitor interaction. There are two possible states: the *primary* state $s = 1$ with probability $q > 0$ - which we interpret as the "normal " state of affairs, with q close to 1 - and the *secondary* state $s = 2$ with probability $1 - q \geq 0$, which is much less likely, for instance a laboratory experiment. After the state is known the producer chooses an amount $x_s \geq 0$ to produce at unit marginal cost. Output represents a public good providing a social benefit to the recipient of $V_s f(x_s)$ where $V_s > 0$ is a measure of the value of the public good and f is smooth and strictly differentiably increasing and concave.[6]

The effect of any individual member in a large group on average output is negligible, so there is a severe free-rider problem. We have modeled this by separating the recipient from the producer. Hence a selfish producer would prefer not to produce at all. We are going to assume that peer pressure can be used to provide producer incentives: production can be monitored and those who fail to produce can be punished. Specifically, in state s the group may establish an output quota y_s and generate a noisy signal $z_s \in \{0, 1\}$ about whether the producer respected the quota (that is $x_s \geq y_s$), where 0 means "good, respected the quota" and 1 means "bad, failed to respect the quota." If the quota is not satisfied the signal takes on the value 1 with probability one and if it

[6] That is $f' > 0$ and $f'' < 0$.

is satisfied it takes on the value 1 with probability $\pi_s < 1$. This simple stark signal technology works well in our quantitative analysis and our qualitative analysis is robust to more general error processes.

Along with the producer and recipient there is an anonymous monitor who if the signal about the producer is bad chooses whether or not to transmit it. If a bad signal is transmitted the group imposes an endogenous utility penalty $P_s \geq 0$ on the producer. This may be in the form of ostracism or some other social penalty. This punishment is also costly for the monitor who bears a cost of $\psi_s P_s$ where $\psi_s > 0$. Notice that since $\psi_s > 0$ selfish monitors will never transmit the signal.

A *social norm in state* s consists of an output quota y_s, a *de jure* output target $Y_s \geq y_s$, and a punishment level P_s. The group faces a mechanism design problem consisting of a choice of social norm for each state and an initial choice of the fraction of the group $0 \leq \phi \leq 1$ who internalize the social norms in the sense we will shortly specify. These individuals are referred to as *acolytes* and the remainder of the group as *opportunists*. Whether or not an individual is an acolyte is private information.

A social norm is only meaningful if group members are willing to adhere to it. In this context that means that it is incentive compatible for acolytes to produce Y_s, opportunists to produce y_s and for acolytes to transmit a bad signal. Incentive compatibility is defined with respect to an internalization penalty $\gamma > 0$: any acolyte who does not follow the social norm suffers a penalty of that amount. In other words, an acolyte producer who fails to hit the *de jure* target or an acolyte monitor who fails to transmit a bad signal loses utility γ. This can be interpreted as guilt for violating the social norm. Opportunists suffer no penalty. When a social norm is followed in state s each type of producer meets the output quota. The probability of generating a bad signal is therefore π_s and the probability that this signal is transmitted and the producer is punished is equal to the probability that the monitor is an acolyte, that is ϕ; the social cost of this punishment is the cost to the producer plus the cost to the monitor $P_s + \psi_s P_s$. Therefore the expected social utility under norm (y_s, Y_s, P_s) given ϕ is

$$U_s = V_s f((1-\phi)y_s + \phi Y_s) - ((1-\phi)y_s + \phi Y_s) - \phi \pi_s (1+\psi_s) P_s.$$

Prior to the realization of the state the group invests in indoctrination: the greater the investment in indoctrination the more acolytes ϕ there will be. Such investment is costly: the social cost is $H\phi$. A *social mechanism* consists of an initial investment ϕ and contingent social norms (y_s, Y_s, P_s). As all group members *ex ante* share the same interest we assume that the group collectively chooses the incentive compatible social mechanism that provides the greatest *ex ante* expected utility to group members. That is, the group collectively chooses the social mechanism that maximizes *social utility* $W = qU_1 + (1-q)U_2 - H\phi$. We refer to this as an *optimal social mechanism.* To avoid an uninteresting special case we will make a generic assumption about the primary state: $H \neq \gamma(1+\psi_1)\pi_1/(1-\pi_1)$.

2.1. Preliminaries

In the sequel the state will be often clear from context: in these cases we will omit the state subscript. We will need repeatedly to solve optimization problems involving output $f(x)$, of the form $\max_\theta Vf(a+b\theta) - c\theta$ subject to $0 \leq \theta \leq \Theta$. For this purpose it is convenient to define a function $g(\mu, x, X)$ called the *demand function* as follows

1. for $\mu > f'(x)$ set $g(\mu, x, X) = x$
2. for $\mu < f'(x+X)$ set $g(\mu, x, X) = x + X$
3. for $f'(x) \geq \mu \geq f'(x+X)$ set $g(\mu, x, X) = [f']^{-1}(\mu)$.

Lemma 1. *The solution to* $\max_\theta Vf(a+b\theta) - c\theta$ *subject to* $0 \leq \theta \leq \Theta$ *is unique and given by* $\theta = (1/b)\left(g((1/V)(c/b), a, b\Theta) - a\right)$. *The function* $g(\mu, x, X)$ *is continuous and increasing*[7] *in* x, X. *It satisfies* $x \leq g(\mu, x, X) \leq x + X$ *and for* $x < g(\mu, x, X) < x + X$ *it is smooth and strictly decreasing in* μ.

Proof. The first part follows directly from the definition and the fact that $f'(x)$ was assumed to be strictly decreasing. For the second part the derivative of the objective is $Vbf'(a+b\theta) - c$. At an interior solution this gives $a+b\theta = [f']^{-1}((1/V)(c/b))$, as stated. If $Vbf'(a) - c < 0$ we have a corner solution at $\theta = 0$. We can write this as $(1/V)(c/b) > f'(a)$ which is to say $g((1/V)(c/b), a, b\Theta) = a$. Finally, if $Vbf'(a+b\Theta) - c > 0$ we have a corner solution at $\theta = \Theta$. We can write this as $(1/V)(c/b) < f'(a+b\Theta)$ which is to say $g((1/V)(c/b), a, b\Theta) = a + b\Theta$. □

Finally, it will be convenient to define $\varphi = (Y - y)/\gamma$. Hence $Y = y + \varphi\gamma$ and expected output is given by $x \equiv (1-\phi)y + \phi Y = y + \phi\varphi\gamma$. In this notation, omitting state subscript, a norm (y, Y, P) can be written as (y, φ, P), and

$$U = Vf(y + \phi\varphi\gamma) - (y + \phi\varphi\gamma) - \phi\pi(1+\psi)P.$$

3. The Second Stage Problem: Optimal Social Norms

For any given ϕ the optimal social norm (φ, y, P) should be chosen in each state to maximize expected utility U. We omit the state subscript. A key idea in the choice of an optimal social norm is encapsulated in the *marginal cost of monitoring*

$$M \equiv (1+\psi)\frac{\pi}{1-\pi}. \tag{3.1}$$

This measures the marginal cost of increasing output y by opportunists due to the need to punish them. It consists of two parts: the first $1+\psi$ is the social cost of punishment, the second $\pi/(1-\pi)$ measures the difficulty of monitoring. Notice that the numerator π plays a key role: it measure the amount of punishment that takes place on the equilibrium path - that is erroneous punishment.

[7] For brevity increasing and decreasing without qualification always mean weakly so.

Theorem 1. *At the optimal solution if $\phi = 0$ then $y = 0$ and φ,P do not matter. When $\phi > 0$ then φ maximizes $Vf(\phi\gamma\varphi) - \phi\gamma\varphi = Vf(x) - x$ so is given by $\varphi = (1/(\phi\gamma))g(1/V, 0, \phi\gamma)$ and*

1. If $Vf'(\phi\gamma) \leq 1$ the optimal solution is first best *with $y = 0, P = 0$.*

2. If $Vf'(\phi\gamma) > 1$ the solution is second best *with $\varphi = 1$ and y maximizes $Vf(y+\phi\gamma)-(1+M)y$ so is given by*

$$y = g\left(\frac{1+M}{V}, \gamma\phi, \frac{(1-\pi)\phi\gamma}{\psi}\right) - \gamma\phi.$$

In all cases

$$P = \frac{y}{\phi(1-\pi)}.$$

Moreover, the maximized utility is concave and increasing in ϕ. Finally, $y \leq (1-\pi)\phi\gamma/\psi$.

The theorem says that we should first use the acolytes to provide output since there is no monitoring cost associated with their providing up to $\phi\gamma$ of output. It may be that this is enough in the sense that the first best of maximizing $Vf(\phi\varphi\gamma) - \phi\varphi\gamma$ is achieved for $\varphi < 1$. If V is small then indeed the first best is obtainable using acolyte production alone. As we increase V eventually at optimum $\varphi = 1$. At this point further increases in V do not increase output until $Vf'(\phi\gamma) - 1$ becomes equal to M the marginal cost of monitoring. Further increases in V then raise output as it becomes optimal to use acolytes as monitors and provide incentives to non-acolytes to produce. In the second stage problem punishment serves as a costly backstop technology to making use of acolytes who have internalized the social norm.

