Interventions when Social Norms are Endogenous: A Critique[☆]

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

Abstract

Lucas's critique of adaptive expectations argues that treating expectations as exogenous when they are endogenous can lead to important policy mistakes. More broadly the critique is of treating the endogenous as exogenous. Here we make the same argument with respect to social norms. Typically social norms are treated as exogenous in the face of substantial evidence that they are endogenous. We present a simple stylized model of endogenous social norms and examine ways in which they can lead to erroneous conclusions. We first examine how misunderstanding of the optimal nature of social norms may lead to misinterpretation of evidence and misguided policy prescriptions. This point we illustrate with two examples. In the first we show how populist working class resentment of the lax supervision of professional classes may be wrongly attributed to political power. In the second we show how the impact and non-impact of double-blind laboratory treatments are properly understood only in the context of optimal social norms. Our second setting is one of social norms that may change in response to changed circumstances. In our first illustrative application we show that how public good production can decrease when it is subsidized - and that this is nevertheless evidence of an increase in welfare. In the second we show how small interventions such as changes in the minimum wage in particular locations may result in no reduction in employment - while large interventions across many locations may lead to substantial employment reduction. We argue that this may explain the discrepancy in the empirical literature between "natural experiments" and time series analysis of minimum wage changes. Finally, we turn to the issue of internalization of social norms. This is widely understood to be important - but like other aspects of social norms internalization is not fixed or magical but endogenous. Here we show how the naive idea that laboratory results carry over directly to the field can lead to mistaken conclusions while at the same time showing how appropriate use of laboratory results can serve as important confirmation for hypotheses about the field.

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 Mattozzi and seminar audiences at WUSTL, Warwick, Queen's, the Paris Institutions Conference, the Zurich Political Economy Conference 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

1. Introduction

We study a theory of endogenous social norms. The theory leads to a mapping from economic fundamentals to social norms. The theme of the paper is that ignoring or failing to understand this mapping is dangerous. It leads to erroneous conclusions about welfare and public policy and it leads to the misinterpretation of laboratory and empirical results.

Empirically the importance of self-enforcing social norms in enabling groups to overcome public goods problems is well-established[4] and there is some theoretical work on the subject. In contrast to work in behavioral economics where pro-social behavior is seen as an intrinsic part of preferences, social norms are endogenous and adapt themselves to circumstances: this can be clearly seen in the cross-cultural experiments of Henrich et al (2001). Here we present a theory of endogenous social norms - but also cultural norms and institutions which we do not regard as intrinsically different - based on the key idea of peer enforcement and collective decision making.

Our model is one of individual behavior - Nash equilibrium with respect to selfish preferences - in which decisions are collective only in the sense that groups have the ability only to coordinate on a particular equilibrium that is mutually advantageous. In this theory pro-social behavior arises because there are penalties for anti-social behavior. We also consider the possibility that individuals find it personally advantageous to internalize social norms - resulting in apparently altruistic behavior despite the fact that individuals have no intrinsic preference for altruism.

Specifically we elaborate on the model of peer incentives introduced in Levine and Modica (2016) and used in Levine and Modica (2017) to study lobbying groups and in Levine and Mattozzi (2017) to study political parties. Here we focus on the issue of peer monitoring - an issue that also arises in the earlier work of Kandori (1992). We study a model of public goods production by group members who are monitored by other group members. Both the producers and monitors face incentive problems: the producers because the group would like to induce them to take costly actions that provide a public benefit and the monitors because they have to choose whether to report accurately their noisy information about the actions they observe. After reports are received the group engages in individually valuable social activities, and bad reports about a producer can be punished by ostracism. However, ostracism can be costly for the punishers as well as the punished.

Our basic hypothesis is that the group designs an incentive compatible mechanism for itself that is mutually beneficial for members. That is, we do not necessarily assume that social norms are left-over from some past meaningful equilibrium - we assume that groups can at some cost change social norms to reflect changed circumstances. In this direction we point to three pieces of evidence: the rapid change in social norms (measured in minutes) concerning the treatment of airplane hijackers that took place on September 11, 2001; the change in social norms (measured in months) concerning public protest that took place in East Germany following the commitment by Gorbachev that military intervention in East Europe was off the table; and the rapid and organized change in social norms (following a debate that lasted over 12 years) that took place in Sweden

[4]Particularly see Olson (1965) and Ostrom (1990) among many others.

when the change was made from left-side to right-side of road driving.[5] In these examples the incentives for change were large and in two of the cases took a substantial amount of time to come to fruition. The view we take is that there are fixed costs associated with introducing or changing social norms and we investigate both the nature of optimal social norms and the consequences of frictions for changes in social norms. This idea is consistent also with evidence such as Bigoni et al (2016) or Belloc, Drago and Galbiati (2016) that sometimes social norms are very persistent: lasting decades or even centuries.

Methodologically this paper is designed to examine a simple theory of optimally chosen social norms that are are either enforced through incentives or internalized (or both) and which, in a dynamic setting, are subject to adjustment cost. Our goal is to show that this theory is consistent with a wide range of phenomena that superficially fail standard tests of rationality. We think that a good theory must not only explain a single fact well but must explain many facts. We shall not provide an in depth analysis or a definitive answer to any particular question, but rather examine the implications of the theory across a broad range of applications, identify the strengths and weaknesses of the theory, and lay the groundwork for future empirical analysis that will validate or invalidate the theory. The theory draws on empirical insights about group behavior documented by Olson, Ostrom and many others. In earlier work (see especially Levine and Mattozzi (2017)) the theory was successfully used for a detailed study of the particular problem of voter turnout. Here our applications are intended to be illustrative rather than definitive. They are intended to document how the failure to understand the endogenous nature of social norms may lead to misleading conclusions about welfare, public policy and both laboratory and empirical results.

We first study one time choice of social norms. In our first application we consider how monitoring differs between the working class and professional class. We argue that the differences that come about because the cost of punishing the monitor depends on how socially close the monitor is to the producer lead to greater laxity of monitoring of professionals - and likely to populist resentment of the professional class by the working class. Our second application shows how the optimality of social norms outside the laboratory may lead to the failure of procedures such as double-blind designed to reduce or eliminate possibility of outside influence.

We then turn to the role of adjustment costs in changing social norms. Our first application is to introducing subsidies for the provision of a public good. Naturally this tends to increase public good output. However in the face of fixed costs for implementing non-trivial social norms external incentives can have a perverse effect as has been noted in the experimental literature: introducing a subsidy may reduce output. This is due to an endogenous change in norms, and behavioral analysis that views social norms as fixed is misleading. An unstated assumption in the literature is that the reduction in output is bad for welfare: we show that increasing the subsidy always increases welfare even if it lowers production of the public good. The key point is that public good production is reduced because a costly monitoring scheme is eliminated and this cost reduction compensates for

[5]See the discussion in Levine (2012) for details of these three cases.

the reduction in public goods output.

We then consider a more general Lucas critique: interventions that are not sufficiently widespread - in the laboratory certainly, but even in "natural" experiments - may not lead to a change in social norms while a broader intervention may change social norms and so have different consequences. Hence interventions that appear effective in the "small" may not work "in the large" or vice versa. We illustrate this idea with an application to the minimum wage.

Finally, we examine the implications of the fact that social norms may be internalized. We model this by assuming that group members may specialize by investing in particular strategies. As a result of this costly investment, the chosen strategy provides a utility benefit when it is used. We assume moreover that the investment is not individually profitable but will be made only when subsidized by the group. One obvious conclusion is that internalization can have large effects: in particular production that is not feasible without internalization may be feasible with internalization. We show moreover that internalization is more important for monitors than for producers and study an application where there are both small and high stakes matches showing how as the value of the public good increases punishment in small stakes matches initially rises then declines (and internalization substitutes for punishment) while in high stakes matches punishment only increases (and internalization is a complement for punishment). Consequently inferences about punishment and internalization drawn from the laboratory (small stakes) may have misleading implications for behavior outside the laboratory (high stakes). We review the Henrich et al (2001) cross-cultural experiments in this context.

We do not provide an extensive literature review in the introduction - rather we comment on the relevant literature in the context of our specific assumptions. Our basic model is a variation on the workhorse principal agent model - albeit one in which there is a monitor and costly punishments. The contribution of this paper is not in the fact that this variation is different from the many principal agent models that have been studied, but rather in the questions about social organization it enables us to address.

2. The Base Model

We study a large group of unit mass.[6]. There are two stages. In the *production stage* members are randomly matched into pairs one as producer and one as monitor. We refer to the pair as partners. The producer may choose to provide effort $e \in \{0,1\}$ to produce a public good at a cost of ec where $c > 0$. The social value of the public good produced is V, hence if all producers provide effort ($e = 1$) the aggregate per match social value of public goods production is V. The producer's choice e generates a noisy signal $z \in \{0,1\}$ where with probability π the signal is wrong $z \neq e$ and with probability $1 - \pi$ the signal is correct $z = e$ and $\pi < 1 - \pi$. This signal is observed by the monitor matched with the producer. The monitor then makes a report $x \in \{0,1\}$, interpreted as a bad or good report.

[6]For technical details on how a continuum model like this works see Ellickson et al (1999).

Following the production stage there is a *social interaction stage.* During this social interaction members may be *ostracized* based on the identity of the producer and the monitor and the report of the monitor. We assume that there is constant marginal cost of ostracism with respect to the number of members ostracized: the cost to a member who is ostracized is normalized to 1, the cost of having a partner ostracized is $h \leq 1$ and the per match social cost of ostracizing one member of that match is $U \geq 1$.

Example. The population is rematched into social subgroups of size $N \geq 4$ for a social meeting. Partners have a probability H of being matched with each other. One member at each meeting is chosen at random and may be excluded. If N members are at the meeting each receives N; if one member is excluded each receives $(1-\nu)N$. A member who is ostracized is excluded from the meeting at which they appear: such a member has a $1/N$ chance of losing the benefit N, so the expected cost of being ostracized is 1. The probability of being present at a meeting where a partner is excluded is H/N and the utility loss is νN so that $h = \nu H$. We may then compute the per match social cost of ostracizing a match partner as the probability that partner is selected at their social meeting $(1/N)$ times the loss of utility to the meeting at which the exclusion takes place $N + (N-1)\nu N$ so that $U = 1 + (N-1)\nu$.

An *ostracism rule* is a probability of being ostracized based on the type of member - producer or monitor - and the report by or about the member. A *strategy* for a member is a type contingent decision on whether or not to produce and what report to make as a function of the signal observed. A *truthful strategy* is to report the observed signal as a monitor. A *social norm* is a truthful strategy together with an ostracism rule. The *default social norm* is the truthful strategy of not producing together with the rule of not ostracizing any members. Notice that if everybody follows this strategy it is a Nash equilibrium. In a *productive social norm* the strategy is the truthful strategy of producing. If a productive social norm is a Nash equilibrium we say that it *implements production* and that the ostracism rule is *incentive compatible.* Notice that ostracism is costly for the monitor: if $h > 0$ the monitor shares part of the cost of ostracizing the producer and so has incentive to let the producer off the hook. Hence to implement production the ostracism rule must provide incentives for the monitors as well as producers.

Our assumption is that the group chooses the *optimal social norm* that maximizes per match *ex ante* utility of any of the ex ante identical group members. We refer to this utility as the *social utility.*

Example. [continued] In the meeting example an ostracism rule can be decentralized in an incentive compatible way. Prior to the meeting the member eligible for exclusion is chosen and the type of that member and the report (if any) about the production match in which that member participated is available to the members in that meeting. There is a public randomizing device that enables meeting members to coordinate their decisions. Based on the reports and the randomization device the members at the meeting vote on exclusion. There is a number $1 < K < N-1$ such that exclusion take place if and only if K or more members of the audience vote for it. With $N \geq 4$ if all vote in favor or all against no voter is decisive, so any exclusion rule is incentive compatible.

Monitoring Technologies. In the web appendix we allow the possibility that there is more than a single monitor and examine three monitoring technologies: *collusive monitors* who can coordinate their reports; *independent monitors* who cannot; and *public information* where monitors are constrained to tell the truth but are anonymous. We show that the case of collusive monitors is qualitatively the same as a single monitor, that public information is equivalent to $h = 0$, and that independent monitors are similar to public information. Hence we may think of a small number of collusive monitors as corresponding to $h > 0$ and public information or a large number of independent monitors as corresponding to $h = 0$.

Implementation Cost. We normalize the social utility from the default social norm to 0. Alternatively the group may try to implement production. If it is possible to do so we say that *production can be implemented*. The only tool for implementation is the ostracism rule. That is, by incurring some part of the social cost U from ostracism it may be possible to obtain V, the social utility from public good production. Given a productive social norm s let $L(s)$ denote the per match expected number of members who are ostracized on the equilibrium path. The corresponding per match expected social utility $W(s)$ is the per match payoff from production V minus the per match cost of production c, minus the expected cost of ostracism: $W(s) = V - c - L(s)U$. In solving the problem of finding an optimal social norm it is useful to separate the problem of the optimal incentives for production from the problem of whether or not to produce at all. To this end note that $W(s) = V - [c + L(s)U]$ and define the *cost of implementation* as $C(s) = c + L(s)U$ - per match production cost c plus expected lost per match social utility $L(s)U$ which is the monitoring cost. If s is to be an optimal social norm it must minimize implementation cost, and implementation will be optimal if and only if $V \geq \min_s C(s) \equiv \underline{C}$.

Heterogeneous Matches. Suppose that production matches are heterogeneous in the sense that $\tau = (V, c)$ varies from match to match but is known to the partners and is known when ostracism decisions are taken. From the formulation of the homogeneous problem as one of designing an ostracism rule to minimize $\min_s C(s)$ and of implementing production whenever $V \geq C$ we see that we can solve this problem separately for each type of match and that the optimal social norm will be given by the minimum cost function $\underline{C}(\tau)$ together with the rule of implementing production exactly when $V \geq \underline{C}(\tau)$.

3. Cost Minimizing Social Norms

We characterize cost minimizing social norms. Let $0 \leq \lambda \leq 1$ and define the *monitoring cost factor*

$$\mu(\lambda, \pi, h) = \frac{\lambda + (1 - \lambda)h}{(1 - 2\pi)(1 - h^2)}.$$

Our main result follows from a more general result in Web Appendix 1:

Theorem 1. *Define*

$$P = \mu(1, \pi, h)c.$$

Production can be implemented if and only if the implementation condition $P \leq 1$ *is satisfied. In this case a cost minimizing ostracism rule exists and satisfies*

1. a producer who gets a good report is not ostracized; a producer who gets a bad report is ostracized with probability P

2. a monitor who files a bad report is not ostracized; a monitor who files a good report is ostracized with probability $Q = hP$.

The cost of implementation is given by

$$C = [U\mu(\pi, \pi, h) + 1]\, c.$$

The remainder of the section is devoted to understanding the implications of the key parameters V, c, h for the structure of optimal social norms. We begin by making some preliminary observations.

- The implementation condition is crucial: if it is satisfied then any sufficiently valuable public good, that is, large enough V, will be produced, and if it is not no public good will be produced no matter how valuable.

- If h is sufficiently close to one the implementation condition fails. The reason for this is a feedback effect: a bigger punishment for the producer implies a bigger punishment for the monitor. The feedback effect is that the latter reduces the incentive for the producer to produce: by not producing she can reduce the probability the monitor is punished for sending a good report. A high degree of social interaction (h) makes this feedback effect very strong and consequently implementation is impossible.

- With private information malicious gossip is valued in the sense that the monitor is less likely to be ostracized when a bad report is filed.

- The cost of implementation is proportional to c the incentive to cheat on the social norm. This is a robust and common result in peer monitoring models: for example Levine and Modica (2016) or Levine and Mattozzi (2017). It follows from the constant marginal cost of ostracism assumption.

