Interventions with Sticky Social Norms: A Critique[☆]

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

Abstract

We study the consequences of policy interventions when social norms are endogenous but costly to change. In our environment a group faces a negative externality that it partially mitigates through incentives in the form of punishments. In this setting policy interventions can have unexpected consequences. The most striking is that when the cost of bargaining is high introducing a Pigouvian tax can increase output - yet in doing so increase welfare. An observer who saw that an increase in a Pigouvian tax raised output might wrongly conclude that this harmed welfare and that a larger tax increase would also raise output. This counter-intuitive impact on output is demonstrated theoretically for a general model and found in case studies for public goods subsidies and cartels.

JEL Classification Numbers: A1, D7, D9

Keywords: social mechanisms, pigouvian taxes, adjustment costs

[☆]First Version: October 14, 2017. We would like to thank the editor, five anonymous referees, Heski Bar-Isaac, Marco Casari, Andrea Ichino, Andrea Mattozzi, Rohini Somanathan and seminar audiences at WUSTL, Warwick, Queen's, the Paris Institutions Conference, the Zurich Political Economy Conference, Delhi School of Economics and the University of Trento. We gratefully acknowledge support from the EUI Research Council and MIUR PRIN 2017 n. 2017H5KPLL_01.

[*]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

This paper shows that outside interventions in environments where groups are governed by social norms can have unexpected consequences. We can illustrate the main idea through a simple story. Consider a negative production externality, for example, fishing in a lake. From the work of Coase (1960), Ostrom (1990), and others, we know that it is likely that fishermen will self-organize and use peer pressure to mitigate the externality. We refer to such an arrangement as a social mechanism. Suppose that a naive planner unaware of the existence of such a mechanism arrives on the scene and observing the negative externality introduces a tax designed to reduce it. The fishermen then have a choice. They can negotiate a new social mechanism. If they do so output will go down as expected. However, bargaining is costly and in the presence of the tax an agreement may not be so valuable, so they may choose not to do this. They may instead maintain the existing mechanism even though it is ill-adapted to the presence of a tax. Alternatively, as the externality is mitigated by the tax anyway, rather than maintaining a costly system of monitoring and punishment they may find it better to revert to non-cooperative behavior. Suppose this is the case. While the tax will tend to lower output, abandoning social incentives will tend to increase it and the overall effect is ambiguous. As we will show output may go up rather than down. This, we imagine, will come as a surprise to the naive planner who will then conclude that the tax is a failure, and perhaps get rid of it. That, however, might also be a mistake, as the increase in output induced by the tax may nevertheless be coupled with increased welfare for the group. The goal of this paper is to determine when such a story might be true, and what other consequences an unanticipated intervention might have in the presence of a social mechanism.

Our model follows Townsend (1994) and Levine and Modica (2016) by modeling the self organization of a group as a mechanism design problem. Our setting is one of a production externality. The group can establish an output quota, it has a noisy monitoring technology for observing whether the quota is followed, and it can punish group members based on these signals. The new feature that this paper incorporates is that social norms may be costly to redesign after an external intervention: this introduces a stickiness in which social norms may be maintained when they are no longer optimal, or abandoned altogether.[4] We study a simple environment with two periods. In the first period the group designs a social mechanism anticipating the second period will likely be the same as the first. In the second period an unanticipated intervention may take place - for example, the introduction of a Pigouvian tax. If there is an intervention the group may, at a cost, design a new mechanism to cope with changed circumstances. It may at no cost choose to maintain the quotas and punishments of the existing mechanism - although individuals will reoptimize in response to changed circumstances. Finally, it may simply abandon any effort to police itself and revert to the "law of the jungle," which is to say to non-cooperative behavior.

[4]Levine (2012) gives evidence that social norms change very quickly when incentives for such a change are strong, while Bigoni et al (2016) and Dell et al (2018) give evidence that social norms can be sticky when incentives for change are weak.

Our general environment applies to a variety of problems including that of a standard externality with a Pigouvian tax, subsidies for the provision of a public good,and the Cournot setting of a cartel. One of the strengths of the approach is that by covering a broad range of settings it enables us to use data from one arena to make predictions about a less studied arena. That is: the problem of colluding business firms in a cartel faced with a negative demand shock is no different than that faced by a group facing a negative externality hit with a Pigouvian tax. We present evidence that a negative demand shock to a cartel can increase output - for exactly the same reason a city with local air pollution controls might increase pollution in response to a federal carbon tax. Similarly our analysis applies to subsidies for public good provision. Here we present evidence in the arena of foreign aid: a study by Bano (2012) shows that in some cases subsidies reduced public good provision and that this was because existing monitoring arrangements were abandoned for the "law of the jungle."

We first consider the case of a Pigouvian tax in a stylized model. Here we provide a complete analysis with closed-form solutions. If the size of the intervention is small the group does not respond at all. There is a threshold at which output jumps. If bargaining cost is small output jumps down with the new norm and remains lower than in the first period. This is the same as we would expect if individuals faced adjustment costs as in the widely used menu cost model of Calvo (1983). However if bargaining cost is large, for a range of interventions output jumps up, declining as the tax goes up to values lower than in the first period. Here as the intervention increases in size the first period norm becomes increasingly dysfunctional until it is better simply to revert to the law of the jungle; and as long as the tax is not too high non-cooperative behavior can result in higher output. This is the counterintuitive outcome: output can move in the wrong direction in response to an intervention.

Unlike individualistic models, whether or not the tax is rebated lump sum to the group matters. In particular if an outside agency intervenes to set a naive Pigouvian tax and keeps the proceeds there will be underproduction as in the standard case. We also study welfare. If the group keeps the tax revenue and production increases this is evidence of a welfare improvement: it means the policy is a success notwithstanding the increase in production it brings about. The nature of the rebate also matters in the attitude of the group towards taxes. If the group keeps the tax revenue it is happy with higher taxes; if the outside agency keeps the taxes if the group is able to do so it may wish to take political action to repeal the tax not withstanding the fact that the tax may be an efficient Pigouvian one.

In the tax setting there is a clear trade-off between public policy in the form of taxes and the use of social norms to mitigate the externality. To an extent this tradeoff has been examined in the literature on regulation stemming from the work of Coase (1960): however the general tendency in that literature is to either argue that the private sector is able to solve the problem, or to argue, as for example Chari and Jones (2000), that the mechanism design problem is insurmountable, and that only public policy can solve the problem. Here we take a more nuanced approach.

The second part of the paper examines the question of when we might expect to see an increase

in output in response to an intervention that from both an individualistic and social perspective "ought" to lower output. Here we allow general functional forms that encompass cartels as well as tax externalities and general monitoring functions. We find that there are three features that lead to anomalous output increases. First, relatively high bargaining costs. This means that when an intervention takes place it is not worth reaching a new agreement. Second, an intermediate size of intervention. If the intervention is small it is not worth changing the existing social norm; if it is large even non-cooperative output will be less than the original quota. Finally, the monitoring function should exhibit left insensitivity, at least approximately: this means that decreasing output below the quota has little effect on the chances of an erroneous signal indicating the quota was violated. Roughly speaking, if we think of the quota as being like a speed limit, say seventy kilometers per hour, this means that the chances of getting a fine are pretty much the same regardless of whether you drive sixty five or forty five.

2. The Model

In each period $t = 1, 2$ identical group members $i \in [0, 1]$ engage in production choosing a real valued level of output $X \geq x_t^i \geq 0$. The utility of a member i in period t depends upon the real valued state $\omega_t \geq 0$, their own output, and the average output of the group $x_t = \int x_t^i di$ according to $u(\omega_t, x_t, x_t^i)$.