Proof. Consider first the production problem. Observe that the probability of being punished is equal to the probability that the monitor is an acolyte times the probability of a bad signal. Hence for an opportunist the cost of meeting target y is $y + \phi\pi P$, while the best alternative of producing zero costs ϕP, resulting in the incentive constraint $y + \phi\pi P \leq \phi P$ or $y \leq \phi(1-\pi)P$. On the other hand whenever it is incentive compatible for an opportunist to produce y it is incentive compatible for an acolyte to produce up to $y + \gamma$, that is up to $\varphi = 1$. Therefore, a norm (φ, y, P) with $0 \leq \varphi \leq 1$ is incentive compatible for both types of producers if and only if $y \leq \phi(1-\pi)P$.

If $\phi = 0$ the only feasible $y = 0$ and $U = 0$ for any $P\varphi$, as in the statement. Now assume $\phi > 0$. Since P should be minimized we get $P = y/[\phi(1-\pi)]$. Incentive compatibility for monitoring requires $\psi P \leq \gamma$, which inserting the value of P from above reads

$$y \leq (1-\pi)\phi\gamma/\psi. \tag{3.2}$$

Now the monitoring cost of output $y + \phi\varphi\gamma$ is $(1+\psi)\phi\pi P = My$, so the objective function is

$$U = Vf(y + \phi\varphi\gamma) - (y + \phi\varphi\gamma) - My. \tag{3.3}$$

This has to be maximized with respect to y, φ subject to the constraints $y \leq (1-\pi)\phi\gamma/\psi$ and $0 \leq \varphi \leq 1$.

Since the objective function is concave and the constraint set convex we see immediately that the maximized objective is concave in ϕ. It is increasing in ϕ: because utility depends only on $x = y + \phi\varphi\gamma$ and y and the feasibility restrictions $x \leq y + \phi\gamma$ and $y \leq (1-\pi)\phi\gamma/\psi$ are both relaxed as ϕ is increased.

From the objective function we see that φ is a dominant technology over y: that is, increasing output by increasing y has an associated monitoring cost of $(1+\psi)\pi y/(1-\pi)$ and φ does not. In particular if at the optimum $\varphi < 1$ then $y = 0$ otherwise output $y + \phi\varphi\gamma$ could be held fixed and utility increased by lowering y and increasing φ. Why choose $\varphi < 1$? Because there is also a resource cost of producing output when the designer faces the *first best problem* of maximizing $Vf(x) - x$. If $Vf'(\phi\gamma) \leq 1$ the solution to this problem is feasible and obtained by taking $y = 0$ and from Lemma 1 with $a = 0, b = \phi\gamma, c = \phi\gamma$ choosing φ as stated in the proposition.

The solution $\varphi = (1/(\phi\gamma))g(1/V, 0, \phi\gamma)$ has the property that $\varphi = 1$ for $Vf'(\phi\gamma) \geq 1$. When that is the case it may be optimal to choose $y > 0$: we should fix $\varphi = 1$ and maximize U in 3.3 with respect to y under the constraint $y \leq (1-\pi)\phi\gamma/\psi$. Applying Lemma 1 - with $a = \phi\gamma, b = 1, c = [1+M)$- the given solution results. From the definition of g if $y > 0$ this solution satisfies

$$Vf'(y + \phi\gamma) - 1 - M \geq 0$$

so $Vf'(y + \phi\varphi\gamma) - 1 > 0$ implying $\varphi = (\phi\gamma)^{-1}\left(g(1/V, y, \phi\gamma) - y\right) = 1 = (\phi\gamma)^{-1}g(1/V, 0, \phi\gamma)$ as it should. □

Observe that total output is $x = y + \phi\varphi\gamma$, and from Theorem 1 $y \leq (1-\pi)\phi\gamma/\psi$. Since $\phi, \varphi \leq 1$ taken together these imply that the greatest possible output is $x \leq (1-\pi)\gamma/\psi + \gamma \equiv \chi$.

3.1. The Secondary

If we assume that $q = 1$ then the secondary does not matter for the first stage problem since *a priori* it has zero probability. Hence from internalization ϕ is predetermined in the secondary and the solution and comparative statics are that of the second stage problem given in Theorem 1. We will later show this is a good approximation when $q < 1$ but not too small.

We focus on the case $Vf'(\phi\gamma) > 1$: if $Vf'(\phi\gamma) \leq 1$ we can attain the first best in the secondary simply by having acolytes produce.

Corollary 1. *If $Vf'(\phi\gamma) > 1$ then total output is increasing in V, ϕ and decreasing in π, ψ. Define $\hat{\phi}$ by*

$$f'\left(\chi\hat{\phi}\right) = \frac{1}{V}\left(1 + (1+\psi)\frac{\pi}{1-\pi}\right).$$

For $\phi < \hat{\phi}$ the quota y,the punishment P and monitoring cost My are increasing in ϕ and for $\phi > \hat{\phi}$ they are decreasing.

Proof. From Theorem 1 we know that if $Vf'(\phi\gamma) > 1$ the solution has $\varphi = 1$ and

$$y = g\left(\frac{1}{V}\left(1 + (1+\psi)\frac{\pi}{1-\pi}\right), \gamma\phi, \frac{(1-\pi)\phi\gamma}{\psi}\right) - \gamma\phi, \quad P = \frac{y}{\phi(1-\pi)}$$

so since total output is $y + \phi\gamma$ the first part follows from Lemmas 1 and 1. From the two cited lemmas it also follows that y is decreasing in ϕ in the interior but increasing at the upper bound. The condition given in the result is the transition between the interior and upper bound. (Note that the upper bound binds when ϕ is small.)

The final part follows from the fact that from Theorem 1 P and from equation 3.1 My are increasing in y. □

4. Optimal Internalization

We first solve the first stage problem in the state $s = 1$ under the assumption that the secondary state does not occur, that is, $q = 1$. We refer to this as the *primary problem*. Again we omit state subscripts as we are dealing entirely with one state.

Substituting P from Theorem 1 objective function is

$$W = U - H\phi = Vf(y + \phi\varphi\gamma) - (y + \phi\varphi\gamma) - My - H\phi$$

with the constraints $y \leq (1 - \pi)\phi\gamma/\psi$ and $\varphi \in [0, 1]$.

Lemma 2. *The optimal social mechanism has* $\varphi = 1$.

Since indoctrination is costly we should produce no more acolytes than necessary: if $\varphi < 1$ then we could achieve the same goal with fewer acolytes.

Proof. Hold output $y + \phi\varphi\gamma = x$ fixed. We see that if $\varphi < 1$ and $\phi > 0$ we can keep output fixed by increasing φ and decreasing ϕ. Since ϕ has marginal cost H and φ has none this strictly increases the objective function. On the other hand if $\phi = 0$ then φ does not matter, so we may as well take it equal to 1. □

Theorem 2. *If* $H < \gamma M$ *then* ϕ *maximizes* $Vf(\phi\gamma) - (1 + H/\gamma)\gamma\phi$ *so is given by* $\phi = (1/\gamma)g\left((1/V)(\gamma + H)/\gamma, 0, \gamma\right)$ *and*

$$y = g\left(\frac{1 + M}{V}, \gamma, \chi - \gamma\right) - \gamma.$$

If $H > \gamma M$ *then* $y = (1 - \pi)\phi\gamma/\psi$ *and* ϕ *maximizes* $Vf((\chi\phi) - \chi\phi - (1 + \psi)\pi\phi\gamma/\psi - H\phi$ *so is given by*

$$\phi = \frac{1}{\chi} g\left(\frac{1}{V}\frac{\chi + (1 + \psi)\pi\gamma/\psi + H}{\chi}, 0, \chi\right).$$

The theorem has two cases depending on the marginal cost of internalization H. If this is low then the situation is much as in the second stage problem: as V increases first produce cheap acolytes simply to produce, pause, then start producing additional acolytes to provide monitoring. If the marginal cost of internalization is high then we always invest in acolytes both to produce and to monitor.

Proof. The partial derivatives of the objective function are

$$\partial W/\partial y = Vf'(y+\phi\gamma) - 1 - (1+\psi)\frac{\pi}{1-\pi}$$

$$\partial W/\partial \phi = \gamma\left(Vf'(y+\phi\gamma) - 1\right) - H = \gamma\left(\partial W/\partial y + (1+\psi)\frac{\pi}{1-\pi}\right) - H.$$

If follows directly that if $H < \gamma(1+\psi)\pi/(1-\pi)$ then $\partial W/\partial\phi \leq 0$ implies $\partial W/\partial y < 0$ hence at the optimum, if $\phi < 1$ so that $\partial W/\partial\phi \leq 0$, we have $\partial W/\partial y < 0$ so $y = 0$. When $y = 0$ Lemma 1 gives the expression for ϕ in the statement. If $\phi = 1$ (with $\varphi = 1$ as well) we can write the objective function as $W = Vf(y+\gamma) - y - \gamma - y(1+\psi)\pi/(1-\pi) - H$. Applying Lemma 1 gives the expression for y. Since $H < \gamma(1+\psi)\pi/(1-\pi)$ if $\phi < 1$ the expression gives $y = 0$ so is valid in both cases. On the other hand $\phi = 1$ iff $\gamma Vf'(y+\phi\gamma) \geq \gamma + H$ in which case $\gamma Vf'(\phi\gamma) \geq \gamma + H$ so $(1/\gamma)g\left((\gamma+H)/(V\gamma), 0, \gamma\right) = 1 = (1/\gamma)\left[g\left((\gamma+H)/(V\gamma), y, \gamma\right) - y\right]$.