3.1. Rotation, Expertise, and Populism

With private information the social closeness between monitor and producer, measured by h, matters. By manipulating the matching process a group may be able to vary the social distance between the monitor and producer. This leads in a natural way to a tradeoff: using monitors with greater social distance makes it less likely they will interact socially with the producer (lower h), but may also make it more difficult to accurately observe the production decision (higher π). Here we examine how the tradeoff depends on the production technology.

One practical method of varying the social distance between monitor and producer is the use of supervisor rather than peer evaluation. In the literature on personnel management a great deal of attention is on which system provides the best incentives. Generally speaking we expect that peers

interact with each other and supervisors interact with each other, but interactions between the two groups is less common - in other words supervisors have greater social distance than peers. Indeed, in some instances peers and supervisors are actively discouraged from interacting: for example in the military officers clubs used to be common to encourage officers to socialize with one another but not with enlisted ranks. We expect then that supervisor evaluation will deliver lower h - albeit at the cost of higher π. Indeed there is data - see for example Kraut (1975)[7] that indicates that peer evaluation is substantially more accurate than supervisor evaluation.

A second method of varying the social distance between monitor and producer is the system of rotation. In the police, for example, police officers who monitor their peers may be periodically moved between precincts to deliberately break social ties. Rotation increases the social distance because police officers know their colleagues less well. Hence we expect that it will lower h - again at the cost of higher π. As in the case of supervisor monitoring a common complaint is that the effectiveness of monitoring is reduced as the monitors have less interaction with and knowledge of the producers.

To study more clearly the trade-off between π and h let us assume a trade-off in the form of a smooth *accuracy function* $\pi = f(h)$ where $f'(h) < 0$ so that increasing h raises monitoring accuracy $1 - \pi$. To focus thoughts, as an illustrative example consider the police as a prototypical working class occupation and surgeons as a prototypical professional occupation. In both cases there is a substantial public goods output in the form of good reputation: corruption or excessive use of force by the police gives all police a bad reputation; lack of effort by doctors results in poorer patient outcomes, reduced demand for the services of doctors and less income for all doctors. In both cases group members have incentives to self-organize to reduce bad behavior.[8] We note moreover that in the police - as in working class occupations more generally - social distance to peers is low but supervisor evaluation is common and rotation is sometimes used as well. By contrast in professions such as the medical profession monitoring is done almost entirely by peers. Can this be explained by differences in the tradeoff $\pi = f(h)$ between the two types of occupations?

What are the economic fundamentals - that is, what does the function $f(h)$ look like in the two cases? We observe first that surgeons require a high level of specialized knowledge - more than a decade of specialized training[9] - while police officers require less than a year.[10] We interpret that to mean that the sensitivity of f to h is much greater for surgeons than for police officers - outsiders are unlikely to have the specialized knowledge needed to evaluate "surgical output" while it is not so difficult for an outsider to evaluate "police output." Specifically denote by f_S the function for

[7] Most studies in this literature look only at the correlation between peer and supervisor rating or the within group correlation of rankings ("reliability"). Kraut (1975), by contrast, looks at peer and supervisor evaluations made at the end of a four week training course and shows that peer evaluation is a far better predictor of subsequent promotions.

[8] A common form of ostracism in the medical profession is to refuse to refer patients to other doctors: see Kinchen et al (2004) and Sarsons (2017) who document that perceived medical skill is the most important factor in surgeon referrals and that bad surgical events lead to reduced referrals.

[9] https://study.com/articles/Surgeon_Career_Summary_and_Required_Education.html

[10] http://work.chron.com/long-train-cop-21366.html

surgeons and f_P that for police: we expect that $|f'_S(h)| > |f'_P(h)|$ - that reducing the social distance and hence increasing the level of expertise of the monitor will make a great deal of difference for the accuracy with which surgical output is observed but not so for police output.

We observe second that for any given level of expertise it is more difficult to observe surgical output than police output: we expect that $f_S(h) > f_P(h)$. Compare, for example, improper behavior by police versus malpractice by doctors: from survey and other data the fraction of bad signals in response to bad behavior is greater for the police (about 3.2%, see Langton and Durose (2013)) than for doctors (less than 1%)[11] indicating better signal quality for police than surgeons.

Recall from Theorem 1 that the implementation cost is given by $C = [U\mu(\pi, \pi, h) + 1]\, c$ so the group's objective is to minimize $\mu(f(h), f(h), h)$ under the constraint $P \leq 1$. The consequences of these fundamentals on cost minimizing h and minimal implementation costs are stated below and proven in Web Appendix 2.

Theorem 2. *Suppose that $0 < f_S(h), f_P(h) < \pi/2$ are accuracy functions and let C_S, C_P denote the corresponding minimal implementation costs. If S has a noisier signal $f_S(h) > f_P(h)$ then $C_S > C_P$. If in addition S has greater signal sensitivity $|f'_S(h)| > |f'_P(h)|$ and c is sufficiently small that there exist unconstrained cost minimizers $\hat{h}_j$ that satisfy the constraint $P_j \leq 1$ then for any cost minimizers $\hat{h}_S, \hat{h}_P$ either $\hat{h}_P = 0$ or $\hat{h}_S > \hat{h}_P$.*

The last part of the result tells us we should see greater social closeness in the monitoring of surgeons than of police. To understand more clearly the consequence of greater implementation cost for surgeons, recall that production is implemented only when $V \geq C$. We imagine that a social norm anticipates that there can be many different values of V hence takes the form of a rule for implementation as a function of V. What the theorem says then is that the range of cases V for which surgeons will be expected to produce the public good is narrower than for police. That is, roughly speaking, the theorem says that surgeons are chummy with their monitors and "get away with" more stuff than police.

The theorem seems to reflect reality. One form of low value public good, for example, is being on time: nobody can have failed to notice that doctors are never on time for appointments, while working class who are late for work are generally punished. These facts are relevant to the political analysis of populism. One of the root causes of populism is working class resentment of professionals.[12] One source of this resentment is the (correct) perception that professionals are laxly monitored: they are chummy with their monitors and get away with more stuff. This is often attributed by the working class to the political power of elites. This analysis shows that instead it may be due to the different nature of monitoring accuracy. Notice that there may indeed be a social problem to be solved: there is no reason to believe, for example, that the cost to surgeons of unnecessary surgery is as great as it is for the patients. Our analysis does point to an appropriate

[11] In Civil (2000) about 3% of cases where malpractice is documented in medical records lead to claims, while the actual incidence of malpractice is estimated to be 4 times higher.

[12] See, for example, Williams (2016).

remedy: not populism, whatever that means, but rather collective punishment for the professional class. What might be politically and practically feasible is hard to say - but, for example, a tax on all surgeons based on the number of fatal surgical accidents would encourage surgeons to tighten their self-regulation.

3.2. Quality and Fairness

Typically a social norm not only specifies whether or not a public good should be produced, but how much or what quality. We now extend the analysis to a choice of production level or quality denoted by $\theta \geq 0$. We assume that we can normalize the units of quality so that the cost of producing at the level θ is given by $c = \theta^2$ and that the value of public good produced is $v\theta$. Given a social norm of θ the individual producer may choose to produce $e\theta$ where $e \geq 0$; if $e = 1$ production is θ and the norm is followed, otherwise it is not. We continue to assume a simple signaling technology with just two signals $z \in \{0, 1\}$ which we think of as meaning "bad, the social norm was not followed" and "good, the social norm was followed." Specifically if $e = 1$, that is, the producer follows the social norm, then with probability $1 - \pi$ the signal is 1 and with probability π the signal is 0. If $e \neq 1$ then with probability $1 - \pi$ the signal is 0 and with probability π the signal is 1. With this structure it is clear that if the producer chooses not to follow the social norm the optimal deviation is to produce 0 since the chances of being punished are the same for any deviation.[13]

The analysis of the simple model is straightforward observing that now c is not exogenous but it is equal to θ^2. The following theorem is proven in Web Appendix 2:

Theorem 3. *With variable quality and quadratic cost the optimal social norm is*

$$\widehat{\theta} = \min\left\{\sqrt{1/\mu(1, \pi, h)}, v/(2(1 + U\mu(\pi, \pi, h)))\right\}$$

One less obvious application of public good "quality" is to social norms of fairness. One type of "production" opportunity is to give away resources obtained through luck. Here θ represents the size of a gift to give away. There are two reasons why gift giving can be a public good. First, there is an insurance motive: when explicit insurance contracts are costly sharing lucky gains can substitute for otherwise lacking insurance markets. Second, excessive effort to "be first" to make a lucky find can be avoided if the winner has to share the find.[14] In this setting of socially valuable gift giving the social norm θ represents a notion of "fairness:" how much to share. Rather than accepting that there is some arbitrary notion of fairness the theory of endogenous social norms directs us to look beyond some intrinsic notion of fairness and ask why should this be socially optimal?

[13] A more refined signaling technology must include at least this incentive constraint, but might have additional incentive constraints - for example, if small deviations are less likely to be detected than large deviations.

[14] The social cost of racing to be first reaches an absurd height in the case of high frequency trading which, unfortunately, cannot be avoided through informal social norms. See Cramton, Budish and Shim (2015).

3.3. Information Leakage and Double-Blind in the Laboratory

In our next application we consider inadvertent information leakage. As example we might think of the Panama papers or the HSBC leak where information about supposedly confidential offshore bank accounts became public. In these cases information becomes public so if we wish we may consider that $h = 0$ although this plays no role in our analysis. To model information leakage we now allow the possibility that not all matches are monitored.

Specifically we assume that each producer has probability η of being monitored by a single monitor and probability $1 - \eta$ of not being monitored.[15] Producers do not know if they are monitored, that is, whether or not information may leak. In studying the role of η we consider that the value of η may vary from match to match so that the social norm specifies contribution level θ and punishment as a function of the (commonly known) value of η for that match. When the match is monitored we say that there is *information leakage*.

The key fact from Web Appendix 1 is that only the level of punishment in Theorem 1 changes: now $P = \mu(1, \pi, h)c/\eta$. That is if there is only a chance η of "getting caught" then the punishment must be increased proportionately. The cost of implementation does not change: the proportionately higher cost of punishment is incurred less frequently and η cancels out of implementation cost entirely. In the case of variable quality this means that the constraint binds for smaller values of θ and in Web Appendix 1 it is shown than

$$\widehat{\theta} = \min\left\{\sqrt{\eta/\mu(1,\pi,h)}, v/(2(1+U\mu(\pi,\pi,h)))\right\}.$$

What does this say about how the optimal social norm depends upon η? Lowering the probability of information leakage η does not change the optimal social norm $\hat{\theta} = v/(2(1+U\mu(\pi,\pi,h)))$ until the threshold $\eta = \mu(1,\pi,h)\left[v/(2(1+U\mu(\pi,\pi,h)))\right]^2$ is reached. As we have observed, up until this point a lower chance of information leakage is compensated by a greater chance of punishment, maintaining incentives without changing monitoring cost. After this point the level of punishment is maintained at $P = 1$ and the optimal $\hat{\theta}$ correspondingly declines.

The idea of information leakage is interesting also with respect to studies particularly of social preferences such as the dictator game (see for example, Tisserand et al (2015)) where it is believed that participants behave altruistically to make a good impression on others - in particular the experimenter. In an effort to eliminate this a double-blind treatment is often used in which neither the other participants nor the experimenters can tell who did and did not donate money.

We wish to propose a rather different interpretation of behavior and motives. Specifically: we believe that what participants are "worried" about is violating a social norm of fairness and getting caught. Participants are assured that their behavior in the laboratory will not be "leaked" to the outside world ("what happens in Vegas stays in Vegas"). In the literature it is generally assumed

[15] Hence if ζ is the mass of monitors the mass of monitors in matches $\eta(1-\zeta)$ should equal the mass of monitors. Web Appendix 1 studies the general case where only a fraction η of matches are monitored, if a match is monitored it is monitored by k monitors, and not all monitors need be matched, so that the general assumption is $\eta k(1-\zeta) \leq \zeta$.

that these representations are believed. We do not believe these representations are true, nor do we believe that participants believe this is the case. We have two reasons for this doubt:

1. Mistakes happen. If hackers can obtain confidential and damaging emails from Yahoo - not to speak of HSBC - what are the chances the experimental records are so secure that they will never leak to the outside world?

2. Even if identities are protected - for example through double-blind - there is a long history of deception in experiments by psychologists who have systematically lied to their subjects. What, for example, is to keep a deceptive experimenter from using a secret camera to record supposedly confidential placement of money into an envelope?

We argue that while through instructions, design, and reputation, the perceived probability of being observed (the one that matters) may be made small, it is unlikely to be made zero. Subjects - rightly - have some concern that if they behave selfishly in the laboratory word of this will get back to their friends outside the laboratory and they will then have an unfortunate reputation for behaving badly when they think nobody is looking.

What does the theory of endogenous social norms tell us? At issue is η the probability of inadvertent information leakage. Consider an effort to reduce perceived η through instructions, design and the like. Here it is crucial to understand the nature of the optimal social norm: if there was a fixed penalty for violating the social norm a reduction in η would decrease giving. By contrast we have seen that the optimal social norm says that up to a point $\hat{\theta}$ does not decrease, but rather the penalty P increases as η decreases in such a way that $\hat{\theta}$ remains constant. Once the threshold $P = 1$ is reached $\hat{\theta}$ begins to decline. Hence as efforts are made to reduce η the (unobserved) probability of ostracism outside the lab will increase but we will see no change in behavior inside the lab. This means that modest efforts to reduce η should have little or no effect, yet strenuous effects can cross the threshold $\eta = \mu(1, \pi, h)\left[v/(2(1 + U\mu(\pi, \pi, h)))\right]^2$ and so substantially reduce giving. We offer this as a possible explanation for the following stylized fact about dictator experiments (see for example, Tisserand et al (2015)): many studies have found that giving ($\hat{\theta}$) in double-blind is the same as in single-blind, while a few - those that make a strenuous effort to reduce η - find that it substantially reduces giving. We suggest that the difference between these studies lies in the extent to which η was successfully reduced: where the effort was modest we would expect no effect, but where the effort was strenuous we would expect an effect. That is: it is not "double-blind" versus "single-blind" that matters - it is how persuasive the double-blind is that matters.

4. Changing Existing Social Norms: a Lucas Critique

So far we have studied what can be described as frictionless social norms: social norms are adopted to maximize social utility. In practice it is neither instantaneous nor costless for large groups to discuss and agree on social norms, and there is always the option of simply settling on the default equilibrium - agreeing to nothing and "letting nature take its course," each individual following their own personal interest without monitoring and ostracism. The existence of frictions has implications for experimental and empirical economics. In particular, interventions - changes

in incentives for producing public goods - will have a different effect depending on whether or not they are sufficient to overcome the "friction" of changing social norms. This can lead to perverse consequences where incentives designed to encourage public goods production instead reduce it because they displace peer monitoring and ostracism. Moreover, small scale interventions - either in the laboratory, in the field, or as measured in a natural experiment - may be insufficient to change social norms and so may provide misleading guidance about large scale interventions which are sufficient to change social norms. This latter is a kind of Lucas critique: when social norms are endogenous data generated with fixed social norms does not tell us about what happens when social norms change in response to policy.

In this section we study policy interventions thought *a priori* to be unlikely so the existing social norm does not incorporate a response. After the intervention one of three things can happen: the existing norm can be maintained, there can be reversion to the default norm, or the group may pay a fixed cost to introduce a new social norm that is responsive to the policy.

We study two examples. In the first example we consider an efficient intervention in the form of a fine (or Pigouvian tax) in a situation where social incentives are aligned between the group and society. A prototypical case is picking up children late from a day care center. In the second example we consider increasing the minimum wage in the setting of a socially enforced cartel.

4.1. A Simple Incomplete Contracting Model

We are interested in a two period model where there is an existing social norm and circumstances change such that it may be optimal to alter the social norm. One question that arises is why ever change the social norm at all: as we have indicated in our discussion of heterogeneous matches there is no reason a social norm should not be contingent on match characteristics, such as different values of η. Here we take the same perspective as that in the incomplete contracting literature: it makes sense to plan in advance for likely contingencies but not for unlikely ones.