The presence of x_t represents an externality: we adopt the convention that the externality is negative. Because of the externality the group collectively faces a mechanism design problem, and we assume that incentives can be given to group members in the form of individual punishments based on monitoring: the group can set a production quota y_t and receives signals of whether or not individual output exceeds the quota. Based on these signals it can impose punishments. Specifically, monitoring generates a noisy signal $z_t^i \in \{0, 1\}$ where 0 means "good, likely respected the quota" and 1 means "bad, likely exceeded the quota." The probability of the bad signal is given by a weakly increasing function $\Pi(x_t^i - y_t)$ defined on the real line. We assume that punishments must take place in the period in which the signal is received, and when the signal is bad the group imposes an endogenous utility penalty of P_t.[5] This may be in the form of social disapproval or even in the form of monetary penalties.[6]

The social cost of the punishment P_t is ψP_t where $\psi > 0$ could be greater or less than one. For example, if the punishment is that group members are prohibited from drinking beer with the culprit that might be costly to the culprit's friends as well as the culprit. In this case $\psi > 1$. Or it might be that the punishment is a monetary fine most of which is shared among the group

[5] In principle punishments could be issued even for a good signal: as incentives depend only on the difference in punishment between the good and bad signal and punishments are costly this will not be part of an optimal mechanism, so for notational simplicity we rule it out.

[6] In principle the group might also impose downward quotas against underproduction. Because the externality is negative if there is any cost associated with this then the group prefers not to do so, so for notational simplicity we do not consider this possibility.

members. In that case there would be very little social loss so we would expect $\psi < 1$. In addition to the social cost of punishment there may also be a cost $\psi_0 \geq 0$ of operating the monitoring system - for example, sending spies to observe output. This cost is only incurred when $P_t > 0$ since if there is no punishment there is no need for monitoring.

The tools available for mechanism design in period t consist of a quota y_t and a punishment for a bad signal P_t. The overall period t utility of a member i is $u(\omega_t, x_t, x_t^i) - \Pi(x_t^i - y_t)P_t$. These utilities define a game for the group members. If the mechanism designer chooses (y_t, P_t) we denote by $X(y_t, P_t)$ the set of x_t such that $x_t^i = x_t$ is a symmetric pure strategy Nash equilibrium of this game. We refer to a triple (x_t, y_t, P_t) with $x_t \in X(y_t, P_t)$ as an *incentive compatible social norm.*[7] If an incentive compatible social norm issues no punishments ($P_t = 0$) we call it *non-cooperative.* The mechanism designer is benevolent and welfare from an incentive compatible social norm (x_t, y_t, P_t) is given by

$$W(x_t, y_t, P_t) \equiv u(\omega_t, x_t, x_t) - \psi\Pi(x_t - y_t)P_t - \psi_0 \cdot 1\{P_t > 0\}.$$

2.1. Adjustment Costs and the Mechanism Design Problem

In the first period the state is a given ω_1. In the second period there are two possibilities: it may be the same as the first period with $\omega_2 = \omega_1$, or an *intervention* may take place in which case $\omega_2 > \omega_1$. If an intervention occurs it is observed at the beginning of the second period. Our focus will be on the case where the chance of intervention is *a priori* regarded as low, that is, the intervention is "unanticipated" or that the mechanism designer is unaware of the possibility of intervention.[8]

In the initial period $t = 1$ the group solves the mechanism design problem of choosing an incentive compatible *initial social norm* (x_1, y_1, P_1) as if the second period will be the same as the first. As there is limited commitment and no connection between the two periods, this amounts to ignoring the second period and maximizing period 1 welfare over incentive compatible social norms.

In period 2 if an intervention has occurred there are three possibilities:

1. (status quo) The initial design (y_1, P_1) can be costlessly maintained, with the designer choosing any x_2 such that (x_2, y_1, P_1) is incentive compatible at ω_2.
2. (non-cooperative) Any non-cooperative social norm $(x_2, y_2, 0)$ may be chosen.
3. (re-optimize) For a fixed cost of $F > 0$ a new incentive compatible social norm (x_2, y_2, P_2) may be chosen.

The fixed costs of adjustment are in the spirit of menu costs in the macroeconomic literature as in Calvo (1983).[9] Here our basic presumption is that reverting to the non-cooperative norm is costless while designing a new social norm is costly. Reverting to a non-cooperative social norm is a decentralized decision: if it is evident that the non-cooperative social norm is superior to the

[7] In the language of contract theory it is an enforcement contract with costly state verification.

[8] See, for example, Modica and Rustichini (1994).

[9] The fixed costs 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.

alternatives there is no need to get together to discuss this and reach an agreement, implicitly everyone has agreed in advance that in this case they will all go their own way. By contrast developing a new social norm cannot be decentralized and the group must be reconvened to agree upon a new social norm.[10]

2.2. Why Do Social Mechanisms Break Down?

A key element of our theory is the possibility that in response to an unanticipated change in circumstances a social mechanism may be abandoned in favor of non-cooperative behavior. This can have counter-intuitive consequences: in particular an adverse intervention that would ordinarily reduce output might instead increase output. Is there evidence that social mechanisms do break down in response to unanticipated changes? Is this due to bargaining costs? One type of social mechanism that has been extensively studied by economists are cartels.

Our theory applied to cartels differs from those most common in the theory of repeated games. In Green and Porter (1984), Rotemberg and Saloner (1986) or Abreu, Pearce and Stacchetti (1990) price wars are a disciplinary device and are the anticipated consequence of real or apparent cheating. In our account, as in the theoretical and empirical account of Harrington and Skrzypacz (2011), cartel discipline is achieved through modest individual penalties for real or apparent cheating. In the empirical literature our model of cartel breakdown appears to be the more relevant one. Indeed, much of the empirical literature, for example the classical study of sugar cartels by Genesove and Mullin (2001), is devoted to debunking the price war model. As an example we quote from the survey by Levenstein and Suslow (2006): "after the adoption of an international price-fixing agreement in the bromine industry, the response to violations in the agreement was a negotiated punishment, usually a side-payment between firms, rather than the instigation of a price war... As repeatedly discovered by these cartel members, the threat of Cournot reversion is an inefficient way to sustain collusion."

In our account, unlike in the repeated game literature, cartel breakdown occurs because of the cost of bargaining in the face of unanticipated changes in circumstances. Again this seems to be the relevant reason for cartel breakdown. Again from Levenstein and Suslow (2006) "Bargaining problems were much more likely to undermine collusion than was secret cheating. Bargaining problems affected virtually every cartel in the sample, ending about one-quarter of the cartel episodes." Their overall conclusion is "cartels break down in some cases because of cheating, but more frequently because of entry, exogenous shocks, and dynamic changes within the industry."

This evidence suggests that social mechanisms do revert to non-cooperative behavior because of the cost of bargaining in the face of changed circumstances. The literature has not addressed the issue of whether as a result, output increases in response to unanticipated adverse changes. Recently,

[10]Notice that we do not allow advance contingency planning. The idea is that to do so is costly. In this we follow the literature on incomplete contracting such as Hart and Moore (1988) and rational inattention such as Sims (2003). Our model is similar to those of unawareness as in Modica and Rustichini (1994) and in the spirit of Tirole (2009) and Dye (1985) or costly contemplation such as Ergin and Sarver (2010).

however, there has been a rather striking natural experiment. In response to the unanticipated reduction in oil demand due to the covid-19 pandemic, OPEC+ attempted to negotiate reduced quotas. On March 8, 2020 bargaining broke down. Subsequently cartel members announced plans instead to increase output, and they did so. During the period December 21, 2019 to March 20, 2020 while the agreement was in effect, and including the period clearly prior to the Covid-19 shock, OPEC output ranged from 27.8 to 28.6 millions of barrels per day. In the following month March 21 to April 20 OPEC output increased to 30.4 mb/d, a more than 6% increase in output.[11] In brief an unanticipated negative demand shock resulted in a substantial increase in cartel output.[12]

3. Pigou

We give a detailed analysis of a Pigouvian tax in a simple quadratic framework. Each individual derives a private benefit from output $U(x_t^i) = (V+1)x_t^i - (V/2)(x_t^i)^2$ up to the satiation point $X = (V+1)/V$ which we also take to be the upper limit on output. The negative externality reduces the benefit by x_t. In addition, the state ω_t represents a Pigouvian tax a fraction of which $0 \leq \alpha \leq 1$ is rebated in a lump sum. Overall individual utility is therefore $u(\omega_t, x_t, x_t^i) = U(x_t^i) - \omega_t x_t^i - (1-\alpha\omega_t)x_t$.We consider a simple monitoring technology: $\Pi(x_t^i - y_t) = \pi > 0$ if $x_t^i \leq y_t$ and $\Pi(x_t^i - y_t) = \pi_B > \pi$ if $x_t^i > y_t$. Define the *monitoring difficulty* $\theta = \pi/(\pi_B - \pi)$. We assume moreover that $\psi_0 = 0$ and $\psi = 1$.