It similarly follows that if $H > \gamma(1+\psi)\pi/(1-\pi)$ then the constraint $y \leq (1-\pi)\phi\gamma/\psi$ binds. Indeed in this case if $\partial W/\partial y \leq 0$ then $\partial W/\partial\phi < 0$, so either $\partial W/\partial\phi < 0$ so $\phi = 0$ or $\partial W/\partial y > 0$ hence the constraint again binds. This gives the objective function

$$W = Vf((1-\pi)\phi\gamma/\psi + \phi\gamma) - ((1-\pi)\phi\gamma/\psi + \phi\gamma) - (1+\psi)\pi\phi\gamma/\psi - H\phi$$

which making use of $\chi = (1-\pi)\gamma/\psi + \gamma$ is as given above, so the final result follows as well from Lemma 1. □

4.1. Continuity

We now consider $q < 1$. The essential point is that for q close to 1 the solution is approximately that of solving the primary problem with $q = 1$.

Corollary 2. *Under the generic assumption $H \neq \gamma(1+\psi_1)\pi_1/(1-\pi_1)$ the primary problem has a unique solution $\hat{\phi}_1$. Let $\hat{\phi}(q)$ be optimal for q. If $q \to 1$ then $\hat{\phi}(q) \to \hat{\phi}_1$.*

Proof. Uniqueness is given in Theorem 2. The full objective function $W = qU_1 + (1-q)U_2 - H\phi$ is continuous and the feasible domain given by $0 \leq \phi \leq 1$, $0 \leq \varphi \leq 1$ and the incentive constraints $y_s \leq (1-\pi_s)\phi\gamma/\psi_s$ from equation 3.2 is compact. By a standard argument this implies that the argmax correspondence is upper hemi-continuous. The result then follows directly from the uniqueness of the solution for $q = 1$. □

From Theorem 1 it is also the case that the solutions of the second stage problems are continuous in ϕ. Hence for q near 1 to a good approximation the level of internalization ϕ is given by the solution of the primary problem and this is predetermined from the perspective of the secondary problem. We shall adopt this point of view henceforth, taking $q = 1$.

We should emphasize the following. Our benchmark producer/monitor/recipient model has been chosen for illustrative purposes. There are many other mechanism design problems that might in practice constitute either the primary or secondary. For example, production might involve joint

effort by several group members, there might be several monitors, output might have both public and private dimensions, and so forth. The key point is that as long as mild regularity conditions are satisfied - the full objective function is continuous in the level of internalization ϕ, the incentive constraints are continuous and the solution to the primary problem with $q = 1$ is unique.

4.2. Comparative Statics of the Primary

To a good approximation, then, the comparative statics of ϕ and of the primary problem are independent of the secondary problem.

Corollary 3. *Internalization ϕ, the production quota y and total output $x = y + \phi\varphi\gamma$, punishment P and monitoring cost (expected cost of punishment) are increasing in V. Total output is increasing in γ and decreasing in π.*

Proof. Internalization and the production quota follow directly from Theorem 2, with total output the immediate consequence.

Punishment is given by $P = y/\left[\phi(1-\pi)\right]$ from Theorem 1. By Theorem 2 if $H < \gamma(1+\psi)\pi/(1-\pi)$ then either $y = 0$ so $P = 0$ or $\phi = 1$ in which case y is increasing in V so P is as well. If $H > \gamma(1+\psi)\pi/(1-\pi)$ then $y = (1-\pi)\phi\gamma/\psi$ so P is independent of V.

Monitoring cost is given by $(1+\psi)\pi y/(1-\pi)$ from equation 3.1, hence the result. The third part by observing from Theorem 2 if $H < \gamma(1+\psi)\pi/(1-\pi)$ then total output is

$$x = g\left(\frac{1}{V}(\gamma + H)/\gamma, 0, \gamma\right) + g\left(\frac{1}{V}\left(1 + (1+\psi)\pi/(1-\pi)\right), \gamma, \chi - \gamma\right) - \gamma$$

and if $H > \gamma(1+\psi)\pi/(1-\pi)$ then total output is

$$x = g\left(\frac{1}{V}\frac{\chi + (1+\psi)\pi\gamma/\psi + H}{\chi}, 0, \chi\right).$$

□

The effect of γ, π, ψ are more complicated because they depend on the elasticity of demand. We denote by g_k the derivative of $g(\mu, x, X)$ with respect to variable $k = \mu, x, X$ and $G(\mu, x, X) \equiv g_\mu(\mu, x, X)/g(\mu, x, X)$. Also let $\mathcal{H} = 1$ if $H > \gamma(1+\psi)\pi/(1-\pi)$ and zero otherwise. Define the demand elasticity

$$\eta = -\frac{1}{V}G\left(\frac{1}{V}\left(1 + \frac{H + \mathcal{H}(1+\psi)\pi\gamma/\psi}{\gamma + \mathcal{H}(1-\pi)\gamma/\psi}\right), 0, \gamma + \mathcal{H}\frac{(1-\pi)\gamma}{\psi}\right).$$

Corollary 4. *There are cutoffs $\eta_\gamma, \eta_\pi, \eta_\psi$ so that for $\eta < \eta_\gamma$ internalization ϕ is decreasing in γ and conversely, while for $\eta < \eta_\pi$ [resp. $\eta < \eta_\psi$] internalization ϕ is increasing in π [resp. ψ] and conversely.*

As the proof is a computation it is proven as Corollary 5 in the Appendix.

4.3. Lessons Learned

The interesting and striking point is the non-monotonicity of punishment and monitoring cost of the secondary in ϕ (Corollary 1). Consider raising V_1 (the value in the primary). This will increase internalization, initially raising punishment and monitoring cost in both the primary and the secondary. However, as V_1 is increased, while punishment and monitoring in the primary continue to increase they decrease in the secondary. If there is a high degree of internalization then it does not make sense to punish everyone to squeeze extra output out of a few opportunists. By contrast in the primary we would never choose the level of internalization so high.

There are several other take-aways from this analysis. First, internalization is essential for monitors: in this model no monitoring can take place without internalization because monitoring is costly and monitors cannot be monitored.[8] It is a ubiquitous problem in mechanism design that getting people to tell the truth about others is problematic. If monitors have incentive to lie, for example, because punishment either is costly or beneficial to them, and they can be identified, then it is possible to make them indifferent by punishing them based on their reports. However, this provides weak incentives for truth-telling and if monitoring itself is costly, there is no incentive to bear that cost. Even a small incentive to tell an undetectable lie can lead to enormous losses - a small amount of internalization by making it strictly optimal for acolytes to tell the truth can have a big impact.[9]

The second take-away is that in this simple model there is a single variable "internalization" ϕ that links problems across states. This has also been called "publicness" and "pro-social." It plays a key role in solving the second stage problem as Theorem 1 shows. One particular implication is that if we can measure ϕ as we do below using laboratory data then it tells us something about the solution of the mechanism design problem outside the laboratory.

The role of internalization also differentiates societies. That is, societies facing different primary problems will choose different levels of internalization and this means that they will choose different solutions to secondary problems: we will give a practical example of this in an application below.

Finally: the difference in solutions between the primary and secondary problem means that we cannot reach simple and direct conclusions based on observing the secondary. For example: if we observe little punishment in the secondary this does not imply that there is little punishment in the primary. Suppose that V_1 is very large. Then from Theorem 2 internalization and punishment in the primary will both be at their upper limits: $\phi = 1$ and $P_1 = \gamma/\psi$. In the secondary, however, as indicated if V_1 is large then in fact we will see little punishment due to high internalization as this is determined in the primary and will be larger than $\hat{\phi}$. Indeed if V_2 is small enough then from Theorem 1 $P_2 = 0$.

[8] Or it is prohibitively expensive to do so: see Levine and Modica (2016) for a model where monitors can be monitored.

[9] This is not a paper about monitoring technology: in addition to monitoring monitors it may be that there are several monitors whose reports can be compared. For a deeper analysis of monitoring monitors see Rahman (2012). We chose this simple technology to make the point that internalization can be essential.

4.4. Laboratory Experiments

We already discussed in Section 4.1 allowing more complicated or different primary and secondary mechanism design problems than the illustrative benchmark model. We observed there that under mild regularity conditions, to a good approximation the level of internalization is determined in the primary but predetermined in the secondary. The classical case of a secondary state is a laboratory experiment. We can be reasonably confident that internalization is determined without reference to the possibility that group members may find themselves under study by social scientists. Hence Corollary 2 as discussed in Section 4.1 says that to a good approximation ϕ is predetermined and subjects will solve the mechanism design problem posed in the laboratory taking this as given. Is there any evidence that they do so? The next two sections are a quantitative examination respectively of an experimental public good contribution game with punishment and of experimental dictator/ultimatum games.