We offer a simple model in the spirit of Tirole (2009) and Dye (1985). Denote by φ a non-negative variable representing the extent of a policy change. In our first example it will be the size of a fine, in our second it will be the fraction of firms covered by a minimum wage. When the group originally designs the social norm it anticipates that with (high) probability $1-p$ there will be no policy intervention and that $\varphi = 0$ but with (small) probability p there will be a policy intervention and $\varphi > 0$. The group then faces a choice: it can design a norm specific to $\varphi = 0$ or it can design a contingent norm specifying what targets and punishments should be for different values of φ. The latter is more complex, requires more computation and discussion. Denote the added cost of doing this by $f > 0$. Alternatively the group can wait and see if there is a policy intervention, designing a responsive norm only after it is clear that $\varphi > 0$. It is natural to think the latter is more expensive than making the plan in advance: not only will there be computation and discussion, but the group must be reconvened to reach an agreement.[16] Denote by $F > f$ the

[16]The fixed cost might well depend on the size of the group: for example Levine and Modica (2017) assume it is proportional to group size. Here we are keeping the size of the population fixed and normalized to 1 - although of

cost of *ex post* design of a responsive probability after φ is known. As the probability that $\varphi > 0$ is p it follows that it is optimal to wait whenever $\varphi > 0$ is sufficiently unlikely in the sense that $p \leq f/F$.

We focus next on contingencies that are are *a priori* unlikely so not planned for in advance. Hence there is a fixed cost F of introducing a new social norm that accounts for $\varphi > 0$. Alternatively the group can simply stick with the existing social norm (not reconvene the group). Thirdly the group can revert to the default norm, where no ostracism occurs. Note that there is an intrinsic asymmetry between the default social norm and implementing production, because the former is completely decentralized: the group need not do any organization, members are simply left on their own to optimize. By contrast implementing production requires agreement over a production level and enforcement scheme. It is natural then to think that it is less costly to switch to the default social norm than to build a new one - and for simplicity we take the cost of switching to the default social norm as zero.

4.2. Fines and Frictions

Consider the setting of Gneezy and Rustichini A (2000) who study the role of incentives in an experiment in which a modest fine is introduced for being late to pick up children at day-care. We model this with variable quality θ with higher values representing the frequency with which the child is picked up on time. We maintain the basic setup of section 3.2 where the value of public good produced is $v\theta$ and the cost is θ^2. Therefore absent any enforcement issues the optimal quality is $\theta^* = v/2$ and we think of this as representing "always pick up the child on time." We then suppose that a fine $\varphi \leq v$ may be assessed for late pickup so that a parent picking up a child on time with frequency θ must pay $\varphi(\theta^* - \theta)$. Notice that the information technology underlying the fine is better than that available for social enforcement: there is no noise in the observation of lateness by the school officials imposing the fine.[17] We assume that the fine simply represents a transfer within the group (the school exists to serve the parents) so there is no conflict over social objectives.

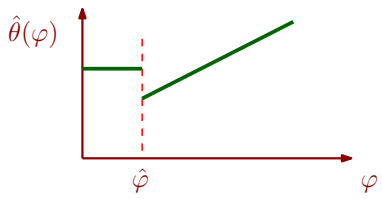
We assume that the imposition of fines is *a priori* thought to have low probability - and is of uncertain duration when introduced - hence it does not make sense to bear the fixed cost of designing a new optimal norm and the only relevant decision by the group is whether to keep the existing social norm adapted to $\varphi = 0$ or to switch to the default social norm. The default norm, given φ, is the θ that minimizes individual cost $\theta^2 + \varphi(\theta^* - \theta)$ that is $\theta = \varphi/2$. Let $\hat{\theta}(\varphi)$ denote the quality produced after fines are introduced. In Web Appendix 2 we show:

Theorem 4. *There exists $\overline{F} > 0$ such that for every $F \geq \overline{F}$ there is a $0 < \hat{\varphi} < v$ such that for $0 \leq \varphi < \hat{\varphi}$ the existing social norm is optimal and output is $\hat{\theta}(\varphi) = \hat{\theta}(0)$, while for $\varphi > \hat{\varphi}$ the*

course in practice it depends on whether we are talking about 100% of the population of New Jersey or the United States.

[17] This is not so uncommon: the IRS, for example, generally knows more about income and tax payments than friends and neighbors.

default social norm is optimal and $\hat{\theta}(\varphi)$ *is continuous, increasing and satisfies* $\hat{\theta}(\hat{\varphi}) < \hat{\theta}(0)$ *and* $\hat{\theta}(v) > \hat{\theta}(0)$. *This is illustrated in the figure below.*



The striking fact is that a drop in public good output with respect to a fine for bad behavior has been observed. Gneezy and Rustichini A (2000), in particular, showed how introducing modest incentives can lead to the discouragement of the activity it is designed to promote: they show that introducing a modest fine for being late to pick up children at a day-care resulted in more parents picking up their children late. This is consistent with the theory here if prior to the fine lateness was punished by a social norm among parents, but with the incentive provided by the fine it was no longer worth implementing a non-trivial social norm and consequently lateness increased.[18]

There are existing behavioral explanations for this phenomenon. A particularly well developed version can be found in Benabou and Tirole (2006): the principal (the experimenter) by means of the incentives provided affects what effort signals about the agent's identity. In the absence of incentives participants find the signaling value of the task sufficiently high (they are prosocial) and provide effort accordingly. When small incentives are introduced this signaling value is diminished as greater effort may now be attributed to greed. The participants then conclude the task is not worth much and so provide little effort.

Notice that the welfare implications of these two theories are quite different. In the behavioral theory increased lateness is unambiguously bad and the intervention with fines is a failure. With endogenous social norms this is not the case. As the fine increases from 0 to $\hat{\varphi}$ nothing changes including welfare. At the switch point $\hat{\varphi}$ increasing the fine unambiguously increases welfare. This follows from the fact that welfare is measured by group utility and the reason the group switches away from the existing social norm to the default social norm is because it increases group utility - the reduction in welfare from less output of the public good is compensated for by reduced monitoring costs. Here a drop in the output of the public good represents an unambiguous increase in welfare - it means the intervention was a success. As φ increases welfare continues to increase until the first best is obtained at the Pigouvian level of $\varphi = v$.

The welfare issue is of importance - for example in the use of incentives (or non-use) for promoting blood donations - see, for example, Meyer and Tripodi (2017).

[18]Our theory does not explain why when the fine was removed parents continued being late - however, the data after the fine was removed is very short in duration so we cannot say whether in a few more weeks or months lateness began to drop. In general we expect the frictions (and time) to agree to a non-trivial social norm to be greater than that needed to revert to the default.

4.3. Intervention in the Large and in the Small: Wages and Employment

Economists generally think of wages and employment being determined competitively in markets in which firms and workers are small players. The larger social groups to which these small players belong, by contrast, can have substantial monopoly or monopsony power. We see this explicitly in the case of trade unions - which indeed exercise monopsony power by enforcing social norms such as "do not work too hard" through the mechanism of peer monitoring and ostracism. In our setting "not working too hard" is a public good because it enables the group to exercise its monopsony power. Here we study large cartels: we accept that millions of farmers through peer monitoring and punishment can overcome the public goods problem[19] involved with lobbying - why cannot hundreds of firms overcome the public good problem of forming a cartel?

Consider the field experiment of Gneezy and List (2006) in which they paid some solicitors a fixed bonus above the market wage and others not. They discovered that initially those with the bonus increased their effort, but over the entire course of the experiment did about the same amount of work per unit of pay as those without the bonus. This is consistent with a social norm in which the wages per unit of effort are part of a social norm: solicitors do the amount of work per pay as called for by the social norm regardless of whether the money is paid as a piece rate or a lump sum.

In this context stickiness has potentially significant consequences. First, if wages or employment are determined by social norms then the presence of fixed costs of changing social norms will have a dynamic effect very similar to that of the menu cost model of Calvo (1983): changes will occur only when economic circumstances change enough to make it worthwhile to pay the fixed cost of changing the norm. Second, the Lucas critique may apply to empirical work studying policy interventions in markets.

A particular case involving interventions is the continuing controversy over the employment effect of the minimum wage. Consider studies by two labor economists, both John Bates Clarke medal winners: David Card and Kevin Murphy. Card and Krueger (1994) provide evidence that changes in the minimum wage have little effect on employment, while Deere, Murphy and Welch (1995) provide evidence that the minimum wage has a substantial effect on employment. These two studies use rather different data: Card and Krueger (1994) use a natural experiment comparing the effect of minimum wage change in one state against nearby states where the minimum wage did not change. Deere, Murphy and Welch (1995) examine the effect of a change to the federal minimum wage. If employment is determined by social norms then changes in a single state may represent a small intervention insufficient to change social norms. By contrast changes in the federal minimum wage may represent a large intervention sufficient to change social norms so there would be a substantial employment effect. This raises the issue of whether studies of the minimum wage might need to look more closely at the extent to which social networks and social norms play a role in determining employment.

[19] See, for example, Levine and Modica (2017)

Specifically, suppose that there is an input restricting cartel. Suppose that demand is linear and the units of output are scaled so that the cartel faces a market price of $\Pi = \Pi_0 - \theta$ where θ is per firm output. (Notice that θ is now a public bad from the cartel point of view.) Output is produced solely using labor. We suppose that that the cartel is small with respect to the labor market, that the competitive wage is ω_c and the minimum wage is $\omega_m > \omega_c$. A fraction of φ of members face the minimum wage and the remainder face the competitive wage. Initially $\varphi = 0$ so that all firms face the competitive wage.

The cartel establishes a quota θ_j for each firm depending on whether it faces the minimum or competitive wage $j = m, c$. It is made up of many firms - so cartel quotas cannot be enforced, for example, by common punishment such as threats of future cartel collapse,[20] rather social means are used to punish defectors. A firm that chooses to deviate from the social norm can produce up to capacity $\overline{\theta}$. We assume that this capacity is such that prior to the formation of a cartel at the competitive equilibrium capacity was exactly equal to demand: that is $\overline{\theta} = \Pi_0 - \omega_c$. We suppose that initially no firms are constrained by the minimum wage and that minimum wage changes are unlikely to happen so it is optimal to design a non-contingent social norm. Then a minimum wage is introduced for a fraction φ of firms. Notice the standard result that if there is no cartel and the industry is in competitive equilibrium the minimum wage will increase price and reduce output and employment.

For simplicity we assume that monitoring costs are relatively low $\mu(1, \pi, h) \leq 1/(\overline{\theta}(\Pi_0 - \omega_c))$. Our first theorem is proven in the web appendix

Theorem 5. *Define* $\alpha = 1/[2(U(\mu(\pi, \pi, h) + 1)]$. *Suppose that the minimum wage is not too large in the sense that* $\omega_m < \omega_c + \alpha(\Pi_0 - \omega_c)$. *Then there exists* $\underline{\varphi} > 0$ *such that for* $0 \leq \varphi \leq \underline{\varphi}$ *we have* $\hat{\theta}(\varphi) = \hat{\theta}(0)$. *Moreover (and regardless of the size of the minimum wage) there is an* $0 < \underline{F}$ *and* $\overline{\varphi} > 0$ *such that if* $F \leq \underline{F}$ *and* $1 \geq \varphi \geq \overline{\varphi}$ *we have* $\hat{\theta}(\varphi) < \hat{\theta}(0)$.

This result is similar to that of the previous section: small increases in φ do not change anything. The existing social norm remains near optimal so it is not worth paying a fixed cost to switch to something slightly better or to jump to the default social norm. In the previous example parents are constrained by the social norm not to be too late and introducing a small norm does not cause them to want to arrive earlier; here firms are constrained by the social norm not to produce less and a modest increase in the wage does not want to reduce their output. Notice on the other hand the second result that with small fixed costs a global change in the minimum wage $\varphi = 1$ leads to a definite drop in output.

Observe that a small intervention does achieve the desired purpose. Consumers are unaffected and welfare does not change, the consequence of increased minimum wage covered is a transfer from the cartel to workers. This is misleading with respect to a large intervention - here consumer and producer surplus as well as employment are all reduced. The only beneficiaries are those workers

[20] See Fudenberg, Levine and Pesendorfer (1998)for a discussion of why common punishment fails in the presence of many players.

who keep their jobs.

We should note that when the fixed cost is large there can be an even more dramatic effect.

Theorem 6. *If the minimum wage is large in the sense that* $\omega_m > \Pi_0$ *then there exists a* $\underline{\pi} > 0, \underline{h} > 0$, *a* $\overline{F} > 0$ *and a* $\overline{\varphi} > \underline{\varphi} > 0$ *such that for* $\pi \leq \underline{\pi}, h \leq \underline{h}$, $F \geq \overline{F}$ *and* $\underline{\varphi} < \varphi < \overline{\varphi}$ *we have* $\hat{\theta}(\varphi) > \hat{\theta}(0)$.

Here an intermediate increase in minimum wage coverage implies a competitive equilibrium in which price is determined by the firms facing a high minimum wage - and this generates substantial competitive rents for the industry. Rather than enforcing an expensive cartel the industry prefers simply to each go their own way and eschew monopoly profits for competitive rents. Notice that this will not be the case for very large φ as competitive rents at $\varphi = 1$ are zero.

5. Internalization

Sociologists and psychologists refer to the process by which an individual learns to adopt the norms of their group as their own as internalization. We model internalization of a social norm from the point of view of an individual as a tradeoff between a costly learning investment and a benefit from adhering to the norm. Consistent with our objective of endogenizing behavior that behavioral economics has taken to be exogenous we adopt a simple model of internalization. We assume that individuals can internalize any strategy - that, roughly speaking, they can invest in learning a rule of behavior. In the current stylized simple setting, learning a behavior rule may seem relatively easy - produce or not produce? tell the truth or not tell the truth? - although the correct ostracism probabilities are perhaps not so trivial - but real social norms deal with a much broader array of more complicated interactions. Indeed, social norms may encompass entire codes of conduct in the sense of Block and Levine (2016) or secret handshakes as in Robson (1990). We also refer to the literature on automata and the complexity of strategies (see for example Abreu and Rubinstein (1988) and the more recent literature on competition Gale and Sabourian (2005)) - a literature that implicitly or explicitly supposes that it is more difficult to implement complicated strategies than simple ones. Once a strategy has been chosen and invested in, our assumption is that the investor receives a benefit from adhering to it.[21] In other words the utility function changes so that doing the things you have learned to do well brings utility. This makes sense also from the perspective of the habit formation literature (see for example, Constantanides (1990), Campbell and Cochrane (1999) or Boldrin, Christiano and Fisher (2001)) - in a broader sense learning to do something often makes it more pleasurable. Just as the utility of fine wine increases with experience, so may the benefit of altruistic giving.

Internalization that is completely neutral between different strategies or based entirely on considerations of complexity does not change anything in a continuum game of the type we are studying

[21] If the cost of investment is positive it would not make sense to invest in a strategy if the only consequence was to feel guilt for not following it. However failure to receive a benefit is equivalent to a loss, so the loss of benefit may indeed be the same as guilt.

where commitment has no value. The key to our results is that it will be less costly to invest in the social norm than in any other strategy. There are two reasons this might be the case. First, there may be a network externality - it may be less costly to learn to use a strategy that everyone else is using than to invent a new strategy. This would be the case, for example, if the strategy is a choice of language. Lower cost for the social norm is similar to the preference for conformity studied by Benassy (1998) and Akerlof and Kranton (2000).

Second, the group may subsidize investment in the social norm. Public schooling is a particular example: nations invest in public schooling which in part teach social norms. Think of the Madrasas of the Taliban: pupils acquire valuable human capital in exchange for learning Taliban social norms. Indeed all public schools work this way - Bowles and Gintis (1976) documented the teaching of social norms in US schools and everyone who has read history post K-12 recognizes the substantial element of national myth taught in school. Indeed, if we examine the history of public education we observe that it originated in Scotland and that the valuable skill of high literacy was taught for religious reasons - to promote a social norm. Here we shall focus on this latter idea - that the group subsidizes investment in the social norm.