To focus thinking, consider first the limiting case in which $\pi = 0$ and there are no monitoring costs so that members can be forced to meet any target $y_t = x_t$. In this case the group simply maximizes the utility $u(\omega_t, x_t, x_t) = x_t\left[V - (1-\alpha)\omega_t - (V/2)x_t\right]$. As all the optimization problems in this section are quadratic we collect the calculations in the Online Appendix, and report the results here. The group chooses the *first best* $x_t^f = (V-(1-\alpha)\omega_t)/V$ with sufficiently large punishments to deter deviation, and the corresponding welfare is $u_t^f = (V-(1-\alpha)\omega_t)^2/(2V)$. To avoid the uninteresting boundary case we assume that $(1-\alpha)\omega_t \leq V$. To assure that the non-cooperative output is higher than the first best we assume in addition that $\omega_2 \leq 1/\alpha$.

In the special case in which $\alpha = 1$, so all the tax is rebated to the group this is the *Pigouvian solution* $x^P = 1$ with $u^P = V/2$. Here we have a policy irrelevance result: when monitoring costs are low and most of the tax is rebated to the group tax policy will have little effect on output of an organized group.

3.1. Individual Optimality and Monitoring Costs

Consider next the problem of choosing an optimal first period social norm or re-optimizing in the second period, each for a given value of ω_t. It is useful to break the problem into two steps and consider first the problem for fixed x_t of choosing (y_t, P_t) to minimize the monitoring cost

[11] Reported in the OPEC Monthly Oil Market Report for March and May 2020.

[12] It should be noted that the marginal cost to Saudi Arabia of extracting a barrel of oil (see knoema.com) is estimated to be less than $3 while even with the substantial price fall that took place, the price remained well above $20 so there is no issue here of a price war in the sense of producing below marginal cost.

$M(x_t) = \pi P_t$ subject to incentive compatibility. With this simple monitoring technology if an individual decides to deviate from x_t there is a unique optimal deviation determined by ignoring the punishment: we denote this by x_t^B. If (x_t, y_t, P_t) is an optimal social norm then we call y_t an optimal quota.

Theorem 1. *The optimal deviation is $x_t^B = (V + 1 - \omega_t)/V$. If $x_t < x_t^B$ then the optimal quota is $y_t = x_t$ and monitoring cost is given by $M_t(x_t) = \theta(u(\omega_t, x_t, x_t^B) - u(\omega_t, x_t, x_t))$.*

Notice that the individual optimum in the absence of penalty is independent of x_t: in other words the non-cooperative social norm also generates output $x_t^N = x_t^B$. Note also that x_t^B decreases linearly in ω_t.

Proof. The only feasible quota is $y_t = x_t$ because for any other quota group members can increase output without changing the probability of being punished. The optimal deviation x_t^B is the maximizer of $u(\omega_t, x_t, x_t^i)$ with respect to x_t^i. Hence the greatest gain from deviating is $u(\omega_t, x_t, x_t^B) - u(\omega_t, x_t, x_t)$. The incentive constraint is therefore $(\pi_B - \pi)P_t \geq u(\omega_t, x_t, x_t^B) - u(\omega_t, x_t, x_t)$. Monitoring cost is minimized when P_t is minimized, so the optimal punishment is determined when the incentive constraint holds with equality. Solving and plugging into monitoring cost yields the result. □

3.2. Optimal Social Mechanisms

Our main interest is in the response in the second period when there is an intervention: which social norm is chosen and what is the consequence for output?

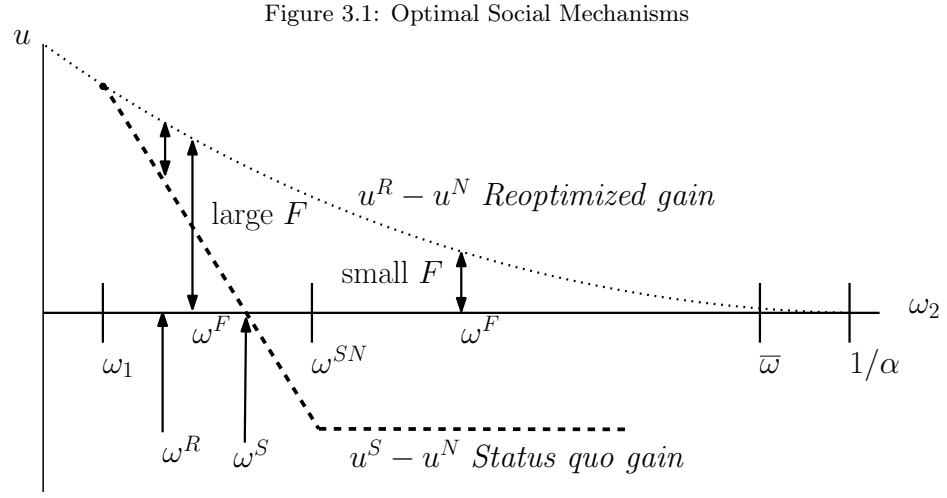
The first period and re-optimal second period problem can be conveniently expressed in terms of monitoring cost as the problem of choosing output x_t to maximize $u(\omega_t, x_t, x_t) - M_t(x_t)$. Denote the solution to the problem by x_t^R, u_t^R. The non-cooperative solution we denote by x_t^N, u_t^N. We must also consider the status quo solution in the second period with solution x_2^S, u_2^S. The situation is described in Figure 3.1. For each of the different social norms as a function of ω_2 we compute the utility gain over the non-cooperative social norm. The dotted reoptimized gain curve shows how much utility is gained over the non-cooperative social norm by reoptimizing; it is necessarily non-negative andreaches a minimum of zero at $\omega_2 = 1/\alpha$. The bold line which coincides with the x-axis is the net utility gain from the non-cooperative social norm over itself, therefore 0.

3.2.1. Status Quo versus Non-Cooperative

The utility gain from the status quo social norm over the non-cooperative social norm is $u_2^S - u_2^N$. This is shown by the piecewise linear dashed *status quo gain* curve. The main result of this paper is that as the tax ω_2 is increased to the point ω^S where the status quo gain curve reaches zero and the group switches social norms, output jumps up.

To see why this is, increase the tax rate starting at $\omega_2 = \omega_1$ where the tax rate in the second period is the same as the first. Here the status quo gain is the same as the reoptimized gain and

Figure 3.1: Optimal Social Mechanisms



so it is strictly positive. This implies that the non-cooperative output x^N is strictly bigger than the status quo output which is the first period output x_1. There are two crucial facts. The first is that as the tax rate increases the status quo output remains stuck at x_1 while the non-cooperative output drops until the two become equal at some tax rate ω^{SN}. The second is that the point at which the group switches from the status quo to the non-cooperative output ω^S lies to the left of ω^{SN} so that the switch induces an increase in output from x_1 to x^N.