In order to move to applications we need to look more closely at the monitoring technology, in particular how a wrong signal about a member's behavior may arise in the laboratory. In each case there is a social norm in the form of an output quota y and the signal z^i on member i's behavior is about whether or not output $x^i \geq y$. There are two possible sources of noise: it may be that x^i is imperfectly observed, which we think is the most common interpretation. However, it could equally well that the social norm y is imperfectly observed. In the laboratory as a rule x^i is perfectly observed, so in our analysis of laboratory experiments we shall take the latter interpretation: π corresponds to uncertainty over the social norm. An ultimatum bargaining experiment is, for example, an unusual event, and two different members of a group may well have different interpretations of how the social norm applies. In the public goods experiments we study, in three different sessions average output ranges from 9.8 to 14.3 which indicates to us that there is substantial uncertainty about what the social norm is.

We should also emphasize that the laboratory is an especially difficult environment for solving mechanism design problems. Agreement over a social norm must be reached without the possibility of discussion and based on limited observation of the behavior of other participants in a small number of matches. We do not think that people instantaneously solve mechanism design problems any more than they instantaneously solve optimization problems. Hence, as is common in the study of equilibrium, we focus on later rounds after learning has taken place. In other words, there is no more reason to presume that participants have successfully solved a mechanism design problem the first time than to imagine that they reach equilibrium the first time they play.[10]

Overview of findings

Before jumping into the details here is an overview of our findings. We study three classes of games in which the subjects are Western college students: a public goods game, the dictator

[10]The literature on level-k beliefs, for example Stahl and Wilson (1995), show clearly that equilibrium play is not a good description of the first round in the laboratory, while repeated strangers treatments often lead to equilibrium even in environments where finding equilibrium is computationally demanding, see, for example, Levine and Palfrey (2007).

game, and ultimatum bargaining. First, it appears that the fraction of acolytes is about 50% and that it takes around ten rounds of play to "solve" the mechanism design problem posed in the laboratory in the sense that utility is higher with an effective use of punishments. The theory works well quantitatively for both the public goods problem and for dictator games. For ultimatum bargaining games the results are mixed. If the only source of the demand for fairness is risk aversion then the theory fairs poorly, but if there is substantial demand for fairness for other reasons the theory fairs well.

5. Public Goods and Punishment in the Laboratory

The classical experiment on the use of punishments to induce contributions to a public good is that of Fehr and Gachter (2000). They study a public goods contribution game with four players. They examine treatments both with and without the possibility of punishment. Participants choose contribution levels $0 \leq x^i \leq 20$ and receive utility $u^i = v_0 - cx^i + v\sum_{j\neq i} x^j$ where $v_0 = 20$, $v = 0.4$, and $c = 0.6$.

We analyze their results for the last of ten rounds in the stranger treatment. As indicated, we examine the final round to allow participants the chance to "learn their way" to a solution. Although Fehr and Gachter (2000) also study a partners treatment this is a repeated game played only once - we know from the work of Dal Bo (2005) that we need repeated repeated treatments in order to observe equilibrium play. Hence we focus on the stranger treatment. We use data averaged across all three sessions.

The average contribution in the no-punishment condition is $x = 1.9$. In the punishment treatment we will shortly describe contributions were much higher, at $x = 12.3$. Can our theory of internalized norms possibly account for such large contributions when there is punishment? Surprisingly the answer is yes: with the costs and consequences of punishment the acolytes can be leveraged to greatly enhance contributions.

5.1. The Punishment Game

We must describe how the punishment treatment works. After contributions are observed participant i can purchase punishment points p_i^j against j. The cost of these points is equal to the number of points up to 2 points, then becomes convex.[11] As we explain later our theory does not suggest purchases greater than 2.43 so we treat the cost of punishment points as linear. Each punishment point against a participant reduce their payoff by 10%: specifically utility at the end of the punishment round is $v^i = (1 - (1/10) \cdot \min\{10, \sum_{j\neq i} p_j^i\})u^i - \sum_{j\neq i} p_i^j$, where the min avoids pushing payoff below zero.

5.2. The Mechanism Design Problem

As indicated, we interpret noise in the signal z^i as due to uncertainty about the social norm which is a quota y for contributions. For simplicity we assume first that all four participants observe

[11] The cost of 3 points is 4, for example.

the same signal z^i. We assume second that if there is a bad signal for any match participant all the acolytes choose a common number of punishment points which we denote by p.

The second row of the table below lists for a particular participant i who has a bad signal the probability that one of the other three has a bad signal.

others with bad signals	0	1	2	3
probability	$(1-\pi)^3$	$3(1-\pi)^2\pi$	$3(1-\pi)\pi^2$	π^3
number punishing	3	2	4/3	1

The final row of the table indicates how many opponents (conditional on being an acolyte) are willing punish i. If i has the only bad signal all three opponents will punish her if they are acolytes (total 3). If there is one other bad signal then the two without bad signal each give half a punishment to the two with bad signals, and the one with a bad signal gives a full punishment to i (she does not punish herself), so total in this case is $1/2+1/2+1$. If there are two other bad signals then the one without a bad signal gives 1/3rd punishment and the two with bad signal each give half a punishment to the other two with bad signals, with total $1/3+2\cdot 1/2 = 4/3$. Finally, if there are three other bad signals then each gives 1/3rd punishment. To compute the expected value, observe that if the numbers in the final row were $3, 2, 1, 0$ the expectation would be $3(1-\pi)$. Hence the actual expectation is $Q = 3(1-\pi) + (1/3)3(1-\pi)\pi^2 + \pi^3 = 3(1-\pi) + \pi^2$. Each individual has probability ϕ of being an acolyte so the expected punishment conditional on having a bad signal is ϕQp.

For an opportunist then the utility from abiding by the social norm of y with average output x is $(1-\pi Qp\phi/10)(v_0 - cy + 3vx)$ and from contributing zero is $(1-Qp\phi/10)(v_0+3vx)$, where notice that the free rider has no punishment cost because she does not punish. Hence the incentive constraint is $(1-\pi Qp\phi/10)(v_0 - cy + 3vx) \geq (1-Qp\phi/10)(v_0+3vx)$.

Next we need to determine how much extra $Y-y$ an acolyte is willing to produce. The fact that there is an expected cost of punishing in the punishment round limits what acolytes will be able to contribute in the first. Specifically, the expected cost of punishing in the punishment round is $(1-(1-\pi)^3)p$. Hence the extra cost that can be carried in the first period is $\gamma - (1-(1-\pi)^3)p$. This gives $Y - y = \left(\gamma - (1-(1-\pi)^3)p\right)/c$.

The mechanism design problem can now be stated: it is to maximize over y, Y, x, p the objective

$$W = (1-(1/10)\phi Qp\pi)(v_0 - cx + 3vx) - (1-(1-\pi)^3)\phi p$$

subject to feasibility $x = y+\phi(Y-y)$, incentive compatibility for the opportunists $(1-\pi Qp\phi/10)(v_0 - cy + 3vx) \geq (1-Qp\phi/10)(v_0+3vx)$ and the two incentive compatibility constraints for the acolytes: $p \leq \gamma$ and $Y - y = \left(\gamma - (1-(1-\pi)^3)p\right)/c$.

Since the objective is linear and increasing in y and the opportunistic incentive compatibility constraint is linear, it follows that the opportunists constraint must hold with equality. Solving it

for x we get

$$x = \frac{(1 - \pi Qp\phi/10)c\phi(Y - y) + (1 - \pi)Qp\phi/10)v_0}{(1 - \pi Qp\phi/10)c - (1 - \pi)(Qp\phi/10)3v}.$$

5.3. Calibration

Quoting Fehr and Gachter (2000), the key fact is that "in the no-punishment condition of the stranger-treatment average contributions converge close to full free-riding over time." In particular the average contribution was $x = 1.9$.[12] Moreover "we call those subjects 'free-riders' who chose ...[to contribute 0]... in more than five periods of the no-punishment... [They constitute] 53 percent in the Stranger-treatment." Adopting this definition it appears that in this population $\phi = 0.47$.

Knowing that in the stranger treatment $x = 1.9$ enables us to compute γ. The value of average contribution is $cx = 0.6 \times 1.9 = 1.14$ and $cx = (1 - \phi) \times 0 + \phi\gamma$, that is acolytes contribute the most they are willing. Solving this yields $\gamma = 2.43$.

We can now solve the mechanism design problem numerically for each value of π. This represents the level of uncertainty over the social norm. There are three targets we can try to match: the first two are output $x = 12.3$, and welfare. Welfare is reported as 10% higher than the token utility of 21.1 received in the treatment without punishment,[13] which is to say 23.3 tokens. The third possible target is the *failure rate*, denoted by R. This is defined by $W = (1 - R)(v_0 - cx + 3vx) - 10R$, where the factor of 10 is there because each punishment point which costs one token buys only a 10% increase in failure. Using $x = 12.3$ and $W = 23.3$ gives $R = 0.11$.

In fact it turns out to be impossible to target the failure rate because no optimal mechanism has a failure rate as high as that in the data, $R = 0.11$. For this reason we compute instead R_d which would be the failure rate if all acolytes choose the maximum punishment $p = \gamma$ on a bad signal.

As utility measured in tokens is not especially interesting, we normalize utility so that it is zero when no public good is produced and one at the maximum possible utility of 32 when everyone donates 20 tokens and there is no punishment. That is, if U is utility in tokens we report welfare $(U - 20)/12$. In these units welfare from no punishment of 21.1 tokens is 0.10 and from punishment of 23.3 tokens is 0.28 respectively. In other words, the mechanism observed in the data is successful in the sense that it yields 0.17 more utility than that without punishment.