We turn next to the specific details of the model. The aim of the section is to characterize the cost minimizing social norm. As before the issue is whether production can be implemented and what are the ostracism probabilities needed to support it.

A Model of Investment in Social Norms . The mechanism design problem (and the induced game) we studied above effectively begins with the group choosing and announcing a pure strategy σ called the social norm. We now assume that after this announcement and before matching, production and monitoring takes place - in particular before one's identity as producer or monitor becomes known - individuals may choose to invest (or specialize) in a pure strategy s of their choice. We refer to this investment as *internalization* of the strategy s. We denote the level of investment by b and assume that the gross individual cost of this investment is $(1+\beta)b$ where $\beta > 0$. The choice of investment is known only to the investor.[22]

The consequence of investing in a strategy is that the investor gets utility from using it: specifically if s is chosen and the terminal node is consistent with s the investor receives a bonus equal to the level of investment $b \geq 0$ for following the prescribed strategy. Notice that since $\beta > 0$ no individual will ever wish to internalize a strategy. However, we allow also the possibility that the group subsidizes investment in the social norm σ. Specifically we assume the group sets a target investment level B and a subsidy level γ so that the net individual cost of investing in the social

[22]In the bargaining literature (see Schelling (1956), Muthoo (1996) and Dutta (2012)) commitment is assumed to be observable. That literature focuses on the strategic advantage of commitment when there are small numbers and an unobservable commitment is useless. Here we assume the commitment is unobservable: if it were observable there would be an additional channel of punishment - ostracism could be based on failing to invest in the social norm. We wish to keep the punishment channel the same as in the base model so we keep the commitment unobservable. The observable case (with noisy signals of investment) is similar to the model of codes of conduct studied in Block and Levine (2016).

norm is$(1+\beta-\gamma)\min\{b,B\}+(1+\beta)\max\{b-B,0\}$.[23]

Essential versus Inessential Indifference. When we solved the cost minimization problem with private information and without the possibility of internalization the solution involved two different types of indifference. The producer is indifferent between producing and not producing while the monitor is indifferent between reporting 0 and 1. The first indifference - that of the producer - is inessential in the sense that we could punish a bit more for a bad signal and the producer would strictly prefer to produce. The indifference of the monitor is essential in the sense that if the monitor is not indifferent between reporting 0 and 1 the monitor will not tell the truth.

One way to see that an indifference is essential is to perturb the model. Suppose that the monitor may choose whether or not to observe the signal and that there is a positive cost for observing the signal. In this case production cannot be implemented: if we make the monitor indifferent between $0, 1$ so willing to tell the truth, it is always better to report randomly and not pay the monitoring cost.

If the group chooses $B > 0$ together with a sufficient subsidy that the social norm is internalized all indifference is inessential. If implementation of production is possible and the group uses the cost minimizing incentives given in Theorem 1 then every group member strictly prefers to produce and strictly prefers to report the truth. Adding a small monitoring cost no longer changes things.

We should note also that implementation of production is always possible with internalization: if the group is willing to incur the cost of a very large B then a very large c may be overcome.

Optimal Internalization

To state our main theorem we distinguish between five different types of social norm and three levels of internalization cost and production cost. The precise definitions can be found in the web appendix along with the proof. The five different types of social norm by level of internalization are:

None: $B = 0$, only incentives are used and we are in the original case of Theorem 1.

Minimal: implementation of production is impossible without internalization and the level of internalization B is the least amount compatible with implementation of production. In particular in this case we always have $P = 1$. If $Q = 0$ we refer to this case as *Honesty with* $P = 1$.

Honesty: the least amount of internalization is used so that no incentives are needed for monitors. In particular in this case we always have $Q = 0$.

Complete: only internalization is used, $B = c$ and $P = Q = 0$.

[23] Providing the subsidy may itself pose a public good problem. We assume that individual contributions that are required to provide the subsidy are public information and that the channel for punishing those who fail to contribute are separate from the social interaction that takes place after the production decision and that the punishment is adequate to provide the necessary incentives. That is: for simplicity we focus on the case where there is no additional public goods problem from subsidizing investment in the social norm and focus on the public goods problem of producing.

Theorem. *It is always optimal for the group to choose $\gamma = \beta$. The type of social norm that minimizes the cost of implementing production depends on the cost of internalization and production as given in the following table:*

	low β	*medium* β	*high* β
low c	*Complete*	*Honesty*	*None*
medium c	*Complete*	*Honesty*	*Minimal*
high c	*Complete*	*Honesty* $P = 1$	*Honesty* $P = 1$

The message that emerges from the theorem is that internalization is more important for monitor incentives than for producer incentives. In particular: if the cost of internalization of the social norm (β) is low then it is best to completely internalize the social norm and dispense with incentives entirely. If the cost of internalization is medium then internalization should be used to induced truth-telling by monitors ($Q = 0$) but incentives used for producers. Finally, if the cost of internalization is high and production can be implemented without internalization (c low) we are in the original case studied in Theorem 1, while if it cannot be (c medium or high) then internalization must be used and will always be used at least to the level at which monitor incentives are not needed.

When social norms are internalized, it is not easy to distinguish - in the laboratory, for example - between social norms and intrinsic preferences. The issue is controversial as writers such as Fehr and Gachter (2000), Bowles et al (2003) and Roemer (2015) point to evidence that while incentives seem important for providing public goods, incentives seem less important for monitors. Moreover, without internalization, the incentives for monitors and for ostracism are weak - relying as they do on essential indifference. Social norms can provide incentives for monitors through repeated rounds of monitoring: a rule of "punish violators and if you fail to do so you are a violator yourself" can be found even in such places as written constitutions for prison gangs as documented in Skarbek (2014). Levine and Modica (2016) provide theoretical results in this direction. However, as the current theory shows, with internalization the need for multiple rounds of auditing is mitigated. Moreover, as we see, unless the cost of internalization is very high, incentives will neither be used or needed for monitors - which may explain why researchers such as Bowles et al (2003) and Roemer (2015) find evidence that they are uncommon.

Heterogeneous Matches

We now study a situation with two types of matches. Absent internalization the problem of each type of match can be solved separately but this need not be the case when investment in a social norm is possible. We must now confront an issue concerning investment in social norms and more broadly in strategies: to what extent is it possible to differentiate between different components of a strategy? If the social norm is to produce and to tell the truth, is the cost of teaching both equal to the sum of the cost of teaching each? In practice we imagine that there is some economy of scale in teaching both and adopt a model in which the social norm must be learned as a whole. Here we

take the extreme point of view: internalization is all or nothing - all components are internalized or none.

We are not going to investigate all possible situations, but rather give an illustrative example. Specifically we consider a setting where there are two types of matches: a fraction $1 > \varphi > 0$ of small stakes matches $\tau = S$ in which there is a single monitor per match and a fraction $1 - \varphi$ of high stakes matches $\tau = H$ in which there is public information - where monitors are constrained to tell the truth but remain anonymous. We return to the variable production level $\theta \geq 0$ with quadratic cost $c = \theta^2$ of section 3.2, so that in both cases the social value of production is $v\theta - \theta^2$. Hence a social norm is of the form θ_i, P_i, Q_i where $i = S, H$ and $Q_H = 0$ as we have assumed that $h = 0$ in the high stakes matches.[24] Small stakes are reflected by a constraint $\theta_S \leq \underline{\theta} > 0$.

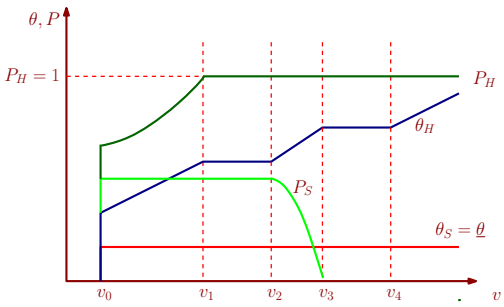
If investment costs are very low there would be complete internalization. Since in practice we generally see the use of punishments for violations of social norms this suggests the investment costs are not so low. Hence we treat internalization as an expensive backstop technology to be used only when needed. Unlike punishment which is limited if society is willing to bear the cost internalization is not. This is consistent with the fact that internalization can be used to get people to do things when even the largest punishment is inadequate. In other words, internalization works in extreme circumstances when punishment is not available - some people will intervene to save strangers at the risk of their own life in circumstances where if they walked away nobody would be the wiser. Moreover even the most horrific of historical punishments are inadequate to deter all crime. Indeed the most horrific punishments were reserved for political crimes carried out because of strongly (large B) internalized social norms of religion.

We observe also that in small stakes matches such as the laboratory we rarely see secondary punishment. As noted above this is often interpreted as an intrinsic behavioral desire for revenge - but it is also an optimal social norm in the "honesty is the best policy" scenario where monitor incentives come from internalization while producer incentives come from punishment. This leads us to focus on the case of intermediate β. In this setting *ex post* after the investment is made we use internalization first and only incentives as needed, so that we might mistakenly conclude that incentives are secondary and used "only as needed." By contrast during the choice of investment the opposite is true: we use punishment first and only internalization as needed. Notice how failing to think through the endogenous nature of internalization leads to a mistaken conclusion about the nature of social norms.

Theorem 7. *For small enough $\underline{\theta}$, intermediate β and intermediate F we have $Q_S = 0$ and the properties of the optimal θ_j, P_j as a function of v are given in the diagram below:*[25]

[24] That public information is equivalent to assuming $h = 0$ is shown in Web Appendix 1. That $Q = 0$ whenever $h = 0$ is shown in Web Appendix 3.

[25] For those who like equations the exact description appears in Web Appendix 3 along with the proof.



Moreover θ_H is decreasing in U for $v_0 \leq v \leq v_1$ and increasing in U for $v_2 \leq v \leq v_3$ while θ_S, P_S are constant in both ranges.

For small v the default social norm is used and there is no internalization. When the first critical threshold is crossed it becomes optimal to produce $\theta_S = \underline{\theta}$ the maximum possible in small matches. The assumption of intermediate β is chosen so that it is desirable to choose B large enough to internalize monitor incentives, that is so that $Q_S = 0$, but no larger. As v increases increased output θ_H is increased by increasing P_H while B remains fixed. Eventually when v is large enough P_H reaches 1. At this point further increases in θ_H are possible only using the costly backstop technology of internalization, that is, by increasing B. This causes a jump in the marginal cost of producing output θ_H in the high stakes matches and consequently θ_H remains constant and we are stuck on at a corner solution. Eventually v is large enough that it becomes desirable to use internalization B to increase output of θ_H in the high stakes matches. Increasing B brings two benefits: it increases θ_H and also lowers P_S. Eventually P_S reaches 0 so that increasing B brings only the benefit of increased θ_H and again for a range of v we are on a corner where nothing changes. Eventually when v is very large it becomes desirable to use internalization solely to increase θ_H in the high stakes matches.

The key point here is that increasing v has quite different effects on the use of punishment against producers in the two types of matches. Both jump up when the default norm is dropped, but in high stakes matches as v increases further punishment increases to 1 and sticks there: internalization is not used as a substitute for punishment. In low stakes matches as v increases further punishment remains constant then eventually declines: internalization substitutes for punishment. As the small stakes are most relevant to the lab and high stakes to the field there is a cautionary note here in thinking that the use of punishment in the lab tells us something about the field.

The difference between the range $v_0 \leq v \leq v_1$ where θ_H is increased using increased punishment and the range $v_2 \leq v \leq v_3$ where θ_H is increased using increased internalization has interesting consequences for the effect of U (the social cost of ostracism) as indicated in the Theorem. In the lower range increased U increases the cost of punishment used to provide incentives to produce the public good θ_H, hence it is optimal to produce less. In the upper range the output of θ_H is supported by internalization and the marginal cost β does not depend on U. However, the cost of punishing in low stakes matches P_S does and when this cost is raised the benefit of increasing B in lowering P_S is greater and so it becomes optimal to produce more of the public good θ_H.

Example. To better understand the role of U and how internalization works let us suppose that the high stakes matches (H) are opportunities to donate blood, while low stakes matches (S) are opportunities to give away small amounts of cash either to beggars in the street or participants in the laboratory. Notice that at some times and places (war) the social value of blood donations (v) might be quite high and in others it might be low.

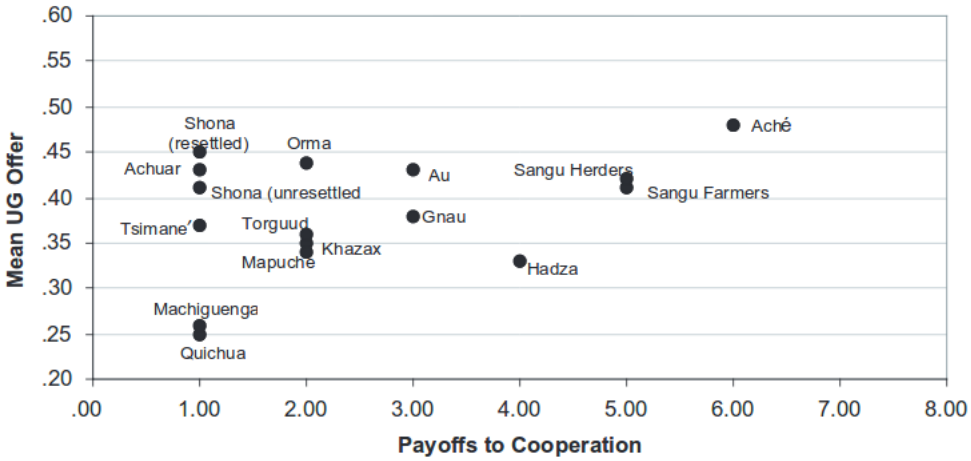
To give a concrete interpretation of U we use the earlier example of social interaction in which members both tell stories and listen to stories. Here $U = 1 + (N-1)\nu$ where N is the number of people at the meeting and ν is the benefit of listening to a story. As a comparative statics exercise consider either a policy or secular change that makes commercial entertainment such as television or the internet less costly. We would not expect this to reduce the social benefit of telling stories, but it might well reduce the social benefit ν of hearing stories as more and cheaper alternatives are available. In other words it might lower the social cost of ostracism U. Hence in the lower range $v_0 \leq v \leq v_1$ cheaper alternative entertainment could raise blood donations θ_H. This is also what a simple theory without internalization and heterogeneous matches would predict.

The upper range is rather different and highlights how internalization works. While in the range $v_2 \leq v \leq v_3$ the optimal internalization would lead to fewer blood donations once the investment in internalization B is made there is no reason to go back make a costly reinvestment in a lower value of B. Note the asymmetry: a large increase in U might well justify throwing away the existing investment and making a larger one. However, for a reduction in U the immediate effect would be nil. Never-the-less, over time, we would expect that there is turnover in any group: some members leave and newcomers arrive. It would make sense to train the newcomers to the lower socially optimal value of B. Hence over time we might expect a gradual reduction in B with a corresponding fall in blood donations - and an increase in the punishment P_S used in small stakes matches.

This result - that reduction in the cost of commercial entertainment might lead to an increase in blood donations when blood donations have relatively low value and a decrease when they have relatively high value - is not an obvious one. Moreover it contains a message about how to make good use of laboratory experiments. If we suspect that an decrease in blood donations has come about because of a decrease in internalization a second consequence is that P_S should increase and this we can test in the laboratory, for example, by observing rejection rates in ultimatum bargaining games.

One implication of Theorem 7 is that punishment in low stakes matches should have an inverted-U shape with respect to v. Is this true? While data about groups with different values of v 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):[26]

[26]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 v. In these societies the public good is largely the insurance resulting from sharing gains received through good luck. The vertical axis is the average offer made by the first mover in the ultimatum game: what fraction of money from the experimenter they choose to share with a partner. It can be interpreted as θ_S the extent of public good production in small stakes matches. As can be seen θ_S seems to jump up at the bottom then remains flat as our theory suggests.