The key to these results is that increases in ω_2 decrease the incentive to deviate. This means that non-cooperative output falls as ω_2 increases. However, as long as the incentive to deviate is positive the status quo output remains stuck at x_1. In other words until we reach ω^{SN} the status quo output is x_1 and the non-cooperative output is larger and decreasing with ω_2. However, at ω^{SN} the status quo must be strictly worse than the non-cooperative social norm because the status quo social norm incurs an unnecessary monitoring cost in order to achieve the same output. This gives the conclusion that the point of indifference between the two norms ω^S lies to the left of ω^{SN}. That is, the switch from status quo to non-cooperative norms takes place where the status quo output is still equal to the first period output x_1 and the non-cooperative output is strictly higher. This upward jump when switching from the status quo to non-cooperative social norm is our main result.

3.3. The Role of Bargaining Costs

If $u_t^R - u_t^N > F$ then it is better to reoptimize than use the non-cooperative social norm and conversely. Since $u_t^R - u_t^N$ is downwards sloping, there is a unique tax rate ω^F where $u_t^R - u_t^N = F$. For lower tax rates $\omega_2 < \omega^F$ it is better to reoptimize, for higher tax rates $\omega_2 > \omega^F$ to use the non-cooperative social norm. The interaction of bargaining costs with the status quo social norm depends on the size of F.

3.3.1. Large Bargaining Costs

We say that F is large when $\omega^F < \omega^S$ The double arrows show the size of F. Observe that in this case F is necessarily larger than $u_t^R - u_2^S$ (the difference between the reoptimized and status quo gain curves) at ω^F, since $u_2^S - u_2^N > 0$ there. And the difference $u_t^R - u_2^S$ shrinks going left. So for tax rates no greater than ω^F the status quo social norm is better than re-optimizing. Since the non-cooperative social norm is better at higher tax rates it is never optimal to re-optimize. As ω_2 increases from ω_1 then we have the following consequences for choice of social norm and output. For $\omega_1 \leq \omega_2 < \omega^S$ the status quo norm is chosen and output remains fixed at x_1. For higher values of $\omega_2 > \omega^S$ it is optimal to switch to the non-cooperative social norm. In the range $\omega^S < \omega_2 < \omega^{SN}$ output at the non-cooperative social norm is higher than x_1. This means that optimal output actually jumps up, then decreases until it again reaches x_1 at $\omega_2 = \omega^{SN}$. After that it falls below x_1.

Output that increases in response to an intervention is the most unexpected and striking feature of our model: we will subsequently investigate how robust a phenomenon it might be. The idea, as indicated in the introduction, is a simple one: with high bargaining costs a change in circumstances can lead to a breakdown of the existing social norm and this can increase output.

3.3.2. Small Bargaining Costs

We say that F is small when $\omega^F > \omega^{SN}$. To the right the non-cooperative social norm is best. However, it may be that $\omega^F > \overline{\omega}$ so this point may never be reached: it depends upon α as we shall discuss subsequently. In case $\omega^F < \overline{\omega}$ there will be a switch from the reoptimized social norm to the non-cooperative social norm at ω^F and output will jump up - but cannot rise so high as x_1 since ω^F lies to the right of ω^{SN}. The point is that as the tax rate increases the non-cooperative equilibrium gets close to the first best anyway, so the gain to reoptimizing is small and not worth bargaining over.

To the left of ω^F the non-cooperative social norm is never used. To see when the status quo social norm is used, we need to find the unique point ω^R where $u_t^R - u_t^S = F$, which as shown in the diagram is also the distance between the reoptimized and status quo gain curves. Since at $\omega_2 = \omega_1$ we have $u_t^R = u_t^S$ this point always lies to the right of ω_1 and from the fact that ω^F lies to the right of ω^{SN} it must also be that ω^R lies to the left of ω^S as shown in the diagram.

Initially then, for $\omega_1 \leq \omega_2 < \omega^R$ the status quo is maintained and output remains fixed. As ω_2 rises into the range $\omega^R < \omega_2 < \omega^F$ the status quo is abandoned in favor of re-optimization. Output jumps down, then continues to declines. Eventually if ω^F is reached it will jump up again to the non-cooperative level, although not as high as x_1, and then again start to decline.

The case in which $\omega^S < \omega^F < \omega^{SN}$ is similar to $\omega^F > \omega^{SN}$: - as ω_2 goes up the transition is from status quo to reoptimization to non-cooperative - except that when ω^F is reached and output jumps up it jumps to a level higher than the original level of output at ω_1. It then declines, eventually falling below the original level.

What is striking in this case is what happens in the vicinity of ω^F. For slightly lower ω_2 output has dropped. For slightly higher ω_2 output has increased. Suppose that two different empirical

studies were conducted, in different but very similar locations, for example. Never-the-less if in one location ω_2 was just below ω^F and in the other just above, the first study would conclude that intervention lowers output while the second would conclude that the intervention raises output. While this may not be a frequent occurrence it is good to be aware of the possibility.

3.4. The Lump Sum Rebate, Overshooting and Tax Repeal

The lump sum rebate is not neutral for either behavior or welfare. In particular when $\alpha = 1$ the group favors taxes over quotas up to the Pigouvian level. To understand this, observe that for $\alpha = 1$ not only must all taxes be rebated to the group but the tax system must be lossless.[13] In this case both taxes and quotas mitigate the externality, but taxes are superior to quotas because a tax unlike a quota can be enforced costlessly. With quotas production is at most at the non-cooperative level for the given tax, and possibly less, but always greater than the Pigouvian output of $x^P = 1$. Contrast this to the situation in which $\alpha = 0$ (say). Here a naive planner might set the Pigouvian tax. If bargaining cost is small and the group reoptimizes this will result in output $x_2^R = 1 - 1/(V(1+\theta))$,that is the group will produce too little in an effort to avoid the tax loss.

This undershooting result, however, understates the potential for error in setting a Pigouvian tax when it is not rebated to the group. A group unlike an individual can take political action. Let us extend the model to allow this possibility. Group utility is convex in the tax rate, so the optimal tax rate is either 0 or so high that output $x_3^R = 0$. A calculation shows that zero output is strictly optimal if an only if $V < \theta/(1+\theta)$.The takeaway is that with $\alpha = 0$ if the cost of organizing to do so is low the group will always repeal the tax: if V is large relative to monitoring difficulty it will eliminate the tax. Perhaps more surprising, if less likely, is that if V is small relative to monitoring difficulty it will set the tax high and shut down production - an extreme form of overshooting.

An interesting example of a group responding to the naive imposition of Pigouvian taxes by engaging in tax repeal is the case of the French "yellow vests." In this instance output x_t^i represents driving speed, while the intervention ω_2 is the inverse of the speed limit. On July 1, 2018 the French Federal Government lowered the speed limit on secondary highways from 90 km/h to 80 km/h ostensibly to reduce highway accidents. The bulk of the impact fell on rural communities where there are no primary highways. Although driving is to an extent anonymous, there are informal social norms, and drivers who are perceived to drive excessively fast are often punished.[14] As drivers observe one another well, we hypothesize that monitoring difficulty θ is relatively low. Moreover, $\alpha < 1$: the speed camera revenue is not returned to rural drivers who receive only an indirect benefit. Finally, F was quite low due to the advent of social media: Facebook played a key role in the organization of the yellow vests. Hence our theory says that if they could do so at low cost they would organize not only a new driving speed norm, but also eliminate the tax. In

[13] Real tax systems like informal enforcement systems have costs associated with them. In countries with strong state capacity the assumption that there are few such costs may be a reasonable approximation: much of tax infrastructure is a sunk cost so irrelevant, and tax authorities have legal access to information such as banking records that the private sector does not.