Below we report the values of π that match each of the targets of output x, welfare, and R (with the last row discussed below).

γ	match	π	p/γ	x	y	welfare	R	R_d
2.43	x	0.28	1.0	12.3	11.6	0.39	0.07	0.07
2.43	welfare	0.32	1.0	10.2	9.6	0.28	0.08	0.08
2.43	R_d	0.65	0.0	1.9	0.0	0.10	0.00	0.11
3.07	all	0.38	1.0	12.2	11.8	0.28	0.11	0.11

[12] Average contributions for both no-punishment and punishment are taken from their Table 3.
[13] Result 8.

The first row where we match on x shows that indeed there are enough acolytes to enforce a quota of 11.6 resulting in an output level of 12.3. However, the value of π for which this is the optimal mechanism is 0.28 for which the optimal mechanism would have a failure rate of only 0.07 against the empirical 0.11, resulting in utility 0.39, higher than observed due to the lesser amount of punishment. We can see this in the other two rows: if we match on welfare then $\pi = 0.32$ which would result in $y = 9.6$ so that it is not incentive compatible for acolytes to produce 11.6; while if we match on failure rate $\pi = 0.65$, which is so much uncertainty about the social norm, the best thing to do is to revert to no punishment.

To summarize, it appears that there is over-punishment. To resolve this we hypothesize that the measured γ is too low. The final row of the table shows how to choose γ to match all three targets: here $\pi = 0.38$. The hypothesized value of $\gamma = 3.07$ corresponds to a contribution of 2.4 tokens in the no punishment game against the empirical estimate of 1.9. As the standard error on the estimate of 1.9 is 4.1 this corresponds to hypothesizing that the actual donation rate without punishment is about one quarter of a standard deviation higher than observed in the data: this seems plausible.

6. Fairness and the Equal Split

In this section and the next we examine two games that have been heavily studied in the experimental laboratory: dictator and ultimatum. In both of these two-player games the first mover receives an endowment X and from it offers an amount x^1 to the second mover. In dictator the decision of the first mover is final; in ultimatum the second mover has the option to reject the offer in which case both get zero. We denote by c^i the amount received by each player: $c^1 = X - x^1, c^2 = x^1$ in dictator or if there is agreement in ultimatum, or zero if the offer is rejected in ultimatum. For both games offers greater than 0 are common, and a $50 - 50$ split is often observed.

What mechanism design problem would result in a 50-50 sharing rule in a dictator or ultimatum game? The answer is that there are several, and indeed we know from the work of Townsend (1994) and Prescott and Townsend (1984) that mechanism design with *ex ante* uncertainty about types creates a strong tendency towards equal division. Here we highlight three forces working towards equal sharing.

Risk and Insurance. Laboratory participants are known to be risk averse over laboratory stakes. If they are *ex ante* identical then it is socially optimal to share unanticipated gains. In particular, in a dictator game if both participants have an identical risk averse utility function $u(c^i)$ then welfare is $u(X - x^i) + u(x^i)$ which is maximized when $x^i = X/2$.

Incentives and Commitment. We know that giving is sensitive to effort (Kahneman, Knetsch and Thaler (1986)). Indeed, even in dictator effort is involved for both parties: the effort in showing up to the laboratory and remaining even when it is discovered that the participant has been

assigned to the role of recipient. When there is joint production and effort is complementary, if all the output accrues to one partner there is a commitment problem: *ex ante* there should be commitment to sharing to provide the partner with incentives to provide effort, but *ex post* the partner who receives the output would prefer to keep it. Social mechanisms can provide the missing commitment. As a simple illustration, suppose there is a joint production function in which output is $y = V(x^1 x^2)^\alpha$ with $\alpha < 1/2$. If h_i is the output share of individual i then individual expected utility is $h_i y - x^i$. Fixing the output shares the optimal individual output is shown in the appendix to be $x^i = (\alpha V)^{1/(1-2\alpha)} (1-h_i)^{\alpha/(1-2\alpha)} h_i^{(1-\alpha)/(1-2\alpha)}$. The social objective function is then

$$V(x^1 x^2)^\alpha - x^1 - x^2 = ((1-h_1)h_1)^{\alpha/(1-2\alpha)} (\alpha V)^{2\alpha/(1-2\alpha)} V(1-\alpha).$$

This has a maximum at $h_1 = 1/2$: that is the optimal incentives are provided by an equal sharing rule.

Prevent Conflict From Competitiveness. One of our first experiences with the social norm of sharing is as children when we are asked to share toys rather than compete and fight over them. Competitiveness from a utilitarian point of view is a utility function (in the extreme case) of $c^i - c^{-i}$: that is an individual is *competitive* if they care about how much better they do than others. Outside the laboratory this may be for a variety of reasons - for example an individual may benefit by weakening an opponent as well as strengthening themselves, and reputation may be enhanced by outperforming others. Competition leading to lower prices, lower costs and innovation has social value. Competition leading to transfer payments does not. There is evidence of competitiveness in the laboratory: for example in Palfrey and Prisbrey (1997) about 15% of participants fail to contribute to a public good when it costs them nothing to do so. Note that this is inefficient from a social point of view, although it is strictly desirable for a competitive individual. It makes sense as well that participants might well "compare notes" after the experiment to see who did the best.

To see how competitiveness matters in ultimatum, observe that a competitive individual will reject all offers less than 50% and accept all offers greater than 50%. As when there are multiple equilibria our assumption is that the best one is chosen, we assume that all offers of exactly 50% are accepted as well. If there are no competitive individuals in the population all offer nothing and get it - there is no inefficient rejection of offers. If there are only competitive individuals in the population all offer 50% and get it (again we assume that the efficient action is taken in the face of indifference) - and again there is no inefficient rejection of offers.

When there is a mixed population, some competitive and some not, there is a problem. In this case all competitive individuals will offer nothing because that offer will be accepted by the non-competitive individuals (competitive recipients would only accept offers exceeding 50% but if you are competitive those splits give you zero utility at best). If less than 50% of the population is competitive then non-competitive individuals will also offer nothing since a better than even chance of X is better than a certain chance of $X/2$. Regardless: some offers will be rejected by competitive individuals and this is inefficient from a social point of view. These rejections can be viewed as a

kind of inefficient conflict.

There are two sources of the social problem created by competitiveness. There are the on-path rejections of offers by competitive individuals and there are the offers of less than 50% by both types of individuals. The mechanism designer can do nothing about the first problem directly- there is no way to impose additional punishment on individuals who reject offers. But the mechanism designer can do something about the second problem. If non-competitive individuals can be convinced to enforce the social norm of an even split by having second movers reject worse offers then demands will be reduced and offers not rejected, restoring efficiency. This has the interesting feature that the solution to inefficient (on-path) rejections is to increase the (off-path) rejections. We should note that it was established in Levine (1986) that second movers rejection rates exceed that consistent with the amount of competitiveness in the first round - a small bit of evidence that participants are "getting it right" in the laboratory.

6.1. Demand for Fairness

The simplest and cleanest model is that of risk. To do a quantitative analysis we need a utility function. Here we take the calibration from Fudenberg and Levine (2011): if c denotes laboratory earnings they suggest that a utility function of the form $1-(1+c/C)^{1-\rho}$ with $C = \$40$ fits the data reasonably well. There is considerable heterogeneity in risk aversion (which we will ignore) and they find that the median coefficient of relative risk aversion ρ is about 1.43. This can be thought of as a measure of the demand for fairness: the greater is ρ the greater the social gain from equalizing income. In dictator, as we shall see, the value of ρ matters little as long as it is positive. By contrast, in ultimatum ρ plays a key role - and $\rho = 1.43$ is not nearly large enough to explain observed behavior through thesocial mechanism theory outlined above- it predicts considerably more selfish behavior than we observe. Since, as we have indicated, there are additional forces creating demand for fairness - incentives and conflict prevention - we do not view this as an important shortcoming. To account for these additional forces we propose to keep the simple clean risk aversion model but for social utility use the CES utility function with $\rho = \rho_r + \rho_f$ where $\rho_r = 1.43$ and ρ_f is a calibrated additional demand for fairness.

6.2. Dictator

Dictator games are relatively easy. There is no possibility of punishment: with the standard $X = \$10$ the theory says that the acolytes should contribute the minimum of \$5 and γ. In Engel (2011)'s meta-study of dictator games "dictators on average give 28.35% of the pie" but for students (the subject population for the public goods and ultimatum experiments we discuss) the meta-regression gives a value of 39.8%. This is remarkably close to 47% of acolytes each giving 50%, which is to say that if $\gamma \geq \$5.00$ the theory predicts what we see in dictator games.

It is worth pointing out that the theory contends equally well with experiments in which there is an additional option to "take" \$5.00 from the second mover. In this case the free riders should indeed take, while the acolytes offers would be $-5.00+\gamma$ or \$5.00 if this is smaller. Indeed, we can use the results of "take" experiments to get an estimate of γ. In List (2007) adding the "take" option resulted

in a drop from giving away \$2.48 to taking of \$1.33 that is a drop of \$2.48+\$1.33 = \$3.81. If we let λ represent the drop in acolytes offers we have $3.81 = \phi\lambda+(1-\phi)5$ so that $\lambda = (\$3.81-\$2.65)/.47 = \$2.47$ which says that acolytes donations drop from \$5 to $\$(5-2.47) = \$2.53$; since the donation is $2.53 = \min\{\gamma-5, 5\}$ this in turn implies a value of $\gamma = \$5.00 + \$2.53 = \$7.53$.