What about punishment P_S? In the ultimatum game the "monitor" is the second mover who must make a costly decision whether or not to punish the first mover. In the laboratory there is no possibility of punishing the monitor for failing to punish: but in our theory this does not matter because monitor incentives in small stakes matches such as the laboratory are entirely internalized with $Q_S = 0$.

What corresponds to π the observational error in the laboratory? Not surely the inability of the second mover to correctly see what offer was made by the first mover. However, social norms specify not merely correct behavior, but rather correct behavior in response to circumstances (the type of match τ). Observational error about the type of match is possible as well as observational error about what happened in the match. An ultimatum bargaining experiment is, after all, an unusual event, and two different members of a group may well have different interpretations of how the social norm applies: this we can model as "observational error" π.

If we make the heroic assumption that the only difference between these different societies is v - and in particular all have the same value of π - then the rejection rate in ultimatum bargaining is a precise measure of P_S. Hence we examine the data on rejections across societies. We find for the two very low v, θ_S groups the Machigeuenga and Quichua there is only one rejection out of the 21 pairs in the Machigeuenga and none in the Quichua. The low value of θ_S and the lack of punishment is consistent with the idea that with very low v the default social norm is being used. On the other hand the theory predicts than when v is very large we should see a high degree of internalization so punishment should be unnecessary in small stakes matches: in fact in the highest v group the Ache

there were no rejections. There is also some casual evidence of a high degree of internalization in this society: "Successful hunters often leave their prey outside the camp to be discovered by others, carefully avoiding any hint of boastfulness." By contrast in the remaining intermediate societies where punishment should be used instead of internalization for producer incentives we find that the rejection rate is 12%. While the data is weak it is consistent with an inverted-U shaped punishment curve for small stakes matches as a function of v.

Our interpretation of the data seen through lens of a theory of endogenous social norms 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. Notice, however, that if we exclude the very lowest value of v, "Payoffs to Cooperation" equal to one, then sharing is relatively flat as v is increased. This is exactly what our theory predicts. Moreover they do not attempt to account for the inverted-U shaped punishment curve. 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 endogenous social norms provides a much more detailed, accurate, and sharper account of what to look for in the data than existing behavioral theories.

6. Conclusion

There is substantial evidence that informally enforced social norms are important, that they are endogenous, that they are sticky, and that they are internalized. We have introduced a simple model of collective decision making subject to incentive constraints that captures these key features. Our goal is to elucidate the implications for economic analysis. We do so through a series of examples and applications. Our first set of applications examines the consequence of the fact that social norms are not arbitrary but optimal. These helps us understand diverse phenomena: why doctors are habitually late but the working class faces penalties for not showing up on time and why improvements in privacy may have little consequence as punishments are optimally adjusted to reflect the lower likelihood of getting caught. Our second set of applications examines the consequences of the fact that social norms are sticky. We show how this can lead policy interventions such as taxes or subsides to have the opposite of the intended effect. Finally, we study internalization showing that while experimental studies may not directly translate into conclusions about the field or about policy appropriate modeling of endogenous social norms and internalization enable us to make useful inferences about the field or policy from experimental data.

We conclude by indicating how the ideas in this paper fit into the broader literature of behavioral economics and cultural norms: we propose indeed that the theory of endogenous social norms may form an important link between these ideas.

Social Norms or Psychology?. While writers such as Bowles et al (2003) and Roemer (2015) point to evolutionary reasons why punishment might be "hard-wired" and while we do not doubt that small children do not need to be taught to punish the theft of a toy, 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. The vast array of social norms we see across time and location indicate to us that most likely they are endogenous reflecting circumstances even if they do tap into intrinsic preferences for "revenge."

An instructive example of the endogeneity and adaptability of social norms is the custom of tipping service providers: this is commonplace in the US and UK, but rare, for example, in Italy. In Italy it works rather the other way around: not only is there no tipping but repeat customers get a discount - kind of a negative tip. In the US and UK there is a definite social sanction for not tipping. Other people at your table as well as the waiters may sneer at you - indeed you may be explicitly told not to return. We would argue that these are not just arbitrary customs, but rather are based on the need for incentives. With low waiter turn-over both within restaurants and within communities social norms among waiters can support good service and tipping is not needed - this is the situation in Italy. With high waiter turn-over and waiters not tied to the local community it is difficult for social norms to support good service, and so tipping is a needed incentive.

Social versus Cultural Norms. We do not see a valid distinction between social and cultural norms: rather the matter seems to us a matter of degree. Although cultural norms require a much larger investment and have a much greater value of commitment, they are part of the same theory as that of social norms. To us the key lies in the idea of investment in strategies and the subsidization by interested parties. Hence: it is easier to learn the language spoken by your parents in your home and parents explicitly teach social norms to their children taking on themselves part of the investment cost.

Here we study a single group. In a political economy setting with competing groups the subsidization of social norms may become strategic as groups compete to encourage individuals to adhere to their preferred social norm. Interest groups fight over school curriculum precisely because they want to promote particular social norms. In economics there has been a tendency to view schooling through the lenses of human capital acquisition - and we agree that schooling is not mere signaling but teaches valuable skills. We should recognize, however, that those skills are a subsidy for learning social norms and that this has an important strategic component.

This idea of the strategic choice of social norms is not new: the idea of social norms that may be acquired horizontally (from peers) or vertically (from parents) has been used by Bisin and Verdier (2001) and Bisin and Verdier (2005) among others to examine the evolution of institutions. The model they use of costly efforts by parents to influence the social norms of their children is compatible with the view here: we think our theory adds an extra dimension to their analysis by emphasizing the endogenous nature of the social norms that are promoted. Their analysis complements ours because it deals with the endogenous formation of groups, a topic which we ignore in this paper.

In this context we should mention as well the possibility of competing social norms. In practice people may belong to several groups. This may not matter to the extent that social norms are incomplete and deal only with behavior relevant to that group: for example, the social norm of

economists deals with how many papers one should referee, but not how often one should attend religious services, while religious social norms deal with the latter but not generally the former. On the other hand there can be competing social norms - for example a Catholic doctor who has a patient wanting an abortion. This raises a complex set of issues that we have studied in part in Dutta, Levine and Modica (2018).

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.

Andrews, Sally, David A. Ellis,Heather Shaw and Lukasz Piwek (2015): "Beyond Self-Report: Tools to Compare Estimated and Real-World Smartphone Use," *PLOSOne*.

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.

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

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

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

Calvo, G. A. (1983): "Staggered prices in a utility-maximizing framework," *Journal of monetary Economics* 12(3): 383-398.

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.

Card, D., and Krueger, A. B. (1994): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *The American Economic Review* 84(4): 772-793.

Civil Justice Research Group (2000): "Medical Malpractice by the Numbers."

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

Cramton, P., E. Budish and J. Shim (2015): "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response, " *Quarterly Journal of Economics*, 130: 1547–1621.

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

Deere, D., Murphy, K. M., and Welch, F. (1995): "Employment and the 1990-1991 minimum-wage hike," *The American Economic Review* 85(2): 232-237.

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.

Dye, R. A. (1985): "Costly contract contingencies," *International Economic Review* 26(1): 233-250.

Ellickson, Bryan, Birgit Grodal, Suzanne Scotchmer and William R. Zame: "Clubs and the Market", *Econometrica* 67: 1185-1217.

Fehr, E., and S. Gächter (2000): "Fairness and retaliation: The economics of reciprocity," *Journal of Economic Perspectives* 14: 159-181.

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

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

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

Edlin, A. S. and C. Shannon (1998): "Strict Monotonicity in Comparative Statics," *Journal of Economic Theory*, 81: 201-219.

Gneezy, U., and List, J. A. (2006): "Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments," *Econometrica* 74(5): 1365-1384.

Gneezy, U., and Rustichini, A. (2000): "A fine is a price," *The Journal of Legal Studies* 29(1): 1-17.

Gneezy, U., and Rustichini, A. (2000): "Pay enough or don't pay at all," *The Quarterly Journal of Economics* 115(3): 791-810.

Hart, O., and Moore, J. (1988). "Incomplete contracts and renegotiation," *Econometrica* 56(4): 755-785.

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.

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

Kinchen, K. S., Cooper, L. A., Levine, D., Wang, N. Y., and Powe, N. R. (2004): "Referral of patients to specialists: factors affecting choice of specialist by primary care physicians," *The Annals of Family Medicine* 2(3): 245-252.

Kraut, A. I. (1975): "Prediction of managerial success by peer and training-staff ratings," *Journal of Applied Psychology* 60: 14.

Langton, Lynn and Matthew Durose (2013): "Police Behavior during Traffic and Street Stops,Bureau of Justice Statistics.

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

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.

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

Sarsons, H. (2017): "Interpreting Signals: Evidence from Doctor Referrals," Working Paper.

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.*

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.

West, E., Barron, D. N., Dowsett, J., and Newton, J. N. (1999): "Hierarchies and cliques in the social networks of health care professionals: implications for the design of dissemination strategies," *Social science & medicine* 48(5): 633-646.

Williams, Joan C. (2016): "What So Many People Don't Get About the U.S. Working Class," *Harvard Business Review.*

Web Appendix 1: Cost Minimizing Social Norms

We state and prove here a more general version of Theorem 1. We study a large group of unit mass in which a fraction $1-\zeta$ of members are randomly chosen to be producers and a fraction ζ to be monitors.[27]. There are two stages. In the *production stage* each producer is matched with a production opportunity and has probability η of being monitored by $k \geq 1$ monitors and probability $1-\eta$ of not being monitored. When a match is monitored we refer to the producer and monitors in a match as *partners*. We do not require that all monitors be matched, so we assume that the mass of monitors in matches is no greater than the mass of monitors: $\eta(1-\zeta)k \leq \zeta$. Producers do not know if they are monitored.

Each producer may use a unit of effort $e \in \{0,1\}$ to produce a public good at a cost of ec where $c > 0$.The social value of the public good is V. Hence if all producers produce ($e = 1$) the aggregate (per capita) social value is $(1-\zeta)V$. The producer's choice e generates a noisy signal $z \in \{0,1\}$ where with probability π the signal is wrong $z \neq e$ and with probability $1-\pi$ the signal is correct $z = e$ and $\pi < 1-\pi$. This signal is observed by every monitor matched with the producer.[28] Each monitor then makes a report $x \in \{0,1\}$.

We distinguish three monitoring technologies: *public information, independent,* and *collusive.* In the case of public information monitors are constrained to tell the truth but remain anonymous. In the collusive case the monitors coordinate their reports on the incentive compatible plan that is most favorable for them. In the independent case we assume $k \geq 2$ so that it is possible to compare reports. We take the cost of having a partner being ostracized in the collusive case as h with $h = 0$ in the public information and independent cases.

Define

$$\mu(\lambda, \pi, h) = \frac{\lambda + (k-\lambda)h}{(1-2\pi)(1-h)(1+kh)}.$$

We first check elementary properties of this function.

Proposition 1. *The function is $\mu(\lambda, \pi, h)$ increasing in all arguments, $\mu \to \infty$ as $\pi \to 1/2$ or $h \to 1$, the isocost curves of $\mu(\pi, \pi, h)$ in h, π space are downward sloping, and the isocost curves of $\mu(\pi, \pi, h)$ in h, π space for fixed h get flatter as π increases.*

Proof. Show $\mu(\lambda, \pi, h)$ is increasing in all arguments.

For λ, π this is obvious. The derivative μ_h has the same sign as the derivative of

$$\frac{\lambda + (k-\lambda)h}{(1 + (k-1)h - h^2 k)}$$

[27]For technical details on how a continuum model like this works seeEllickson et al (1999).

[28]Of course it may be that monitors receive signals that are imperfectly correlated and it may be that the quality of the signal improves when there are more monitors: The implications of different types of monitoring is beyond the scope of this paper and has been discussed extensively in the literature on mechanism design - see for example Cremer and McLean (1988) and Rahman (2012).

which is

$$\frac{(k-\lambda)\left(1+(k-1)h-h^2k\right)-(\lambda+(k-\lambda)h)\left(k-1-2hk\right)}{\left(1+(k-1)h-h^2k\right)^2}.$$

The numerator of this expression is

$$\begin{aligned}&(k-\lambda)\left(1+(k-1)h-h^2k\right)-(\lambda+(k-\lambda)h)\left(k-1-2hk\right)\\ =&k(1-\lambda)+(k-\lambda)h^2k+2\lambda hk>0\end{aligned}$$

Examine the isocost curves

An isocost curve is written as $\mu(\pi,\pi,h)=D$. These are downward sloping since μ is increasing. Now solve for isocost curve

$$\pi+(k-\pi)h-D(1-2\pi)\left(1+(k-1)h-h^2k\right)=0$$

Then find the slope of the isocost curve

$$\begin{aligned}\frac{d\pi}{dh}&=-\frac{k-\pi-D(1-2\pi)((k-1)-2kh)}{1-h+2D\left(1+(k-1)h-h^2k\right)}\\ &=-\frac{k-\pi-\frac{\pi+(k-\pi)h}{(1-2\pi)(1+(k-1)h-h^2k)}(1-2\pi)((k-1)-2kh)}{1-h+2\frac{\pi+(k-\pi)h}{(1-2\pi)(1+(k-1)h-h^2k)}\left(1+(k-1)h-h^2k\right)}\end{aligned}$$

Some elementary algebra shows that

$$\frac{d\pi}{dh}=-\left(1-2\pi\right)\frac{\left(k+k(k-1)h-h^2k^2-kh((k-1)-2kh)\right)-k(1-h)^2\pi}{\left(1+(k-1)h-h^2k\right)\left(1-h+2kh\right)}$$

and since the product of two positive functions decreasing in π is decreasing in π the derivative is increasing in π. □

Proposition 2. *As ostracism is socially costly we should ostracize only as needed for incentive compatibility: in particular if all reports on the producer are good there should be no ostracism; if all monitors file a bad report there should be no ostracism. Define $P=(\mu(1,\pi,h)/\eta)c$. Implementation is possible if $P\leq 1$ and in the public information and collusive case only then. In this case P is the probability a producer with all bad reports is ostracized and Q, the probability that a collusive monitor filing a good report is ostracized, is given by*

$$Q=\frac{h}{1+h(k-1)}P$$

and $C=\left[U\mu(\pi,\pi,h)+1\right]c$. When $P>1$ implementation may be possible in the independent case.

Proof. Let $p(x)$ be the probability the producer is ostracized when all monitors report x and let $q(x)$ be the probability that the monitors are ostracized when all monitors report x.

The on equilibrium path probability of ostracism per capita is the probability a randomly drawn

member is ostracized. A fraction $\eta(1-\zeta)$ are producers who were monitored and a fraction $\eta(1-\zeta)k$ are monitors so this probability is $\eta(1-\zeta)\pi(p(0)+kq(0))+\eta(1-\zeta)(1-\pi)(p(1)+kq(1))$ so the expected number ostracized per match is

$$L(s)=\eta\pi(p(0)+kq(0))+\eta(1-\pi)(p(1)+kq(1)).$$

When monitors tell the truth a producer who produces receives utility

$$-c-\eta\left(\pi(p(0)+khq(0))+(1-\pi)(p(1)+khq(1)\right)$$

and for not producing receives

$$-\eta\left((1-\pi)(p(0)+khq(0))+\pi(p(1)+khq(1)\right)$$

so the incentive constraint is

$$p(0)-p(1)+kh[q(0)-q(1)]\geq\frac{c}{\eta(1-2\pi)}.$$

In the case of public information monitors cannot be ostracized so $q(0)=q(1)=0$ and it follows that $p(1)=0$. Setting $P=p(0)$ we see that the incentive constraint should hold with equality and this can be written as

$$P=\frac{c}{\eta(1-2\pi)}.$$

Hence $P=(\mu(1,\pi,h)/\eta)c$ with $h=0$.

In the case of independent monitors disagreement can be punished by ostracizing the entire match so telling the truth is incentive compatible for any $q(0),q(1)$. From the incentive constraint of the producer we see then that we should take $p(1)=q(1)=0$. We see also that increasing $P=p(0)$ damages the objective by $\eta\pi$ per unit of improvement in the incentive constraint, while increasing $q(0)$ damages the objective by $\eta\pi/h$ per unit of improvement in the incentive constraint. Hence as long as $P\leq 1$ we should take $q(0)=0$ so that this case is the same as public information.