[14] While fictional, the Damián Szifron film "Relatos Salvajes" illustrates the idea well.

fact the yellow vests did act to "repeal" the tax. The rate of traffic camera destruction jumped by 400% and in the year following about 75% of all traffic cameras in France were destroyed.[15]

3.5. Welfare and Late Parents

What is particularly striking is that with the full lump sum rebate ($\alpha = 1$) welfare is unambiguously increasing in the tax rate (up to the Pigiouvian level $\omega_2 = 1$). In particular if bargaining costs are non-negligible so that output jumps up - at ω^F or ω^S as the case may be - the increase in output nevertheless increases welfare. That is, despite the increased externality due to higher output welfare increases because of the decrease in monitoring costs.

An interesting case in point is the study of Gneezy and Rustichini (2000). They studied the introduction of modest fine for picking up children late at a day-care center. They observed that this resulted in more parents picking up their children late - the opposite of the expected and intended effect. In our terminology the intervention is the level of the fine. Initially there was no fine $\omega_1 = 0$, then one was imposed $\omega_2 > 0$. As there was no prior warning or discussion of the fine, it is reasonable to think it was unanticipated. Moreover, as the fine was introduced suddenly and without explanation it might well have been anticipated to be of short duration (as in fact it was) so that it would not be worth renegotiating to identify the re-optimal social norm reducing lateness. Hence our theory predicts if ω_2 were chosen slightly larger than the switching point indeed more parents would pick up their children late.

Authors including Gneezy and Rustichini (2000) and Benabou and Tirole (2006) who have discussed the increased lateness have assumed that this resulted in a drop in welfare. A day-care center, however, is a closed system in which the school is supported by fees from the parents and different schools compete with each other. Implicitly, the money from fines either reduces what parents have to pay, or increases the services they receive. In other words, in this setting we think $\alpha = 1$. If this is the case then the assumption that welfare decreased is wrong: in fact it went up. This highlights the importance of knowing whether social norms are involved and the role of the lump sum rebate.

Other theories than ours have been used to explain the increase in lateness: one of the best worked out is that of Benabou and Tirole (2006). Their idea is that in the absence of fines, picking up children on time serves a valuable self-signaling purpose of virtue. With fines, the signaling value of being on time is lowered enough that it becomes worthwhile to be a little late and pay the fine. In contrast in our account prior to the fine there was an informal system of enforcement. Teachers scolded parents who were late and complained to their peers and other parents about people who were persistently late. After the fines were introduced this stopped and parents simply paid their fines. That is, there was punishment before but not after. While this is plausible we do not know whether or not it was the case, and hence we do not have direct evidence about the merits of our

[15] Private communication from Pierre Boyer. Our account is based on Boyer et al (2019) who documents both the link between the change in speed limit and the yellow vest movement, as well as the systematic way in which that group organized itself.

theory versus that of Benabou and Tirole (2006). As the welfare analysis for the two theories is opposite it is of importance to know.

The key lesson here involves the way in which field experiments are conducted. It was possible for Gneezy and Rustichini (2000) to have arranged the experiment to observe punishment before and after. This could have been done by direct observation of teacher behavior at the pickup point - did they scold parents before, but not after? It could also have been done by a before and after survey instrument asking parents and teachers about their expectations of the response to late pickup. In other words: it would be desirable if field experiments where social norms might be involved attempted to ascertain the presence of informal punishments and if this was changed by intervention.

In existing analyses an upward jump in output in response to a Pigouvian tax is regarded as a failure of policy. The goal of the policy is to reduce output in the face of an externality. But that analysis may miss the mark. If there are informal punishments and α is large, increased output is an indication that the policy has a desirable effect. While the increase in output has a negative consequence for welfare, overall welfare goes up because by switching to the non-cooperative norm the cost of monitoring is avoided and this more than makes up for the loss from increased output.

3.6. Costly Bargaining in the First Period

Implicitly we have assumed that while there is a cost F of introducing a re-optimal social mechanism in the second period there is no fixed cost of choosing the first period optimal social norm. In the the type of applications we have in mind we think this assumption makes sense. The "first period" represents a long ongoing situation while the "second period" represents an unanticipated break with the past. In an ongoing situation there is time for experimentation with different mechanisms and discussion of what might be the best mechanism. Over time people meet for all sorts of reasons and it is of low cost to discuss among other things the implementation of a social mechanism. In the experimental research of Fehr and Gachter (2000) we see that the solution to a social mechanism problem develops slowly over time. All of this suggests that in the "first period" an optimal social mechanism is likely to be developed. By contrast in the immediate aftermath of an unanticipated change developing a re-optimal social mechanism would require crash meetings and leisurely experimentation would have to be replaced by a more careful assessment of the situation. All of this suggests that it makes sense to think of re-optimization as more costly than the initial optimization. It also suggests there might be a "period three" after the unanticipated change in which the use of the status quo or non-cooperative mechanism in the second period is replaced by a reoptimal social norm.

While we think this assumption makes sense it is by no means crucial to the analysis.[16] As shown in the diagram $u_2^R - u_2^N$ is decreasing in the tax ω_2 . Suppose in fact that F applies in the first period. It will be optimal to introduce the optimal social norm in the first period provided

[16]It may be, for example, that it is easier to agree to a revised contract than to create one whole cloth, so ultimately the ratio of costs between the first and second period is an empirical one.

that $u_1^R - u_1^N \geq F$. This simply means that ω_F lies to the right of ω_1 and this is exactly the case we have studied. That is: to the right of ω_F it is optimal to use the non-cooperative social norm in place of the reoptimal social norm, and to the left it is optimal to use the reoptimal social norm. As long as $\omega_1 < \omega_F$ it will be optimal to introduce the optimal social norm in the first period.

4. When Does Output Increase?

We now consider more general utility and monitoring functions. Our goal is to find conditions under which an intervention induces an increase rather than a decline in output. For this to be the case we know that F must be reasonably large so that the switch when it takes place is to the non-cooperative equilibrium not to the re-optimal social norm. We must also consider the role of the fixed cost of monitoring ψ_0. If this is too large then it will already be optimal to use the non-cooperative equilibrium in the first period and the model is a standard one. We are interested in the case where ψ_0 is not too large. Hence by F is large, ψ_0 not too large we mean that F is large enough that it is not optimal to reoptimize and that ψ_0 is small enough that it is optimal to optimize in the first period.

In the analysis which follows we will see that the key property of monitoring technology is that of *left-insensitivity*: lowering output below the quota does not reduce the chances of being caught. This was the case in the Pigou example. Our ultimate goal will be to establish that output increases in response to shocks can occur when the monitoring technology is left insensitive or at least approximately left insensitive.

4.1. Regularity of Utility

We impose relatively standard conditions on $u(\omega_t, x_t, x_t^i)$ that incorporate the convention that the externality is negative and ensure that the Nash equilibrium and social optimization problem in the absence of monitoring are well-behaved.

To capture the convention that the externality is negative we assume that $D_2u(\omega_t, x_t, x_t) < 0$. To capture the convention that increasing ω_t mitigates the externality we assume that such increases reduce the individual incentive to produce more, that is, $D_{31}u(\omega_t, x_t, x_t^i) < 0$. These are conventions in the sense that we could work as well with positive externalities by changing the sign of x_t^i and in the sense that it does not matter which direction of change in ω_t mitigates the externality.

We next make assumptions that guarantee that both the social planner problem and individual optimizations problems are well-behaved. As indicated these are standard.

First, we assume that the social objective $u(\omega_t, x_t, x_t)$ is concave, that is, $D_{22}u(\omega_t, x_t, x_t) + 2D_{23}u(\omega_t, x_t, x_t) + D_{33}u(\omega_t, x_t, x_t) < 0$. In addition we assume that there is an interior *social optimum* $0 < x_t^f < X$ where $D_2u(\omega_t, x_t^f, x_t^f) + D_3u(\omega_t, x_t^f, x_t^f) = 0$. Because the objective is concave, this is unique.