6.3. Ultimatum

We review the mechanism design problem, assuming that the endowment X is 10. Individual utility is given by $u(c) = 1-(1+c/C)^{1-\rho_r}$ and social utility by $w(c) = 1-(1+c/C)^{1-\rho_r-\rho_f}$ where $C = 40$ and $\rho_r = 1.43$. In our reporting we will continue normalize social utility to be measured as a fraction of the maximum, that is, we will multiply by $1/w(5)$ (5 being the equal split of $X = 10$). We denote by q the probability an acolyte rejects an offer on a bad signal, and continue to denote by π the error rate in the signal process. As in our dictator data we discovered γ considerably above \$5 we assume that acolytes are willing to reject any unfavorable offer and are willing to offer \$5 (which is Y in this case) which for efficiency reasons they should do. Hence the objective function is

$$(1-q\phi\pi)\left[\phi w(5) + (1-\phi)(1/2)(w(y)+w(10-y))\right]$$

and should be maximized with respect to q, y subject to the constraints that $q \leq 1$ and that for opportunists the utility of conforming to the social norm and offering y is better than deviating and offering zero: $(1-q\phi\pi)u(10-y) \geq (1-q\phi)u(10)$. Since this must hold with equality at the optimum we can compute

$$q = \frac{u(10)-u(10-y)}{\phi u(10)-\phi\pi u(10-y)}.$$

We will now engage in a calibration exercise based on data from Roth et al (1991) used in an earlier behavioral calibration exercise by Levine (1986). We use the USA $X = \$10$ tenth round data. This is reproduced in the data Appendix.

In the public goods game we had $\phi = 0.47$. Here there is a large jump in the number of offers from \$4.25 to \$4.00 and the fraction of offers \$4.25 or less is 48%. This is consistent with the idea that generally acolytes make offers close to \$5.00, so we shall continue to take $\phi = .47$. From the data the mean offer larger or equal to \$4.25 is \$4.74, quite close to the calibrated value of \$5.00. In the data the overall mean offer is equal to $x = \$4.07$.

Our next step is to calibrate ρ_f. The smallest value of ρ_f for which realized utility is greater than the utility from no punishment is $\rho_f = 8.27$, considerably larger than the coefficient of relative risk aversion. In our calibration we will take a slightly higher value $\rho_f = 8.57$ (which results in $\rho_r + \rho_f = 10$): the results are not terribly sensitive to value of ρ_f provided it is greater than 8.27.

As we did in the public goods experiment we can now compute the optimal mechanism as a function of π. We can target output $x = \$4.07$, welfare, or the rejection rate $R = \phi q\pi$. As in the public good experiment it turns out to be impossible to target the rejection rate because no optimal mechanism has a rejection rate as high as that in the data, $R = 0.16$. For this reason we target instead $R_d = \phi\pi$ which would be the rejection rate if all acolytes reject with probability one

on a bad signal. We report the results for the three targets below.

match	π	q	x	y	welfare	R	R_d
x	0.21	0.72	4.07	3.25	0.91	0.07	0.10
welfare	0.55	0.22	2.69	0.65	0.83	0.05	0.23
R_d	0.34	0.53	3.43	2.05	0.86	0.09	0.16

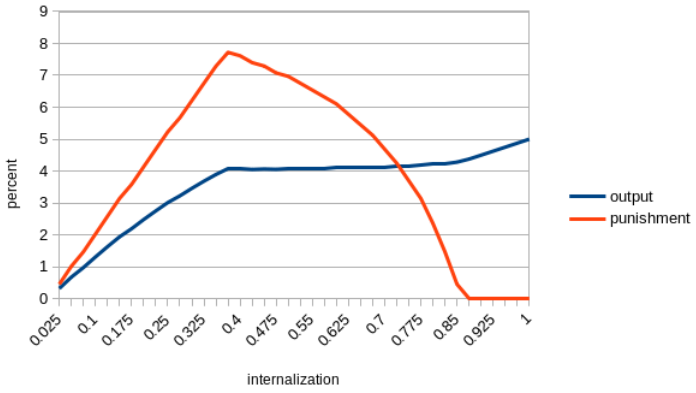
As in the public good experiment if we target output we get a rejection rate 0.07 considerably lower than in the data. To explain the relatively low welfare in the data requires a large value $\pi = 0.55$. Examining the R_d target is instructive. If $\pi = 0.34$ then the optimal mechanism calls for a quota of 2.05 supported by $q = 0.53$. In the data a higher quota of 3.25 is enforced by a higher value of $q = 0.72$. However: the loss from this suboptimal mechanism is low - the optimal mechanism yields a gain of only $0.03 \approx 3.5\%$ over the suboptimal mechanism used. Moreover, over time confusion over the social norm as measured by π should be declining. It might be that with continued play π might fall to 0.21 in which case acolytes would punish less on the equilibrium path, the output target 4.07 would be met, and welfare would increase substantially by 0.08 (from 0.83 to 0.91). The key point is there is reason to believe that this might be the case. Although few ultimatum experiments have been conducted with more than 10 rounds of play, Duffy and Feltovich (1999) did so and they report rejection rates. They find a substantial persistent drop in the rejection rate beginning between round 10 and 20. We refer here particularly to their Figure 3 and the rejection rate for \$3.00 demands (which is roughly the social norm). This drops from about 35% in rounds $6-10$ to about 15% in rounds $11-15$ and remains at roughly that level for the remaining 25 rounds. They do not report an overall rejection rate but it appears after the 10th round to be well less than 10% as opposed to about 16% in our tenth round data. That is, while the observed mechanism features suboptimal punishment it may approach optimality in few rounds.

6.4. Learning and Welfare

In the public goods experiment above it was only in the final two rounds of ten that the punishment mechanism beat the no punishment mechanism. In ultimatum there is evidence that after ten rounds play had not converged to the optimal mechanism. Is must be kept in mind that the laboratory environment is a demanding one for a group to implement a mechanism because there is no possibility of communications. We know that coordination problems are much more easily resolved when there is cheap talk (see, for example, Cason, Sheremeta and Zhang (2012)) - and we observe as well that in the Fehr and Gachter (2000) public good experiment when the game was played with partners rather than strangers - so it was more evident to the group what the social norm was - the group far more effectively implemented a good social norm - inducing contributions close to the first best of 20. One consequence of this is that while there is evidence that participants are struggling and eventually succeeding in finding an optimal social norm it is almost certainly not worth the effort over ten rounds: a few rounds of gain at the end do not make up for the losses accrued during the learning process.

7. Cross Cultural Ultimatum

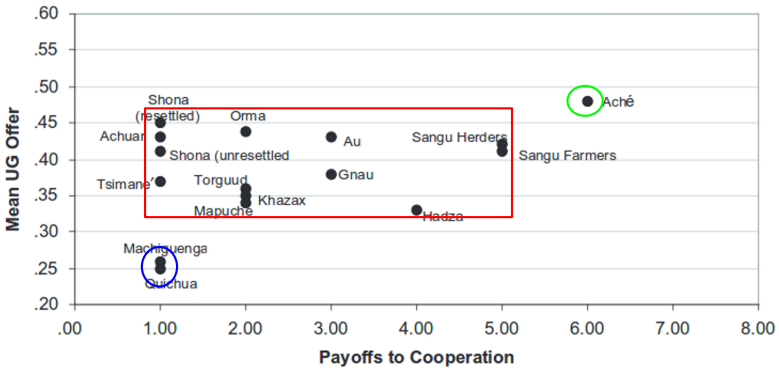
A key element of our theory is that it creates a connection between the primary and secondary through internalization. One take-away is that we cannot simply infer the solution of the primary from the secondary - that is, in general, we cannot make inferences about social norms used in the broader society from those observed in the laboratory. We can, however, go the other way. Specifically, if we observe the level of internalization across societies the theory tells us how this should be reflected in the laboratory. Consider varying V_1 the "importance" of the primary. As this increases we know from Theorem 2 that internalization ϕ should increase. Corollary 1 tells us the resulting impact in the laboratory for a simple public goods experiment: we should see greater output x_2, but also punishment and monitoring cost should initially increase then decrease. While we do not have similar analytic results for ultimatum bargaining we can do a numerical calculation. Let us take our preferred calibration with $\pi_2 = 0.21$ and $\rho = 10.0$ and "large enough" γ and solve the mechanism design problem as a function of internalization: this is shown in the graph below with output $x_2 = y_2 + \phi(Y_2 - y_2)$ and punishment $q\phi\pi_2$ (reported in percent).



There are two notable features. First, like in the simple secondary problem of Corollary 1 punishment is not monotone: it initially increases as more acolytes are available to punish, then declines as with many acolytes there is little reason to pay the cost of punishing the few opportunists. Second output initially rises quite rapidly, flattens out, then rises rapidly again. As can be seen, this is driven by the punishment: after punishment peaks and starts declining output does not rise much, rather the increased internalization is used to reduce punishment costs. Once this is exhausted because punishment is no longer used output begins to rise more rapidly.