In the collusive case monitors must be indifferent between all reporting good and all reporting bad. Reporting all bad results in$-hp(0)-(1+h(k-1))q(0)$ and reporting all good results in $-hp(1)-(1+h(k-1))q(1)$, so the incentive constraint is

$$h(p(0)-p(1))=(1+h(k-1))(q(1)-q(0)).$$

Hence the producer constraint can be written either as

$$(p(0)-p(1))\left(\frac{(1-h)(1+kh)}{(1+h(k-1))}\right)\geq\frac{c}{\eta(1-2\pi)}$$

implying that $p(1) = 0$ or as

$$\left(\frac{(1-h)(1+kh)}{h}\right)[q(1) - q(0)] \geq \frac{c}{\eta(1-2\pi)}$$

implying that $q(0) = 0$.

Setting $P = p(0)$ and $Q = q(1)$ we see also that the incentive constraint should hold with equality so that

$$P\left(\frac{(1-h)(1+kh)}{(1+h(k-1))}\right) = \frac{c}{\eta(1-2\pi)}.$$

Hence $P = (\mu(1, \pi, h)/\eta)c$.

Also plugging in from above, the incentive constraint for the collusive monitors is

$$Q = \frac{h}{1+h(k-1)}P.$$

Turning to cost, since $C(s) = L(s)U + c$ and $L(s) = \eta(\pi P + k(1-\pi)Q)$ we have

$$\begin{aligned} C &= \eta U\left(\pi P + k(1-\pi)Q\right) + c \\ &= \eta U\left(\pi + k(1-\pi)\frac{h}{1+h(k-1)}\right)\frac{1}{\eta(1-2\pi)}\frac{1+h(k-1)}{(1-h)(1+kh)}c + c \\ &= \left[U\frac{\pi\left(1+h(k-1)\right) + (1-\pi)hk}{(1-2\pi)(1-h)(1+kh)} + 1\right]c \end{aligned}$$

The numerator is

$$\begin{aligned} \pi(1+h(k-1)) + kh(1-\pi) &= \pi + \pi hk - \pi h + kh - \pi kh \\ &= \pi - \pi h + kh = \pi + (k-\pi)h \end{aligned}$$

so indeed $C = [U\mu(\pi, \pi, h) + 1]\, c$.

Notice that if $P > 1$, then there is no way to satisfy the producer constraint, while simultaneously making the monitors indifferent since $p(0) - p(1)$ must be less than 1. □

With $P > 1$ implementability may be possible with independent monitors. The reason is that we can create additional incentive for the producer by ostracizing the monitors when they truthfully file bad reports. As shown in the proof this is more costly than ostracizing the producer so only useful if $P > 1$ so that ostracizing the producer does not provide adequate incentive for production. Ostracizing the monitors for the crime of the producer is an example of collective punishment: for example the punishment of the family members of a terrorist. As our applications where collusion among monitors is not feasible involve public information, we do not present details of the independent monitor case.

Web Appendix 2: Comparative Statics

Theorem (*Theorem 2 in the text*). *Suppose that* $0 < f_S(h), f_P(h) < \pi/2$ *are accuracy functions and let* C_S, C_P *denote the corresponding implementation costs. If* S *has a noisier signal* $f_S(h) > f_P(h)$ *then* $C_S > C_P$. *If in addition* S *has greater signal sensitivity* $|f_S'(h)| > |f_P'(h)|$ *and and* c *is sufficiently small that there exist unconstrained cost minimizers* $\hat{h}_j$ *that satisfy the constraint* $P_j \leq 1$ *then for any cost minimizers* $\hat{h}_S, \hat{h}_P$ *either* $\hat{h}_P = 0$ *or* $\hat{h}_S > \hat{h}_P$.

Proof. The first result follows because the fact that f_S lies entirely above f_P makes the optimum $\hat{\pi}_S = f_S(\hat{h}_S)$ strictly feasible for f_P. This implies P must have strictly lower minimum cost.

Second note that it is indeed the case that for c sufficiently small the cost minimizers satisfy the constraint. The key point is that the unconstrained optimal $\hat{h}_j$ do not depend on c hence choosing an unconstrained optimal $\hat{h}_S, \hat{h}_P$ we see the when c is small enough both satisfy the constraint.

For the final result consider first the unconstrained problem. Let $\xi(h, \pi)$ denote the slope of the isocost curve of $\mu(\pi, \pi, h)$ through h, π. Since $\xi(h, \pi) = -\mu_h/(\mu_\lambda + \mu_\pi)$ and $d\mu(f(h), f(h), h)/dh = (\mu_\lambda + \mu_\pi)f' + \mu_h$ we have

$$\frac{1}{\mu_h(f_j(h), f_j(h), h)} \frac{d\mu(f_j(h), f_j(h), h)}{dh} = -f_j'(h)/\xi(h, f_j(h)) + 1$$

By assumption $|f_S'(h)| > |f_P'(h)|$ and from Proposition 1 the isocost curves of $\mu(\pi, \pi, h)$ in h, π space for fixed h get flatter as π increases, and again by assumption $f_S(h) > f_P(h)$. Hence since $-f_j'(h)/\xi(h, f_j(h))$ is negative it is lower at S than at P. The result now follows from Edlin and Shannon (1998) Theorem 1: while they assume that

$$\frac{d\mu(f_j(h), f_j(h), h)}{dh}$$

is strictly lower since μ_h is strictly positive their proof goes through with the weaker condition.[29] Since this is true for unconstrained solutions and unconstrained solutions are assumed to exist the set of constrained solutions is a subset of the set of unconstrained solutions so the result holds also for the constrained solutions. □

We give the next theorem for the case in which there is only a probability η that a match is monitored (but continue to assume $k = 1$).

Theorem (*Theorem 3 in the text*). *With variable quality and quadratic cost the optimal social norm is*

$$\widehat{\theta} = \min\left\{\sqrt{\eta/\mu(1, \pi, h)}, v/(2(1 + U\mu(\pi, \pi, h)))\right\}$$

[29] In fact in their case they need the derivative strictly higher since they are dealing with maximization rather than minimization.

Proof. By Theorem 1 we know that for values of θ such that $\eta < \theta^2\mu(1,\pi,h)$ production is not incentive compatible. For values of θ such that $\eta \geq \theta^2\mu(1,\pi,h)$, production is incentive compatible. In this case, the optimal social norm would maximize $v\theta - \theta^2 - U\mu(\pi,\pi,h)\theta^2$. This function is increasing in θ (and positive), for sufficiently small θ. Therefore the optimal social norm would be the smaller of $v/(2(1+U\mu(\pi,\pi,h)))$, the interior maximum and the upper bound, $\sqrt{\eta\,(1-2\pi)\,(1-h^2)}$. □

Fines

Theorem (*Theorem 4 in the text*). *There exists $\overline{F} > 0$ such that for every $F \geq \overline{F}$ there is a $0 < \hat{\varphi} < v$ such that for $0 \leq \varphi < \hat{\varphi}$ the existing social norm is optimal and output is $\hat{\theta}(\varphi) = \hat{\theta}(0)$ and for $\varphi > \hat{\varphi}$ the default social norm is optimal and $\hat{\theta}(\varphi)$ is continuous, increasing and satisfies $\hat{\theta}(\hat{\varphi}) < \hat{\theta}(0)$ and $\hat{\theta}(v) > \hat{\theta}(0)$.*

Proof. By our notation, the optimum with respect to the existing social norm (with no fines, $\varphi = 0$) is given by $\hat{\theta}(0)$. For sufficiently large F it will never be optimal to introduce any other social norm besides the default. As φ increases from 0 we must check that the existing social norm continues to implement production. Notice that under the existing social norm the punishment probability P does not change, but it may no longer be adequate to give the producer reason to produce: this depends on how utility from deviating and sticking change with φ. With $\varphi > 0$ the producer deviates not to zero but to $\theta = \varphi/2$ where individual cost $\theta^2 + \varphi(\theta^* - \theta)$ becomes $-\varphi^2/4 + \varphi\theta^*$ so the fine increases the utility from the optimal deviation by $\varphi^2/4 - \varphi\theta^*$; on the other hand sticking to $\hat{\theta}(0)$ has the extra cost given by the fine $\varphi(\theta^* - \hat{\theta}(0))$ so utility from sticking increases by $\varphi\hat{\theta}(0) - \varphi\theta^*$. Hence the existing social norm continues to implement production as long as $\varphi < 4\hat{\theta}(0)$. Increasing φ in this range and keeping with the existing social norm keeps group welfare fixed. On the other hand group welfare from the default social norm is $\varphi(2v - \varphi)/4$, increasing in φ (since $\varphi \leq v$). Let $\hat{\varphi}$ be the point at which the two welfares are equal.

We observe: $\hat{\varphi} < 2\hat{\theta}(0)$ (in particular this occurs in the range where the existing social norm is implementing production). To see this, observe that at $\varphi = 2\hat{\theta}(0)$ output under the default social norm $\varphi/2 = \hat{\theta}(0)$. Since the default social norm achieves the same output of the public good at zero monitoring cost, it is strictly preferred to the existing social norm. Since the default social norm welfare is increasing in φ, and is 0 at $\varphi = 0$, it must be that $\hat{\varphi} < 2\hat{\theta}(0)$. Notice the implication that at $\hat{\varphi}$ output drops from $\hat{\theta}(0)$ to $\hat{\varphi}/2$.

Finally, since the marginal cost of monitoring is positive, at the existing social norm it is $\hat{\theta}(0) < \theta^* = v/2$. Then from $\hat{\varphi} < 2\hat{\theta}(0) < v$ we get that $\hat{\theta}(v) = v/2 > \hat{\theta}(0)$. □

Minimum Wage

Theorem (*Theorem 5 in the text*). *Define $\alpha = 1/[2(U(\mu(\pi,\pi,h)+1)]$. Suppose that the minimum wage is not too large in the sense that $\omega_m < \omega_c + \alpha(\Pi_0 - \omega_c)$. Then there exists $\underline{\varphi} > 0$ such that for $0 \leq \varphi \leq \underline{\varphi}$ we have $\hat{\theta}(\varphi) = \hat{\theta}(0)$, a constant. Moreover (and regardless of the size of the minimum wage) there is an $0 < \underline{F}$ and $\overline{\varphi} > 0$ such that if $F \leq \underline{F}$ and $1 \geq \varphi \geq \overline{\varphi}$ we have $\hat{\theta}(\varphi) < \hat{\theta}(0)$.*

Proof. We examine first the optimal cartel where all firms are identical, that is, $\varphi \in \{0,1\}$. If the quota is θ_j the price cost margin is $M = \Pi_0 - \theta_j - \omega_j$. Then incentive for a firm to violate the social norm and produce $\overline{\theta}$ is $M(\overline{\theta} - \theta_j)$. Let $\mu = \mu(\pi, \pi, h)$ and $\mu_1 = \mu(1, \pi, h)/\eta > \mu$. The feasibility condition for implementing a quota is $\mu_1 M(\overline{\theta} - \theta_j) \leq 1$, or $\mu_1(\Pi_0 - \omega_j - \theta_j)(\overline{\theta} - \theta_j) \leq 1$. At the competitive equilibrium $\theta_j = \Pi_0 - \omega_j$ this is certainly satisfied; the LHS is decreasing in θ_j and by assumption the condition is satisfied at $\theta_j = 0$ so the constraint does not bind. Consequently cartel profits are given by $(\Pi_0 - \omega_j - \theta_j)\theta_j - U\mu M(\overline{\theta} - \theta_j)$, which can be written as

$$(U\mu + 1)(\Pi_0 - \omega_j - \theta_j)\theta_j - U\mu(\Pi_0 - \omega_j - \theta_j)\overline{\theta}.$$

The profit derivative is

$$(U\mu + 1)(\Pi_0 - \omega_j - 2\theta_j) + U\mu\overline{\theta}$$

and this is easily seen to be decreasing. Evaluated at the competitive equilibrium it is $(U\mu+1)(\Pi_0 - \omega_j) - U\mu\overline{\theta} - 2\overline{\theta}$. When $j = c$, that is the initial situation, we have $\Pi_0 - \omega_c = \overline{\theta}$ so this is negative, so it is optimal to form a cartel and the optimal quota is

$$\theta_c = \frac{1}{2}\left(\Pi_0 - \omega_c + \frac{U\mu}{U\mu + 1}\overline{\theta}\right) = (1 - \alpha)\,\overline{\theta}.$$

We return to the case $j = m$ subsequently.

Since utility from the existing social norm and the optimal social norm are both continuous for small enough φ it is optimal not to pay the fixed cost and maintain the existing social norm θ_c. We need to check that high cost firms still wish to produce, however. If all firms produce to quota the price is

$$\begin{aligned}\Pi &= \Pi_0 - \overline{\theta}\,(1 - \alpha) = \Pi_0 - (\Pi_0 - \omega_c)\,(1 - \alpha) \\ &= \omega_c + \alpha(\Pi_0 - \omega_c) > \omega_m\end{aligned}$$

so all firms indeed want to produce. So industry output does not change for small φ. This proves the first result, where $\hat{\theta}(0) = \theta_c$.

Now we turn to $j = m$. The profit derivative at the competitive equilibrium $\theta = \Pi_0 - \omega_m$ is $-(U\mu + 1)(\Pi_0 - \omega_m) + U\mu\overline{\theta}$. This is positive if and only if

$$\frac{1}{2}\left(\Pi_0 - \omega_m + \frac{U\mu}{U\mu + 1}\overline{\theta}\right) > \Pi_0 - \omega_m$$

so that the optimal cartel is

$$\theta_m = \min\left\{\Pi_0 - \omega_m, \frac{1}{2}\left(\Pi_0 - \omega_m + \frac{U\mu}{U\mu + 1}\overline{\theta}\right)\right\}$$

from which it follows that $\theta_m < \theta_c$. Since the optimal social norm for $\varphi = 1$ is strictly better than maintaining the existing social norm θ_c, for F sufficiently small and large enough φ it will be

adopted and output will fall. □

Theorem (*Theorem 6 in the text*). *If the minimum wage is large in the sense that* $\omega_m > \Pi_0$ *then there exists a* $\underline{\pi} > 0, \underline{h} > 0$, *a* $\overline{F} > 0$ *and a* $\overline{\varphi} > \underline{\varphi} > 0$ *such that for* $\pi \leq \underline{\pi}, h \leq \underline{h}$, $F \geq \overline{F}$ *and* $\underline{\varphi} < \varphi < \overline{\varphi}$ *we have* $\hat{\theta}(\varphi) > \hat{\theta}(0)$.

Proof. With $\omega_m > \Pi_0$ high cost firms are always priced out of the market, so we can ignore them. With a high enough F the only choice is between the existing social norm (that is θ_c restricted to the fraction $1-\varphi$ of producing firms) and dropping the social norm entirely in favor of the default social norm, the competitive equilibrium. We prove the result holds at $\pi, h = 0$ with $\underline{\varphi} = 1/3$, $\overline{\varphi} = 1/2$; the general result then follows from the fact that θ_c and cartel profits are continuous in π, h.