Second, we ensure that the non-cooperative mechanism has a unique and well-behaved symmetric pure strategy equilibrium. For existence we require concavity in own action $D_{33}u(\omega_t, x_t, x_t^i) < 0$. To ensure that the equilibrium is well-behaved we add the regularity condition that $D_{33}u(\omega_t, x_t, x_t^i) +$

$D_{23}u(\omega_t, x_t, x_t^i)$ has the same sign (negative) as $D_{33}u(\omega_t, x_t, x_t^i)$ (it suffices that $D_{23}(u(\omega_t, x_t, x_t^i) \leq 0$). This implies that the non-cooperative mechanism has a unique equilibrium output level and that it lies above the social optimum. Lastly, we assume that $D_3u(\omega_1, X, X) < 0$ so that non-cooperative output is interior at ω_1.

Our final assumption is that there is a sufficiently large intervention $\overline{\omega}$ that the corresponding non-cooperative output is lower than at the social optimum in the first period. Specifically, we assume that $D_3u(\overline{\omega}, x_1^f, x_1^f) \leq 0$.

When these assumptions are satisfied we say that *utility is regular*. We assume from this point on that this is the case. It is easily checked that it is so in the Pigou analysis of the previous section. The next result is completely standard.

Theorem 2. *If utility is regular then for $\omega_t \geq \omega_1$ there is a unique non-cooperative output level $x_t^N > x_t^f$ strictly decreasing in ω_t when positive. Moreover the first best x_t^f is weakly decreasing in ω_t.*

For some results we will require an additional condition. We say that utility is *separable* if $D_{33}u(\omega_t, x_t, x_t^i) = \kappa(x_t^i)$, that is, independent of ω_t, x_t. This is certainly true for any quadratic utility function $u(\omega_t, x_t, x_t^i)$ such as that in the Pigouvian example.

4.2. Cournot

We check that the standard Cournot model without entry has utility that is both regular and separable utility. Utility is $u(\omega_t, x_t, x_t^i) = (p(x_t) - \omega_t)\, x_t^i - c(x_t^i)$; so in this context ω_t is a negative shock to demand. We suppose as standard that $p'(x_t) < 0$, $c'(x_t^i) > 0$, $c''(x_t^i) < 0$, and for a monopolist the objective $(p(x_t) - \omega_t)\, x_t - c(x_t)$ is strictly concave. We list below our assumptions about the externality and show that they are satisfied in the standard Cournot model: for comparison we show also that they are satisfied in Pigouvian case where $\alpha\omega_t < 1$.

	Cournot	Pigou
$D_2u(\omega_t, x_t, x_t) < 0$	$p'(x_t)x_t < 0$	$-(1 - \alpha\omega_t) < 0$
$D_{31}u(\omega_t, x_t, x_t^i) < 0$	$-1 < 0$	$-1 < 0$
$D_{23}u(\omega_t, x_t, x_t^i) \leq 0$	$p'(x_t) < 0$	0
$D_{33}u(\omega_t, x_t, x_t^i) = \kappa(x_t^i)$	$-c''(x_t^i)$	$-V$

4.3. Properties of Optimal Norms

Our goal is to study two properties. The first property is the existence of the relevant social norms.

Definition 1. Property (E) is said to hold if for all $\omega_t \geq \omega_1$ a reoptimal social norm (x_t, y_t, P_t) exists, and if for any (x_1, y_1, P_1) optimal with respect to ω_1 and any $\omega_2 \geq \omega_1$ a unique welfare maximizing incentive compatible social norm (x_2, y_1, P_1) exists. The latter is the *status quo social norm.*

The property central to the paper is the upwards jump.

Definition 2. Property (U) is said to hold if property (E) holds and for any (x_1, y_1, P_1) optimal with respect to ω_1 there exists an open interval of ω_2 defined by $\omega_1 < \omega_a < \omega_2 < \omega_b$ such that for any such ω_2 the optimal choice of social norm is the non-cooperative social norm and $x_2^N > x_1$.

Whether these properties hold depend upon the monitoring technology, which we discuss next.

4.4. Monitoring Technology

Recall that in the Pigou example we took monitoring to be represented by a step function: the probability of being caught was $\Pi(x_t^i - y_t) = \pi$ for $x_t^i - y_t \leq 0$ and $\pi_B > \pi$ for $x_t^i > y_t$. This is clearly left insensitive: below the quota the chances of being caught are not reduced. It is not, however, smooth, having a discontinuity at zero. To arrive at sensible generalizations we examine some concrete examples of three different types of error, whose salient features we will try to capture. The first is a *gross error,* which can be thought of as a constant, and as such it is obviously smooth and left insensitive. The second is a *measurement error,* which is the additive error most commonly used in economic models; this can be is smooth and not left insensitive but also vice versa. Finally, there are *secret sales* which as we will see are left insensitive but not smooth, as in the Pigou example.

Let us start with a simple example of enforcing a speed limit with a radar system. *Gross error* is an error that is independent of the speed, for example the wrong car is identified by a license plate reader and the individual who receives the fine is not the person that committed the offense. Let us assume that the probability that the wrong car is observed is $0 < \pi/q < 1$ and the probability wrong car is speeding is $0 < q \leq 1$. Then the probability of gross error is π. As this is constant it is both left insensitive and smooth. In a context where the penalty is a fine which benefits other group members, as may be the case in cartels, members have an incentive to make false accusations about violations by other members, and this can be an additional source of gross error.

Measurement error is the usual type of error considered in economic models: if the actual speed is x_t^i the observed speed is $x_t^i + \eta_t$ where the random error η_t is independent of y_t and normal with mean 0. The radar system reports the driver if the observed speed exceeds a threshold y_t. That is, a bad signal is received if $\eta_t > -\left(x_t^i - y_t\right)$. If there is no gross error the bad signal occurs with probability $1 - H(-(x_t^i - y_t))$ where $H(\eta_t)$ is the normal cdf. The overall probability of a bad signal is $\Pi(x_t^i - y_t) = \pi + (1 - \pi/q)\left(1 - H(-(x_t^i - y_t))\right)$. In particular $\Pi(x_t^i - y_t)$ is smooth, but it is not left insensitive. A crucial consideration is this: with smooth measurement error incentive compatible output x_t need not be equal to the quota y_t; and since x_t is chosen endogenously the "quota" itself is really just a shifter of the signalling technology. That is in $\Pi(x_t^i - y_t)$ changing y_t simply changes the benchmark against which x_t^i is measured.

With this in mind, suppose as another example that measurement error η_t is uniform on $[-\gamma, \gamma]$ for $\gamma > 0$. Since y_t is simply a benchmark, we assume that here the radar system reports the driver if the the observed speed $x_t^i + \eta_t$ exceeds a threshold $y_t + \gamma$. Letting $h = x_t^i - y_t$, then $\Pi(h)$ is continuous: it is constant and equal to π for $h \leq 0$, it is linear for $0 \leq h \leq 2\gamma$ and it is constant

and equal to $\pi + 1 - \pi/q$ for $h > 2\gamma$. In this case measurement error is left continuous but not smooth (the derivative is discontinuous at 0 and at 2γ).