While data about groups with different values of ϕ is scarce we do have the famous study of Henrich et al (2001). Here we reproduce data concerning an ultimatum game experiment from Figure 5 in Henrich et al (2005):[14]

[14]We omit data from one group, the Lamalera because deception was used.



The horizontal axis "Payoffs to Cooperation" is an ethnographic variable based on the extent to which each society is judged to benefit from cooperation - or to say the same thing - the importance of public goods in each society. It is conceptually the same as the primary value V_1. The vertical axis is the average offer $x_2 = y_2 + \phi(Y_2 - y_2)$ made by the first mover in the ultimatum (secondary) game. Let us assume that the only difference between these different societies is in fact the primary value of V_1. This determines internalization ϕ as an increasing function from Theorem 2. Hence the horizontal axis may also be taken to measure ϕ. There is in fact some anecdotal evidence to support this: according to Henrich et al (2005) in Ache, the group with the highest value of V_1, "Successful hunters often leave their prey outside the camp to be discovered by others, carefully avoiding any hint of boastfulness." This sounds like a high value of ϕ.

Our reading of the data is that it is consistent with the idea that output rises rapidly at first, is rather flat, then rises rapidly again. The Machigeuenga and Quichua in the blue circle have substantially lower mean offers than any other groups: we take this to mean that while having V_1 very similar to the their groups rated at 1.00 they have slightly lower V_1. For the middle groups marked with the red box output is rather flat as a function of V_1 after jumping up very quickly from the two groups in the blue circle. Finally we see that there is another upward jump for the group with the highest value of V_1, the Ache.

What about punishment and monitoring cost? Here we have the data on rejections across societies. According to the theory the very low V_1, ϕ groups (blue circle) and very high V_1, ϕ groups (green circle) should have very low rejection rates: the former because there are few acolytes to carry out punishment and the latter because there are so many acolytes that punishment is not needed. Indeed: for the two very low V_1, ϕ groups the Machigeuenga and Quichua there is only one rejection out of the 21 pairs in the Machigeuenga and none in the Quichua. In the highest $V_1\phi$ group the Ache there were no rejections. By contrast the red box societies with an intermediate number of acolytes should use punishment to support internalization. Indeed for those societies the average rejection rate is 12%. It appears that indeed punishment initially increases with internalization then decreases.

Our interpretation of the data seen through lens of social mechanisms with internalization is rather different than that taken by Henrich et al (2005). Their view is that greater objective

incentive for cooperation outside the laboratory leads to greater fairness inside the laboratory. This does not predict the lack of monotonicity of offers and punishment that our theory predicts and that we observe in the data. While 14 observations of widely differing societies and a handful of ultimatum games played in each society under difficult conditions cannot be too persuasive, the theory of social mechanisms provides a much more detailed, accurate, and sharper account of what to look for in the data.

8. Conclusion

We conclude by indicating how the ideas in this paper fit into the broader literature of experimental and behavioral economics. Writers such as Bowles et al (2003) and Roemer (2015) point to evolutionary reasons why punishment might be "hard-wired." Experimentalists such as Fehr and Gachter (2000)similarly argue that intrinsic preferences for reciprocal altruism "do unto others as they have done unto you" are observed in the laboratory. We do not doubt that small children do not need to be taught to punish the theft of a toy. Never-the-less social norms must - and do - specify punishment levels scaled to the nature of the offense, the benefit of deviating, and the chances of getting caught. Hence our approach of treating the choice of punishment as the solution to a mechanism design problem. In particular in our setting acolytes carry out punishments because they are useful in solving the social problem of public goods provision, not because of an intrinsic desire for revenge.

We examine a particularly simple stark theory of internalization based on warm glow giving and study the trade-off between the use of incentives and internalization. We show that the idea that in the laboratory participants solve mechanism design problems subject to uncertainty but making good use of internalization is consistent with what we see. In particular we find that internalization is important in alleviating the need to provide incentives to monitors.

References

Abreu, D. and Rubinstein, A. (1988), "The structure of Nash equilibrium in repeated games with finite automata," *Econometrica* 1259-1281.

Akerlof, George A., and Rachel E. Kranton (2000) "Economics and identity," *The Quarterly Journal of Economics* 115(3): 715-753.

Andreoni, J. (1990): "Impure altruism and donations to public goods: A theory of warm-glow giving," *Economic Journal* 100: 464-477.

Belloc, M., F. Drago and R. Galbiati (2016): "Earthquakes, religion, and transition to self-government in Italian cities," *The Quarterly Journal of Economics* 131: 1875-1926.

Bénabou, Roland, and Jean Tirole (2006): "Incentives and prosocial behavior," *The American Economic Review* 96(5): 1652-1678.

J.P. Bénassy (1998): "Conformism and multiple sycophantic equilibria", in P. Howitt and A. Leijonhufvud (eds), *Money, Markets and Method*, Edward Elgar.

Bigoni, M., S. Bortolotti, M. Casari., D. Gambetta and F. Pancotto (2016): "Amoral familism, social capital, or trust? The behavioural foundations of the Italian North–South divide," *The Economic Journal* 126:1318-1341.

Bisin, A., and Verdier, T. (2001):"The economics of cultural transmission and the dynamics of preferences," *Journal of Economic theory* 97(2): 298-319.

Bisin, A., and Verdier, T. (2005): "Cultural transmission," The New Palgrave Dictionary of Economics.

Block, J. I., and Levine, D. K. (2016): Codes of conduct, private information and repeated games," *International journal of game theory*, 45: 971-984.

Boldrin, M., Christiano, L. J., and Fisher, J. D. (2001): "Habit persistence, asset returns, and the business cycle," *American Economic Review*: 149-166.

Bowles, S., and Gintis, H. (1976): Schooling in capitalist America (Vol. 57). New York: Basic Books.

Gintis, H., Bowles, S., Boyd, R. and Fehr, E. (2003): "Explaining altruistic behavior in humans," *Evolution and Human Behavior*, 24: 153-172.

Cason, T. N., Sheremeta, R. M. and Zhang, J. (2012): "Communication and efficiency in competitive coordination games," *Games and Economic Behavior* 76: 26-43.

Coase, R. H. (1960): "The Problem of Social Cost," *Journal of Law and Economics* 3: 1-44.

Campbell, J. Y. and Cochrane, J. H. (1999): "By force of habit: A consumption-based explanation of aggregate stock market behavior," *Journal of political Economy*, 107: 205-251.

Constantinides, G. M. (1990): "Habit formation: A resolution of the equity premium puzzle," *Journal of political Economy*, 98: 519-543.

Duffy, J. and Feltovich, N. (1999): "Does observation of others affect learning in strategic environments? An experimental study," *International Journal of Game Theory* 28: 131-152.

Cremer, J. and McLean, R. P. (1988): "Full extraction of the surplus in Bayesian and dominant strategy auctions." *Econometrica* 56: 1247-1257.

Bó, P. Dal (2005): "Cooperation under the shadow of the future: experimental evidence from infinitely repeated games," *American Economic Review* 95: 1591-1604.

Dutta, Rohan (2012): "Bargaining with Revoking Costs," *Games and Economic Behavior*, 74: 144-153.

Dutta, Rohan, David K. Levine and Salvatore Modica (2018): "Damned if You Do and Damned if You Don't: Two Masters," mimeo EUI.E., and S. Gächter (2000): "Fairness and retaliation: The economics of reciprocity," *Journal of Economic Perspectives* 14: 159-181.

Engel, C. (2011): "Dictator games: A meta study," *Experimental Economics* 14: 583-610.

Feddersen, T., A. Sandroni (2006): "A Theory of Participation in Elections," *American Economic Review* 96: 1271–1282.

Fehr, E. and S. Gächter (2000): "Cooperation and Punishment in Public Goods Experiments," *American Economic Review* 90: 980-994.

Fehr, Ernst and Klaus M. Schmidt (1999): "A Theory of Fairness, Competition and Cooperation", *Quarterly Journal of Economics* 114: 817-868

Fudenberg, D. and D. K. Levine (2011): "Risk, Delay, and Convex Self-Control Costs," *AEJ Micro* 3: 34–68.

Fudenberg, Drew, David Levine and Eric Maskin (1994): "The Folk Theorem with Imperfect Public Information," *Econometrica* 62(5): 997-1039.

Fudenberg, D., D. K. Levine and W. Pesendorfer (1998): "When are Non-Anonymous Players Negligible," *Journal of Economic Theory* 79: 46-71.

Frank, R. H. (1988): *Passions Within Reason: The Strategic Role of the Emotions*, WW Norton and Co.

Frank, R. H., T. Gilovich and D. T. Regan (1993): "Does studying economics inhibit cooperation?" *Journal of Economic Perspectives* 7: 159-171

Frank, R. H., T. D. Gilovich, T. D. and D. T. Regan (1996): "Do economists make bad citizens?" *Journal of Economic Perspectives* 10: 187-192.

Gale, D and Sabourian, H. (2005): "Complexity and competition," *Econometrica*, 73: 739-769.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001): "In search of homo economicus: behavioral experiments in 15 small-scale societies," *The American Economic Review* 91(2): 73-78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2005): "'Economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies," *Behavioral and Brain Sciences* 28: 795-815.

Jackson, M. O. (2010): *Social and Economic Networks*, Princeton university press.