When $\pi, h = 0$ there is no monitoring cost and so the cartel solution with $\varphi = 0$ is the monopoly solution $\theta_c = \overline{\theta}/2 = \hat{\theta}(0)$. For $\varphi > 0$ high cost firms are priced out and low cost firms produce to quota giving cartel output of $(1-\varphi)\overline{\theta}/2$ and cartel profits are $(1/4)(1+\varphi)(1-\varphi)\overline{\theta}^2$. The competitive output is $(1-\varphi)\overline{\theta}$ and the corresponding competitive rents are $\varphi(1-\varphi)\overline{\theta}^2$. We see immediately that form $\varphi < 1/2$ we have competitive output greater than $\hat{\theta}(0)$. Moreover we see that for $\varphi > 1/3$ competitive rents are greater than cartel profits giving $\underline{\varphi} = 1/3$. □

Web Appendix 3: Internalization

We first prove theorem 5:

Theorem (Theorem 5 in the text). *It is always optimal for the group to choose* $\gamma = \beta$. *The type of social norm that minimizes the cost of implementing production depends on the cost of internalization and production as given in the following table:*

	low β	*medium* β	*high* β
low c	*Complete*	*Honesty*	*None*
medium c	*Complete*	*Honesty*	*Minimal*
high c	*Complete*	*Honesty* $P = 1$	*Honesty* $P = 1$

We start by giving the details of the different norms appearing above in the table below. We first define the cost thresholds:

$$\underline{c} = (1 - 2\pi)\,(1 - h^2), \qquad \overline{c} = h + (1 - 2\pi) > \underline{c}.$$

In particular c low means $c < \underline{c}$, medium between $\underline{c}$ and $\overline{c}$ and high above $\overline{c}$. The two relevant thresholds for β will be denoted by $\underline{\beta}, \overline{\beta}$ and will be specified in the course of the proof. In the ranges where the norms appear in the table above the values in the one below are in the required ranges ($B \geq 0, 0 \leq P, Q \leq 1$).

Norm type	B	P	Q
None	0	$c/\underline{c}$	$hc/\underline{c}$
Minimal	$(c-\underline{c})/[1+h(1-2\pi)]$	1	$h-B$
Honesty	$hc/\overline{c}$	$c/\overline{c}$	0
Honesty $P=1$	$c-(1-2\pi)$	1	0
Complete	c	0	0

We start with a couple of preliminary observations.

Lemma 1. *$\gamma = \beta$ is optimal.*

Proof. Indeed if $\gamma < \beta$ then the individual would invest $b = 0$ not B; for $\gamma \geq \beta$ we have the individual would set $b = B$ and the net investment cost to the group per capita is B times $\gamma - [1-(1+\beta-\gamma)] = \beta$. □

From this it follows that the overall cost of implementation for the group, to be minimized, is

$$C = c + U\left[\pi P + (1-\pi)Q\right] + \beta B.$$

As to the constraints, the producer constraint is essentially the one derived in Appendix 1 modulo the addition of the benefit B from production, and it becomes $\eta(1-2\pi)(P-hQ) \geq c-B$. The monitor must want to tell the truth whatever signal he gets. If the signal is bad truth-telling has utility $-hP+B$ while lying yields $-Q$ so the constraint is $-hP+B \geq -Q$ or $Q \geq hP-B$; if the signal is good truth-telling yields $-Q+B$, lying $-hP$ so the relevant constraint is $-Q+B \geq -hP$ or $Q \leq B+hP$. We next show that the latter is not binding at the optimum - the relevant problem is to induce the monitor not to let the producer off the hook when the signal is bad.

Lemma 2. *The constraint $Q \leq B+hP$ is not binding at the optimum.*

Proof. Suppose by contradiction that the constraint binds at the optimum, $Q = B+hP$. Consider the ostracism probabilities $P' = P-\epsilon$ and $Q' = Q - \frac{\epsilon}{h}$. For sufficiently small ϵ P', Q' continues to satisfy both the producer and monitor incentive constraints only now with $Q' < B+hP'$. But this has a lower cost of implementation than P, Q, a contradiction. □

Thus the problem to be solved is

$$\min C \quad s.t.\ Q \geq hP-B,\ (1-2\pi)(P-hQ) \geq c-B,\ 0 \leq P,Q \leq 1.$$

In terms of Q the incentive constraints can be combined into

$$hP-B \leq Q \leq \frac{1}{h}\left(P - \frac{c-B}{1-2\pi}\right) \quad 0 \leq P,Q \leq 1$$

So feasibility requires that P, Q must be in the cone in the positive unit square of the (P,Q) plane spanned by the two lines $hP-B$ and $\frac{1}{h}\left(P-\frac{c-B}{1-2\pi}\right)$ on the right of their intersection. The

intersection of the two lines can be computed as

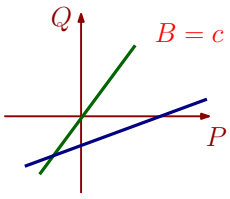
$$P = \frac{c - [1 + h(1 - 2\pi)]B}{\underline{c}} \qquad Q = \frac{hc - \overline{c}B}{\underline{c}}$$

and note that for given c it is determined by B, and that as B increases it shifts down and to the left. From the fact that the cost is increasing in B, P and Q one easily deduces that

Lemma 3. *At the optimum $B \leq c$.*

Proof. The figure below, which illustrates the case $B = c$, makes it clear that the choice $(P, Q) = (0, 0)$ is feasible for any $B \geq c$, any $B > c$ just gives higher cost than $B = c$. □

Figure 6.1: *Complete* case



So we concentrate on $B \leq c$ in the sequel. The next observation is that at the optimum the constraint $Q \leq \frac{1}{h}\left(P - \frac{c-B}{1-2\pi}\right)$ cannot be slack, in other words the solution must lie on the steeper line:

Lemma 4. *At the optimum $Q = \frac{1}{h}\left(P - \frac{c-B}{1-2\pi}\right)$.*

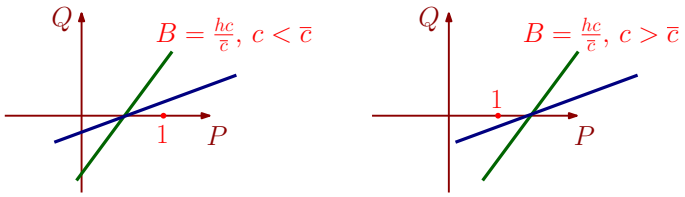
Proof. If $P > hQ + \frac{c-B}{1-2\pi}$ one can hold Q fixed and lower P thus reducing cost. □

At this point the geometry of the problem makes it easy to find its solution. The key simple insight is that since the cost is increasing in B, P and Q, for any choice of B you want to move down and to the left along the steep line as much as you can (for this lowers the cost). If the choice is $B = c$ as illustrated above you can move down to $(P, Q) = (0, 0)$. And clearly if the choice of B places the intersection in the unit square $0 \leq P, Q \leq 1$ then the optimal P, Q is at the intersection. The rest of the proof consists of repeated applications of this idea, plus the fact that the implied choices of P and Q given B make the cost function piecewise linear in B, hence the choice of B becomes trivial.

The various cases in the statement emerge by observing that as B goes down from c, at some point the intersection hits the horizontal axis, where $Q = 0$. This occurs when $B = hc/\overline{c}$ (smaller than c since $\overline{c} > h$). At that point it is always $P \geq 0$ - because $hc/\overline{c} \leq c/[1+h(1-2\pi)]$ which follows from $h \leq 1$ - but not always $P \leq 1$: from $Q = hP - B$ we get $P = B/h = c/\overline{c}$ so $P \leq 1 \iff c \leq \overline{c}$. So we have the two cases in the picture below:

At this point we can start spelling out the solution.

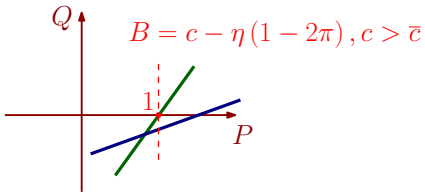
Figure 6.2: On the left panel it is *Honesty*



Lemma 5. *Suppose $c > \overline{c}$. Then if $\beta < \frac{\pi U}{1-2\pi} \equiv \underline{\beta}$ optimal $B = c$ and $P = Q = 0$ (the* Complete *case); if $\beta > \underline{\beta}$ optimal $B = c - (1 - 2\pi)$ and $P = 1, Q = 0$ (the* Honesty $P = 1$ *case).*

Proof. With $c > \overline{c}$ the lowest B for which the feasible cone intersects the unit square is where the steep line has $Q = 0$ and $P = 1$; this is for $B = c - \eta(1 - 2\pi) > 0$:

Figure 6.3: This is *Honesty* $P = 1$



So if $c > \overline{c}$ all possible choices of B from $B = c$ down to $B = c - (1 - 2\pi)$ have $Q = 0$ and $P = \frac{c-B}{1-2\pi}$; therefore the cost is

$$\begin{aligned} C &= c + U\pi\frac{c - B}{1 - 2\pi} + \beta B \\ &= c\left[1 + \frac{\pi U}{1 - 2\pi}\right] + \left[\beta - \frac{U\pi}{1 - 2\pi}\right] B \end{aligned}$$

whence if $\beta < \frac{\pi U}{1-2\pi} \equiv \underline{\beta}$ we set $B = c$ (with $P = Q = 0$), otherwise $B = c - (1 - 2\pi)$ (with $P = 1, Q = 0$). In other words if β is low the cost minimizing norm is *Complete*, otherwise it is *Honesty* $P = 1$. □

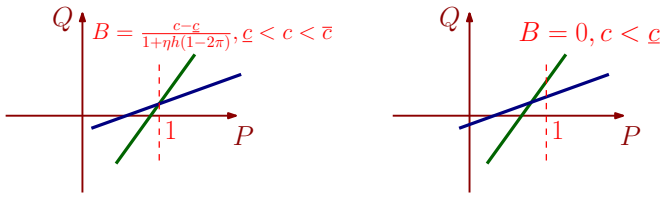
In the case $c < \overline{c}$ as we have seen the intersection with $Q = 0$ has $P < 1$. Then the lowest feasible B is the value where the intersection has $P = 1$ that is $B = (c - \underline{c})/[1 + h(1 - 2\pi)]$ if this is positive that is if $c > \underline{c}$ (as in the left panel of the picture below), or zero if this is negative, in which case $P = c/\underline{c}$ (and $Q = hc/\underline{c}$) (see the right panel):

We will next see that if β is low or medium it is not the lower bound of B which is optimal:

Lemma 6. *If $c < \overline{c}$ and $\beta < \underline{\beta}$ the optimal norm is* Complete.

Proof. If $c < \overline{c}$ we can lower B from c to $hc/\overline{c}$ with $P = \frac{c-B}{1-2\pi}$ while keeping $Q = 0$, then further down at the cost of raising Q besides P (because from then on the optimal choice is the intersection). With $P = \frac{c-B}{1-2\pi}$ and $Q = 0$ the cost is the one we saw above so if $\beta < \underline{\beta}$ setting $B = c$, which is the

Figure 6.4: On the left we have *Minimal*; on the right it is *None*



Complete case, is better than any other choice of B from the interval $[hc/\overline{c}, c]$. What about setting B lower than $hc/\overline{c}$? Here the optimal choice is the intersection where the cost is

$$\begin{aligned} C &= c + U\left[\pi P + (1-\pi)Q\right] + \beta B \\ &= c + U\left[\frac{\pi c - \pi[1+h(1-2\pi)]B}{\underline{c}} + \frac{(1-\pi)hc - (1-\pi)\overline{c}B}{\underline{c}}\right] + \beta B \\ &= c\left[1 + U\frac{\pi + (1-\pi)h}{\underline{c}}\right] + \left[\beta - U\frac{\pi[1+h(1-2\pi)] + (1-\pi)\overline{c}}{\underline{c}}\right] B \end{aligned}$$

which is decreasing in B if $\beta < \overline{\beta} \equiv U\frac{\pi[1+h(1-2\pi)]+(1-\pi)\overline{c}}{\underline{c}}$. This is indeed the case since $\beta < \underline{\beta} < \overline{\beta}$. Therefore setting B less than $hc/\overline{c}$ does even worse than setting it equal to $hc/\overline{c}$. So B would optimally be set equal to c, the Complete norm. □

Lemma 7. *If $c < \overline{c}$ and $\underline{\beta} < \beta < \overline{\beta} = U\frac{\pi[1+h(1-2\pi)]+(1-\pi)\overline{c}}{\underline{c}}$ then optimal norm is* Honesty*: $B = hc/\overline{c}$, $P = B/h = c/\overline{c}$ and $Q = 0$.*

Proof. If $\beta > \underline{\beta}$ the cost goes down as B decreases from c to $hc/\overline{c}$ so either we want to stop at $B = hc/\overline{c}$ or lower B further along the intersection. As computed in the previous lemma at the intersection the cost becomes

$$C = c\left[1 + U\frac{\pi + (1-\pi)h}{\underline{c}}\right] + \left[\beta - U\frac{\pi[1+h(1-2\pi)] + (1-\pi)\overline{c}}{\underline{c}}\right] B$$

Therefore if

$$\beta < U\frac{\pi[1+h(1-2\pi)] + (1-\pi)\overline{c}}{\underline{c}} \equiv \overline{\beta}$$

then the cost is decreasing in B so it is best to stop at $B = hc/\overline{c}$ - where $P = B/h = c/\overline{c}$ and $Q = 0$ - that is *Honesty*. □

The lower bound of B (see the last picture) becomes optimal for $\beta > \overline{\beta}$:

Lemma 8. *If $c < \overline{c}$ and $\beta > \overline{\beta}$: if $c > \underline{c}$ the optimal norm is* Minimal*; if $c < \underline{c}$ it is* None*.*

Proof. We see from the expression in the previous proof that if $\beta > \overline{\beta}$ the cost is increasing in B along the intersection so we want to lower B as much as possible, that is: to $B = (c - \underline{c})/[1+h(1-2\pi)]$ if $c > \underline{c}$ - with $P = 1$ and $Q = h - B > 0$ - which is *Minimal*, as in the left panel of the last figure; or to $B = 0$ if $c < \underline{c}$ - with $P = c/\underline{c}$ and $Q = hc/\underline{c}$ - which is *None*, as in the right panel. □

The last four lemmas end the proof of the theorem. We next turn to the heterogeneous matches model.

Internalization with Heterogeneous Matches

We now consider a setting where there are two types of matches: a fraction $1 > \varphi > 0$ of small stakes matches S in which there is a single monitor and $1-\varphi$ high stakes matches H in which there is public information. In both cases the social value of production is $v\theta - \theta^2$. Hence a social norm is of the form θ_i, P_i, Q_i where $i = S, H$ and $Q_H = 0$ as we have assumed that $h = 0$ in the high stakes matches. Small stakes are reflected by a constraint $\theta_S \leq \underline{\theta}$.

We start with the assumptions we use:

Assumption 1. *Intermediate β*

$$\underline{\beta} \equiv \mu(\pi,\pi,0)U < \beta < \left[(1-\varphi)\mu(\pi,\pi,0) + \varphi\left(\mu(\pi,\pi,h) + \frac{1-\pi(1-h)}{1-h^2}\right)\right]U \equiv \overline{\beta}$$

Define

$$\Theta_H = \frac{\left(1+U\pi/(1-2\pi)\right)\left(\varphi(1-2\pi)+(\varphi-1)h\right)+h\left(\beta+1\right)}{\varphi(1-2\pi+h)\left(1+U\pi/(1-2\pi)\right)} > 1.$$

Assumption 2. *Small $\underline{\theta}$*

$$\underline{\theta} < \sqrt{\frac{1-2\pi}{\Theta_H^2 - \frac{h}{1-2\pi+h}}}$$

Assumption 3. *Large enough F*

$$F > \left(1+U\pi/(1-2\pi)\right)\Theta_H\left((1-\varphi)\Theta_H+\varphi\right) \equiv \underline{F}$$

We can now state the result we wish to prove:

Proposition 3. *For $\varphi \geq \underline{\varphi}$, $\underline{\beta} \leq \beta \leq \overline{\beta}$, and small $\underline{\theta}$ there is $\overline{F} > \underline{F}$ such that for $\underline{F} \leq F \leq \overline{F}$*

1. *$Q_S = 0$*
2. *Let $v_0 > 0$ be the largest root of*

$$\frac{1-\varphi}{4(1+U\pi/(1-2\pi)}v^2 + \underline{\theta}\varphi v - \left[\left(1+U\pi/(1-2\pi)\right)\underline{\theta}^2\varphi - \left[U\pi/(1-2\pi)-\beta\right]h\underline{\theta}^2/\overline{c} + F\right] = 0.$$

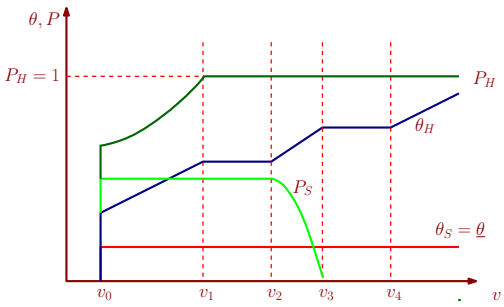
Then for $v < v_0$ the default social norm is used and for $v > v_0$ production is implemented.