Finally, we consider *secret sales*. A natural way to enforce a quota is to require transparency: that output or sales be done in such a way that they are easily observed. This is a common rule in cartels.[17] If a member adheres to the quota there is no reason not to comply with the transparency requirement. On the other hand if a member wishes to violate the quota then they will try to conceal their sales in order to avoid being punished. Hence the key monitoring problem is to determine whether or not secret sales took place. This naturally gives rise to left insensitivity: if the quota is adhered to no secret sales are made and negative signals reflect only gross errors, for example, false or mistaken accusations of making secret sales. If the quota is violated then secret sales take place: if a member is engaging in under-the-table transactions there is a chance word will leak out and they will be detected. The simple Pigou monitoring technology is an example. Generally speaking, however, we would expect that the more secret sales take place, the greater the chance of getting caught. We would also expect that there would be diminishing returns: as secret sales increase the chances of being caught increase at a decreasing rate. Secret sales are given by $h = x_t^i - y_t$ if this is positive, so we can model the probability of being caught by a function $H(h)$ which is zero for $h \leq 0$, that jumps up at zero as there is some chance that word leaks out about under-the-table dealings, and is increasing and concave for $h > 0$ to reflect the increased chance of getting caught with diminishing returns. Allowing for gross error, the overall monitoring technology is then

$$\Pi(h) = \pi + (1 - \pi/q) \cdot \mathtt{1}\{h > 0\} H(h).$$

Notice that this satisfies the property of left insensitivity but is not smooth.

We now want to make a general assumption about $\Pi(h)$ that captures these examples. We have seen concavity for positive values, possibly left insensitivity, and possibly discontinuity or non-differentiability at zero. Notice that unless $\Pi(h)$ is constant it cannot be either concave or convex since no non-constant function bounded below on the real line is concave and no non-constant function bounded above on the real line is convex. Indeed, the boundaries force in a certain sense convexity to the left and concavity to the right. The simplest assumption consistent with this is that there is a single inflection point: that to the left of the inflection point $\Pi(h)$ is convex and to the right concave. This corresponds to a measurement error that has a single-peaked density. We slightly weaken the single inflection point assumption to allow for the piecewise linearity seen in the uniform measurement error example. Specifically:

Definition 3. We say *monitoring is regular* if $h = 0$ is the smallest number for which $\Pi(h)$ is concave to the right and for $h \leq 0$ we have $\Pi(h)$ smooth and weakly convex while for $h > 0$ it is smooth and weakly concave. We do not assume that the function is differentiable or even continuous at 0; we do assume that for $h > 0$ we have $\Pi(h) > \Pi(0)$ and that $\Pi(h) = \Pi^+(h)$ which

[17] See, for example, Genesove and Mullin (2001).

is a smooth, weakly concave function. [18] Finally, we define $\pi = \lim_{h \to -\infty} \Pi(h)$ and, this is crucial, require that $\pi > 0$.

We assume from this point on that monitoring is regular.

In analyzing the mechanism design problem a key role is played by the *monitoring cost* function $M(x_t) \equiv \psi \min_{y_t, P} P\Pi(x_t - y_t)$ subject to the incentive constraint that

$$u(\omega_t, x_t, x_t) - P\Pi(x_t - y_t) \geq u(\omega_t, x_t, x_t^i) - P\Pi(x_t^i - y_t)$$

for all $0 \leq x_t^i \leq X$. With this function we can formulate the re-optimization problem as maximizing $u(\omega_t, x_t, x_t) - M(x_t)$. Unfortunately, even if $\Pi(h)$ is smooth $M(x_t)$ is not a particularly pleasant object: since $\Pi(h)$ cannot be convex $M(x_t)$ is not in general convex either, so that $u(\omega_t, x_t, x_t) - M(x_t)$ is not in general concave or even single-peaked. Never-the-less we will establish that several of the key results from the quadratic Pigou model carry over to the general model.

We next formally define the key property of left insensitivity:

Definition 4. We say that regular monitoring is *left insensitive* if for $h \leq 0$ we have $\Pi(h) = \pi$.

Observe that left insensitivity does not require that $\Pi(h)$ be discontinuous at zero, but does require (via concavity on the right) that the derivative be discontinuous at zero. Thus smoothness of $\Pi(h)$ precludes left insensitivity. The Pigou monitoring was regular and left insensitive.

We indicated above that with a smooth monitoring technology x_t need not be equal to y_t. However for regular monitoring the solution y_t to the monitoring cost problem is never smaller than x_t, from the Appendix we have:

Lemma 1. *Any monitoring cost minimizing y_t satisfies $y_t \geq x_t$. Moreover, in the left insensitive case the monitoring cost $M(x_t)$ is non-increasing.*

This says that the incentive constraint on x_t forces the choice of monitoring technology y_t to lie to the right of it. That is, the solution lies in the convex part of the Π function.

4.5. Monitoring and Output

Our first result does not depend on the assumption that F is large. We let $\mathring{x}_2(\omega_2)$ denote the optimal second period output (which recall can be status quo, re-optimal or non-cooperative for different values of ω_2). Propositions 1 to 3 of the Appendix imply

Theorem 3. *If monitoring is left insensitive or utility separable*[19] *then property (E) holds. Moreover, if x_1 is optimal first period output there is $\omega_1 < \omega_a \leq \omega_b < \overline{\omega}$ such that if $\omega_2 > \omega_b$ then $\mathring{x}_2 < x_1$ and for generic ψ_0 if $\omega_1 \leq \omega_2 \leq \omega_a$ then $\mathring{x}_2 \leq x_1$.*

[18] The auxiliary function Π^+ is needed because the right derivative of Π is undefined at zero when there is a discontinuity there. By $\Pi'(h)$ we always mean the left derivative as this is always well-defined.

[19] It is only for this existence result that separability is needed. While that property holds in our examples, it is far from necessary for existence, and we discuss this further in the Online Appendix.

This theorem leaves open the possibility of a gap between ω_a and ω_b where output is greater than x_1. And indeed our main result is that if there is enough left insensitivity and high bargaining costs F then it is necessarily the case, that is, property (U) holds.

Theorem 4. *If monitoring is left insensitive then property (U) holds. Moreover in the low intervention case of Theorem 3 where $\omega_1 \leq \omega_2 \leq \omega_a$ then there is a right neighborhood of ω_1 where $\mathring{x}_2 = x_1$.*

Proof. Left insensitivity forces $x_1 = y_1$. Indeed Lemma 1 shows that in general cost minimization forces $x_1 \leq y_1$; and in the left insensitive case, if $x_1 < y_1$ then for violations $x_1 < x_1^i < y_1$ the punishment probability does not increase so incentive compatibility fails.

Proposition 2 and Lemma 3 in the Appendix establish that with left insensitivity the status quo does not change as long as the non-cooperative equilibrium lies to the right (that is $x_2^S = x_1$ for $D_3u(\omega_2, x_1, x_1) \geq 0$). This shows that there is a range $\omega_1 \leq \omega_2 \leq \omega_a$ where $\mathring{x}_2 = x_1$ as asserted. Indeed the status quo is better than non-cooperative by continuity because it is strictly better at ω_1, and better than reoptimizing for F large enough.

Next from Lemma 1 $M(x_t)$ is non-increasing, which implies $x_1 \geq x_1^f$. Hence using the assumptions $D_3u(\overline{\omega}, x_1^f, x_1^f) \leq 0$ and $D_{31}(\omega_t, x_1, x_1) < 0$ we may define ω^{SN} as the unique solution to $D_3u(\omega^{SN}, x_1, x_1) = 0$. Because monitoring cost at the status quo is strictly positive the status quo is strictly worse than the non-cooperative social norm at ω^{SN}; and since utility from both the status quo and non-cooperative social norms are continuous in ω_2 it follows that there is a left neighborhood of ω^{SN} in which the non-cooperative social norm is strictly better. Finally, observe that for $\omega_2 < \omega^{SN}$ we have $x_2^N > x_1$ (from $D_{31}u(\omega_t, x_t, x_t^i) < 0$). This establishes property (U). □

4.6. Smooth Monitoring

Left insensitivity is inconsistent with $\Pi(h)$ being smooth. On the other hand smoothness arises naturally with measurement errors, so it is not a case we can dismiss. To focus thinking consider a regular $\Pi(h)$ and the family of monitoring technologies $\Pi(h/\sigma)$. For small σ this amounts to a "small" additive error. In the limit with small additive error and fixed gross error we approach a step-function technology, that is a model with left insensitivity. We would like to know that our result, property (U) in particular, is robust to $\sigma > 0$. We extend the idea of $\Pi(h/\sigma)$ with small σ in the following way.