Kahneman, D., J. Knetsch, and R. Thaler (1986): "Fairness as a Constraint on Profit-Seeking: Entitlements in the Market," A*merican Economic Review* 76: 728-741.

Kandori, M. (1992): "Social norms and community enforcement," *The Review of Economic Studies* 59(1): 63-80.

Levine, D. K. (1998): "Modeling altruism and spitefulness in experiments," *Review of Economic Dynamics* 1: 593-622.

Levine, David K. (2012): *Is behavioral economics doomed?: The ordinary versus the extraordinary* Open Book Publishers.

Levine, D. K. and A. Mattozzi (2017): "Voter Turnout with Peer Punishment," EUI

Levine, David and Salvatore Modica (2016): "Peer Discipline and Incentives within Groups", *Journal of Economic Behavior and Organization* 123: 19-30

Levine, David and Salvatore Modica (2017): "Size, Fungibility, and the Strength of Lobbying Organizations", *European Journal of Political Economy* 49: 71-83

Levine, D. K., and T. R. Palfrey (2007): "The paradox of voter participation? A laboratory study," *American Political Science Review* 101: 143-158.

List, J. A. (2007): "On the interpretation of giving in dictator games," *Journal of Political Economy* 115: 482-493.

Meyer, Christian Johannes and Tripodi, Egon, Sorting into Incentives for Prosocial Behavior (October 24, 2017). Available at SSRN: https://ssrn.com/abstract=3058195

Muthoo, Abhinay (1996): "A Bargaining Model Based on the Commitment Tactic," *Journal of Economic Theory* 69: 134-152.

Olson Jr., Mancur (1965): *The Logic of collective action: public goods and the theory of groups*,

Harvard Economic Studies.
Ostrom, Elinor (1990): *Governing the commons: The evolution of institutions for collective action*, Cambridge university press.
Palfrey, T. R. and Prisbrey, J. E. (1997): "Anomalous behavior in public goods experiments: How much and why?" *American Economic Review*, 829-846.
Ponemon, L. A. (1993): "Can Ethics Be Taught in Accounting?" *Journal of Accounting Education* 11: 185-209.
Prescott, E. C. and Townsend, R. M. (1984): "Pareto optima and competitive equilibria with adverse selection and moral hazard," *Econometrica*, 21-45.
Rahman, David (2012): "But Who Will Monitor the Monitor?", *American Economic Review* 102(6): 2767-2797.
Rand, D. G., Greene, J. D., and Nowak, M. A. (2012): "Spontaneous giving and calculated greed," *Nature*, 489: 427-430.
Robson, A. J. (1990): "Efficiency in evolutionary games: Darwin, Nash and the secret handshake.," *Journal of theoretical Biology*, 144: 379-396.
Roemer, John (2015): "Kantian optimization: An approach to cooperative behavior," *Journal of Public Economics* 127(C): 45-57.
Rogers, V. and A. Smith, A. (2008): "An Examination of Accounting Majors' Ethical Decisions Before and After an Ethics Course Requirement," *Journal of College Teaching and Learning*, 5: 49-54.
Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M. and Zamir, S. (1991): "Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study," American Economic Review: 1068-1095.
Schelling, Thomas C. (1956): "An Essay on Bargaining," *The American Economic Review* 46(3): 281-306.
Skarbek, D. (2014): "The social order of the underworld: How prison gangs govern the American penal system," O*xford University Press.*
Stahl, D. O., and P. W. Wilson (1995): "On players´ models of other players: Theory and experimental evidence," *Games and Economic Behavior* 10: 218-254.
Tirole, J. (2009): "Cognition and incomplete contracts." *American Economic Review* 99(1): 265-94.
Tisserand, J. C., Cochard, F., and Le Gallo, J. (2015): "Altruistic or Stategic Considerations: A Meta-Analysis on the Ultimatum and Dictator Games," Besançon: CRESE, Université de Franche-Comté.
Tangney, J. P., J. Stuewig and D.J. Mashek (2007): "Moral emotions and moral behavior," *Annual Review of Psychology* 58: 345-372.
Tangney, J. P. and R. L. Dearing, R. L. (2003): *Shame and Guilt*. Guilford Press.
Townsend, R. M. (1994): "Risk and insurance in village India," *Econometrica*, 539-591.
Turner, J. H. and J.E. Stets (2005): T*he Sociology of Emotions*, Cambridge University Press.

Theory Appendix

Corollary 5. *There are cutoffs $\eta_\gamma, \eta_\pi, \eta_\psi$ so that for $\eta < \eta_\gamma$ internalization ϕ is decreasing in γ and conversely, while for $\eta < \eta_\pi$ [resp. $\eta < \eta_\psi$] internalization ϕ is increasing in π [resp. ψ] and conversely.*

Proof. For $H < \gamma(1+\psi)\pi/(1-\pi)$ we have

$$\phi = \frac{1}{\gamma} g\left(\frac{1}{V}\left(1+\frac{H}{\gamma}\right), 0, \gamma\right)$$

so ϕ is constant in π and ψ; as to γ we have

$$\frac{\partial \phi}{\partial \gamma} = -\frac{1}{\gamma^2} g\left(\frac{1}{V}\left(1+\frac{H}{\gamma}\right), 0, \gamma\right) + \frac{1}{\gamma}\left[-\frac{H}{V\gamma^2} g_\mu\left(\frac{1}{V}\left(1+\frac{H}{\gamma}\right), 0, \gamma\right) + g_X\left(\frac{1}{V}\left(1+\frac{H}{\gamma}\right), 0, \gamma\right)\right].$$

Now at the upper bound $g_\mu = 0$ and $g_X = 1$ so the above derivative is zero. In the interior case $g_X = 0$ so

$$\frac{\partial \phi}{\partial \gamma} = -\frac{g\left(\frac{1}{V}\left(1+\frac{H}{\gamma}\right), 0, \gamma\right)}{\gamma^2}\left(1 - \frac{H}{\gamma}\eta\right)$$

and the conclusion follows directly.

For $H > \gamma(1+\psi)\pi/(1-\pi)$

$$\phi = \frac{1}{(1-\pi)\gamma/\psi + \gamma} g\left(\frac{1}{V}\left(1 + \frac{(1+\psi)\pi\gamma/\psi + H}{(1-\pi)\gamma/\psi + \gamma}\right), 0, \frac{(1-\pi)\gamma}{\psi} + \gamma\right)$$

In the interior where $\phi < 1$, letting

$$\xi = \frac{(1+\psi)\pi\gamma + \psi H}{(1-\pi)\gamma + \psi\gamma},$$

we have

$$\phi = \frac{1}{\chi} g\left(\frac{1}{V}(1+\xi), 0, \chi\right)$$

$$\frac{\partial \phi}{\partial \chi} = -\frac{1}{\chi^2} g\left(\frac{1}{V}(1+\xi), 0, \chi\right)$$

$$\frac{\partial \phi}{\partial \xi} = \frac{1}{\chi}\frac{1}{V} g_\mu\left(\frac{1}{V}(1+\xi), 0, \chi\right)$$

Now for $\theta = \gamma, \pi, \psi$:

$$\begin{aligned}\frac{\partial \phi}{\partial \theta} &= -\frac{1}{\chi^2} g\left(\frac{1}{V}(1+\xi), 0, \chi\right)\frac{\partial \chi}{\partial \theta} + \frac{1}{\chi}\frac{1}{V} g_\mu\left(\frac{1}{V}(1+\xi), 0, \chi\right)\frac{\partial \xi}{\partial \theta} \\ &= \frac{1}{\chi^2} g\left(\frac{1}{V}(1+\xi), 0, \chi\right)\left[-\frac{\partial \chi}{\partial \theta} + \chi\eta\frac{\partial \xi}{\partial \theta}\right].\end{aligned}$$

Here b is increasing in γ and decreasing in ψ, π. On the other hand ξ defined above is increasing in π; for ψ we have

$$\frac{\partial \xi}{\partial \psi} = \frac{(H+\pi\gamma)((1-\pi)\gamma + \psi\gamma) - ((1+\psi)\pi\gamma + \psi H)\gamma}{((1-\pi)\gamma + \psi\gamma)^2} = \gamma\frac{H(1-\pi) - 2\pi^2\gamma}{((1-\pi)\gamma + \psi\gamma)^2}.$$

We know $H > \gamma(1+\psi)\pi/(1-\pi)$ so, since $1 + \psi > \pi$ this is positive, that is ξ is increasing in ψ.

Also, ξ is decreasing in γ because

$$\frac{\partial \xi}{\partial \gamma} = \frac{(1+\psi)\pi\left((1-\pi)\gamma+\psi\gamma\right) - \left((1+\psi)\pi\gamma+\psi H\right)(1-\pi+\psi)}{\left((1-\pi)\gamma+\psi\gamma\right)^2} = \frac{-\psi H(1-\pi+\psi)}{\left((1-\pi)\gamma+\psi\gamma\right)^2} < 0.$$

The stated cutoffs spell out which term dominates in the various cases. □

Data Appendix

offer (output)	observations	rejections
1	1	0
1.75	1	0
2	4	2
2.50	5	1
3	10	2
3.25	5	4
3.50	6	1
3.75	5	1
4	30	5
4.25	9	0
4.50	17	5
4.75	5	0
5	31	0
5.25	1	0