3. *There exist additional cutoffs $v_0 < v_1 < v_2 < v_3 < v_4$ given by*

$$v_1 = 2\left(1+U\pi/(1-2\pi)\right)\sqrt{1-2\pi+h\underline{\theta}^2/\overline{c}}, \qquad v_2 = \frac{1-\varphi-\varphi U\pi/(1-2\pi)+\beta}{(1-\varphi)\left(1+U\pi/(1-2\pi)\right)}v_1$$

$$v_3 = \frac{2\left(1-\varphi-\varphi U\pi/(1-2\pi)+\beta\right)\sqrt{\underline{\theta}^2+(1-2\pi)}}{1-\varphi}, \quad v_4 = \frac{1-\varphi+\beta}{1-\varphi-\varphi U\pi/(1-2\pi)+\beta}v_3$$

such that the following diagram is true



with the functions characterizing the optimal norm in the intervals where this is not constant given explicitly in the following table.

	$v_0 \leq v \leq v_1$	$v_2 \leq v \leq v_3$	$v > v_4$
θ_H	$\frac{v/2}{1+U\pi/(1-2\pi)}$	$\frac{(1-\varphi)v/2}{1-\varphi-\varphi U\pi/(1-2\pi)+\beta}$	$\frac{(1-\varphi)v/2}{1-\varphi+\beta}$
P_H	$\frac{\theta_H^2-h\underline{\theta}^2/\bar{c}}{1-2\pi}$	1	1
P_S	$\frac{(1-h/\bar{c})\underline{\theta}^2}{1-2\pi}$	$\frac{\underline{\theta}^2-\theta_H^2+(1-2\pi)}{1-2\pi}$	0

We prove this through a series of Lemmas. First, letting $\varphi_S = \varphi$ and $\varphi_H = 1-\varphi$ the objective function is

$$\sum_{i=S,H} \varphi_i \left[v\theta_i - [c_i + U(\pi P_i + (1-\pi)Q_i)]\right] - \beta B.$$

Lemma 9. *We must have $Q_i = 0$ and $B \geq h\theta_S^2/\bar{c}$. B should be as small as consistent with this inequality. The feasibility constraints $P_i \leq 1$ are given by $B \geq \theta_i^2 - (1-2\pi)$.*

Proof. This is a variation on the intermediate β analysis in the single-type of match case studied above. Recall the monitoring cost factor μ from Theorem 1: the condition

$$\beta < \left[(1-\varphi)\mu(\pi,\pi,0) + \varphi\left(\mu(\pi,\pi,h) + \frac{1-\pi(1-h)}{1-h^2}\right)\right] U$$

from Assumption 1 is exactly the condition that the marginal cost of investment β is less than the corresponding reduction in monitoring cost when the incentive constraints hold in both types of matches with exact equality. Hence to minimize cost for fixed θ internalization B must be at least large enough that $Q_S = 0$. For $i = S$ at the intersection of the two constraints from $Q_S = [hc_S - \bar{c}B]/\underline{c}$ we see that at $Q_S = 0$ we have $B = hc_S/\bar{c} = h\theta_S^2/\bar{c}$ (which is the value of B at Honesty), thus it cannot be smaller than that.

The condition

$$\frac{\pi}{1-2\pi} U < \beta$$

also from Assumption 1 is exactly the condition that the marginal cost of investment β is larger than the corresponding reduction in monitoring cost when production incentive constraint holds

with exact equality and $Q_i = 0$. Hence B should be larger than needed for $Q_S = 0$ only if necessary to satisfy the production incentive constraint.

The third part follows from $P_i = (c_i - B)/(1 - 2\pi)$ for $Q_i = 0$. □

Lemma 10. *Letting $\varphi_S = \varphi$ and $\varphi_H = 1 - \varphi$ the objective function is*

$$\sum_{i=S,H} \varphi_i \left[v\theta_i - (1 + U\pi/(1-2\pi))\theta_i^2 \right] + \left[U\pi/(1-2\pi) - \beta \right] B$$

Proof. Follows from the fact that the objective function with $Q_i = 0$ is

$$\sum_{i=S,H} \varphi_i \left[v\theta_i - (c_i + U\pi P_i) \right] - \beta B$$

by plugging in the value of P_i when $Q_i = 0$. □

Lemma 11. *The solution of the unconstrained problem (0 constraints) is given by*

$$\theta_H^0(v) = \frac{v/2}{1 + U\pi/(1-2\pi)}$$

$$\theta_S^0(v) = \frac{\varphi(1 - 2\pi + h)}{(1 + U\pi/(1-2\pi))\left(\varphi(1-2\pi) + (\varphi - 1)h\right) + h\left(\beta + 1\right)}(v/2) < \theta_H^0(v)$$

and

$$P_H^0(v) = \frac{\theta_H^0(v)^2 - h\theta_S^0(v)^2/\overline{c}}{1 - 2\pi}, \quad P_S^0(v) = \frac{\theta_S^0(v)^2 - h\theta_S^0(v)^2/\overline{c}}{1 - 2\pi} < P_H^0(v)$$

Proof. Without constraints from Lemma 9 $B = h\theta_S^2/\overline{c}$ must hold with equality. The solution follows from plugging this into the objective function and solving the first order conditions. This is transparent for θ_H. For $\theta_S$we have the first order condition

$$\varphi\left(v - 2(1 + U\pi/(1-2\pi))\theta_S\right) + 2\left(U\pi/(1-2\pi) - \beta\right)\frac{h\theta_S}{1 - 2\pi + h} = 0$$

from which the result follows. □

Denote by $\underline{v}$ the unique solution of $\theta_S^0(v) = \underline{\theta}$. From Lemma 11 we see that this is equal to

$$\underline{v} = 2\left(1 + U\pi/(1-2\pi)\right)\Theta_H\underline{\theta}.$$

Next we show that when the $v \leq \underline{v}$ the fixed cost is large enough that the default social norm is optimal:

Lemma 12. *Under Assumptions 3 and 2 when in the unconstrained problem with $v \leq \underline{v}$ the default social norm is optimal and $P_H^0(v) < 1$.*

Proof. Substitute the solution of the unconstrained problem from Lemma 11 into the objective function from Lemma 10 noting that for a quadratic objective $A_1\theta - A_2\theta^2$ the maximized objective

function is $A_1^2/(4A_2)$ to find that the utility from implementing production is equal to

$$\begin{aligned}&(v^2/4)\left(\frac{1-\varphi}{1+U\pi/(1-2\pi)}+\frac{\varphi^2(1-2\pi+h)}{\varphi(1+U\pi/(1-2\pi))\,(h+1-2\pi)-h\,[U\pi/(1-2\pi)-\beta]}\right)\\ =&(v^2/4)\left(\frac{1-\varphi}{1+U\pi/(1-2\pi)}+\frac{\varphi}{(1+U\pi/(1-2\pi))\,\Theta_H}\right).\end{aligned}$$

Substituting $\underline{v}$ shows that Assumption 3 is exactly that F is larger than this utility from implementing production.

Observing that it is increasing and plugging in $\underline{v}$ we find the condition $P_H^0(\underline{v})<1$ can be written as

$$(\Theta_H\underline{\theta})^2-\frac{h\underline{\theta}^2}{1-2\pi+h}<1-2\pi$$

which is satisfied by Assumption 2. □

Next we consider the constrained problem in which $\theta_S=\underline{\theta}$; here the solution is immediate:

Lemma 13. *The solution of the $\theta_S=\underline{\theta}$ constrained problem (1 constraint) is*

$$\theta_H^1(v)=\frac{v/2}{1+U\pi/(1-2\pi)}$$

$$P_H^1(v)=\frac{\theta_H^1(v)^2-h\underline{\theta}^2/\overline{c}}{1-2\pi},\quad P_S^1=\frac{(1-h/\overline{c})\underline{\theta}^2}{1-2\pi}$$

Lemma 14. *We have v_0 given by the largest root of*

$$\frac{1-\varphi}{4(1+U\pi/(1-2\pi))}v_0^2+\underline{\theta}\varphi v_0-\left[(1+U\pi/(1-2\pi))\underline{\theta}^2\varphi-[U\pi/(1-2\pi)-\beta]\,h\underline{\theta}^2/\overline{c}+F\right]=0$$

and

$$v_1=2\,(1+U\pi/(1-2\pi))\sqrt{1-2\pi+h\underline{\theta}^2/\overline{c}}$$

and there exists $\overline{F}>\underline{F}$ such that for $\overline{F}>F>\underline{F}$ we have $\underline{v}<v_0<v_1$.

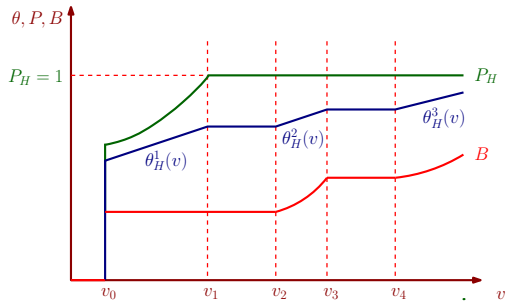
Proof. The first condition is for indifference between the default social norm and implementing production when the constraint $\theta_S=\underline{\theta}$ binds. Observe that the value of v_0 depends on the fixed cost F. Importantly, at $\underline{F}$ we get $v_0=\underline{v}$. So for $F>\underline{F}$ we have $\underline{v}<v_0$. The second condition is for $P_H=1$ also when the constraint $\theta_S=\underline{\theta}$ binds. The existence of $\overline{F}$ follows from the fact that at $\underline{F}$ and $\underline{v}$we have $P_H<1$ by the last part of Lemma 12. So for F close enough to but greater than $\underline{F}$, we get v_0 which is greater than but sufficiently close to $\underline{v}$. Since P_H is continuously increasing in v, $P_H<1$ must still hold at v_0 for v_0 close enough to $\underline{v}$. Therefore for such values of F, we get $v_0<v_1$. □

Before proceeding with the remainder of the proof consider the following intuition of the flat portion of θ_H between v_1 and v_2. As v increases from v_0 we have $Q_S=0$ and we have made an assumption (Assumption 1) that insures that when $Q_S=0$ the most cost effective way of increasing

θ_H is by increasing P_H not by increasing internalization B. Hence as v increases P_H goes up until eventually it reaches 1 and can be increased no further: we label this point v_1. At this point further increases in θ_H are possible only by increasing B and as we just noted the marginal cost of such an increase is greater than that from increasing P_H. Hence it is strictly undesirable to further increase θ_H once v_1 is reached. Eventually, however, as v increases further the additional marginal benefit of θ_H justifies increasing B. Once this point is reached θ_H is determined as the optimal solution to the problem of maximizing utility with the two constraints $\theta_S = \underline{\theta}$ and $P_H = 1$. Denote the solution to this problem $\theta_H^2(v)$. If this is greater than $\theta_H^1(v_1)$ the value of θ_H where P_H first equals 1 this means it must be worth paying the extra B to get more θ_H. Moreover, since there is a unique value of B for which $\theta_H = \theta_H^1(v_1)$ with $\theta_S = \underline{\theta}$ and $P_H = 1$ we see that when $\theta_H^2(v) = \theta_H^1(v_1)$ the two solution are the same; the group is indifferent between the two, so in fact this must be the switch point v_2 where it first becomes optimal to start increasing θ_H again.

We illustrate the remainder of the proof with a table and drawing:

	$v_0 \le v \le v_1$	$v_2 \le v \le v_3$	$v > v_4$
θ_H	$\theta_H^1(v) = \frac{v/2}{1+U\pi/(1-2\pi)}$	$\theta_H^2(v) = \frac{(1-\varphi)v/2}{1-\varphi-\varphi U\pi/(1-2\pi)+\beta}$	$\theta_H^3(v) = \frac{(1-\varphi)v/2}{1-\varphi+\beta}$
P_H	$P_H^1(v) = \frac{(\theta_H^2(v))^2 - h\underline{\theta}^2/\overline{c}}{1-2\pi}$	1	1
B	$h\underline{\theta}^2/\overline{c}$	$\theta_H^2 - (1-2\pi)$	



Lemma 15. *The solution of the problem constrained by $\theta_S = \underline{\theta}$ and $P_H = 1$ (2 constraints) is*

$$\theta_H^2(v) = \frac{(1-\varphi)v/2}{1-\varphi-\varphi U\pi/(1-2\pi)+\beta}$$

$$P_S^2(v) = \frac{\underline{\theta}^2 - \theta_H^2(v)^2 + (1-2\pi)}{1-2\pi}$$

and this is smaller than for the solution constrained only by $\theta_S = \underline{\theta}$. The cutoffs v_2 and v_3 are

$$v_2 = \frac{1-\varphi-\varphi U\pi/(1-2\pi)+\beta}{(1-\varphi)\left(1+U\pi/(1-2\pi)\right)}v_1\,, \quad v_3 = \frac{2\left(1-\varphi-\varphi U\pi/(1-2\pi)+\beta\right)\sqrt{\underline{\theta}^2+(1-2\pi)}}{1-\varphi}$$

Proof. Choosing the optimal θ_H we have $B = \theta_H^2 - (1-2\pi)$ and the objective function becomes

$$(1-\varphi)\left(v\theta_H - (1+U\pi/(1-2\pi))\theta_H^2\right) \\ + \varphi\left[v\underline{\theta} - (1+U\pi/(1-2\pi))\underline{\theta}^2\right] + (U\pi/(1-2\pi)-\beta)\left(\theta_H^2 - (1-2\pi)\right)$$

giving the proposed solution. By Assumption 1 for $v \leq v_1$ this is smaller than for the solution $\theta_H^1(v)$ constrained only by $\theta_S = \underline{\theta}$ indicating that the optimum is that the constraint $P_H = 1$ should not bind. When $P_H = 1$ the constrained solution $\theta_H^2(v)$ is less than $\theta_H^1(v)$ so that the cutoff v_2 where $\theta_H^2(v) = \theta_H^1(v_1)$ is necessarily to the right of v_1 and it is optimal to invest in additional B above v_2. The solution $\theta_H^2(v)$ remains valid as P_S decreases until it reaches 0 which defines v_3. □

The final flat segment occurs when $P_S = 0$: once that constraint binds the benefit of increasing B drops as it no longer reduces P_S and increases θ_H but only increases θ_H. Again when v is large enough as given in the final Lemma it will be desirable to start increasing B.

Lemma 16. *The solution of the problem constrained by $\theta_S = \underline{\theta}$, $P_H = 1$ and $P_S = 0$ (3 constraints) is*

$$\theta_H^3(v) = \frac{(1-\varphi)v/2}{1-\varphi+\beta} < \theta_H^2(v)$$

and the cutoff v_4 is given by

$$v_4 = \frac{1-\varphi+\beta}{1-\varphi-\varphi U\pi/(1-2\pi)+\beta} v_3$$

Proof. The objective function is

$$(1-\varphi)\left(v\theta_H - (1+U\pi/(1-2\pi))\theta_H^2\right) + \varphi(v\underline{\theta} - \underline{\theta}^2) + ((1-\varphi)U\pi/(1-2\pi)-\beta)\left(\theta_H^2 - (1-2\pi)\right)$$

with the first order condition giving the solution. Since $\theta_H^3(v) < \theta_H^2(v)$ is transparently true it follows that the cutoff v_4 is given when v rises sufficiently that $\theta_H^3(v_4) = \theta_H^2(v_3)$. □