Definition 5. We say that $\Pi^n \to \Pi$ if

(1) Π^n and Π are regular, Π^n is smooth, and Π is left insensitive and discontinuous

(2) $\Pi^n(h) > \Pi(h)$ for $h < 0$ and $\Pi^n(h) < \Pi(h)$ for $h > 0$ and

(3) for all $\epsilon > 0$ the functions Π^n converge uniformly to Π on the set $|h| \geq \epsilon$.

Part (1) says that the target is a left insensitive discontinuous monitoring technology such as that in the Pigou section, in the secret sales example, or as in the case of the limit of $\Pi(h/\sigma)$. Part

(2) says that Π^n is a noisier technology than Π and since it is smooth can be thought of as $\Pi(h)$ plus an additive error with a continuous density. Part (3) says that the additive error is "small."[20]

Theorem 5. *If F is large, ψ_0 not too large, utility is separable and $\Pi^n \to \Pi$ then for sufficiently large n property (U) holds.*

This result is proven as Theorem 8 in the Appendix. It follows from the left insensitive case Theorem 4 and the following approximation theorem showing that convergence of the monitoring technology implies convergence of the monitoring cost:

Theorem 6. *Suppose that utility is separable and $\Pi^n \to \Pi$, $x_1^n \to x_1 < x_1^N$. If y_1^n is cost minimizing then $y_1^n \to x_1$, $\Pi^n(x_1^n - y_1^n) \to \pi$, $(\Pi^n)'(x_1^n - y_1^n) \to \overline{\Pi} > 0$ and finite and the monitoring cost $M^n(x_1^n) \to M(x_1)$.*

This is proven as Theorem 7 in the Appendix. The slope condition $(\Pi^n)'(x_1^n - y_1^n) \to \overline{\Pi}$ highlights how the smooth Π^n is different than Π even for very large n. With small additive error it is not a good idea to have punishment increasing very rapidly with respect to small violations of the social norm x_1^n: this would lead to frequent "accidental" and costly punishments. Rather at the social norm punishment should initially be somewhat forgiving to avoid large punishments for small errors. Notice that with smooth Π^n it is necessary that the punishment satisfy the first order condition, that is $P^n(\Pi^n)'(x_1^n - y_1^n) = D_3u(\omega_1, x_1^n, x_1^n)$. This shows how, in a certain sense, the problem with a smooth monitoring technology is harder than with left insensitivity: with a left insensitive monitoring technology the designer need worry only about deviations to higher output, and in the discontinuous case, only deviations to substantially higher output. By contrast with a smooth monitoring technology the designer must not choose the punishment too high because doing so would encourage individuals to deviate to lower output. This highlights a sense in which left insensitivity is desirable - an increasing probability of punishment for $h < 0$ simply makes the mechanism design problem harder.

The fact that the first order condition must be exactly satisfied with a smooth monitoring technology has a second consequence described in Lemma 3 in the Appendix. It means that when $\omega_2 > \omega_1$, holding fixed y_1^n, P_1^n, since $D_3u(\omega_2, x_1^n, x_1^n) < D_3u(\omega_1, x_1^n, x_1^n)$ the status quo equilibrium must shift to the left - it is no longer constant as it is with a left insensitive monitoring technology. In other words for larger ω_t output in the status quo social norm declines: this means that it may remain better than the non-cooperative norm regardless of the size of the intervention, and even if there is a switch to the non-cooperative norm the increase in output may not be enough to raise output above x_1.[21] The key to proving Theorem 5 is to show that when there is "near" left insensitivity these things do not happen.

[20] An alternative would be to assert that for $h \neq 0$ we have $\Pi^n(h) \to \Pi(h)$, that is, pointwise convergence. In fact because the functions in question are monotone and bounded this is equivalent to part (3). The uniform condition obviously implies the pointwise condition. That the converse is true is a technical fact outside the scope of this paper, but see Levine and Mattozzi (2019).

[21] We are grateful to a referee who pointed this out in more or less these words.

5. Subsidies and Public Goods

From the point of view of behavior, subsidizing a public good is not different from taxing a negative externality. In this case x represents a reduction in the quantity of public good provided with $x = X$ being zero provision of the public good and $x = 0$ maximal provision of the public good. Increasing ω_2 corresponds to increasing the subsidy. Hence Theorems 4 and 5 give conditions under which a subsidy will lead to increased "non-provision" of the public good x, which is to say reduced provision. In the case of foreign aid, it is sometimes asserted that subsidies provided by foreign governments and NGOs do exactly this. A good case study is Bano (2012), based on extensive fieldwork in Pakistan complemented by survey data.

Bano (2012) examines public goods that were provided through voluntary efforts with socially provided incentives for contribution. These public goods were primarily welfare related and ranged from health care and education to the defense of political rights. She conducted a detailed study of three organizations, the People's Rights Movement (a political organization), the Edhi Foundation (the largest welfare organization in Pakistan), and the Jamiat ul Uloom al-Shariah, a madrasa that provides a free Islamic education to four hundred students. She documents that volunteers provided public goods not because of altruism or self-signalling but in response to an informal system of social incentives. As in our model this is based on monitoring: examples include informal observation of which ambulance service delivered most frequently, and more formal systems such as the use of receipts to monitor donations. Incentives were social in nature: those who were thought not to pull their weight received less respect and were less likely to be invited to social events such as weddings. As can be seen the narrative fits our model.

Subsequently donor organizations attempted to increase public good provision through subsidies in the form of salaries to contributors. In Bano (2012)'s case studies this led to the unraveling of the provision of social incentives and to decreased provision of the public good. She first documents this for four voluntary organizations. In one case she indicates that "[t]he Maternity and Child Welfare Association... almost collapsed with the influx of such aid." Similarly six community based organizations in Sindh engaging primarily in charity and welfare saw a substantial decrease in provision following the arrival of aid from Oxfam. Finally she discusses the collapse of the Asthan Latif Welfare Trust after the arrival of UNICEF aid. In each case she demonstrates that the reduction in public good provision came about because monitoring and social incentives were abandoned in response to formal incentives and that in the absence of these social incentives volunteer effort dried up.

The bottom line is that Bano (2012)'s evidence fits our model. A public good was provided with social incentives (our status quo social norm). A subsidy was introduced (an unanticipated increase in ω_2) and the social incentives ended (reversion to our non-cooperative social norm) and public good provision declined - as our model predicts.

6. Conclusion

We have studied self-organization by groups to overcome externalities. We find that unanticipated interventions may have counter-intuitive consequences. In particular, adverse circumstances may cause output to go up rather than down when an existing social norm is abandoned and non-cooperative behavior takes its place. Never-the-less this may increase welfare.

We identify three conditions under which output increases rather than decreases. First, bargaining cost should be high so that when an intervention takes place it is not worth reaching a new agreement. Second, the intervention must be of intermediate size. If the intervention is small it is not worth changing the existing social norm; if it is large even the non-cooperative output will be less than the original quota. Third, monitoring cost should exhibit approximate left insensitivity. This means that as long as the status quo is preserved output changes little so that a switch to the non-cooperative norm leads to an increase in output.

Finally our model has a message for field experiments: it is practical and important to assess existence of social norms. The presence and role of self organized enforcement before and after an intervention can be ascertained either by direct observation or by survey. Without such information we cannot be certain about the policy implications of the response to an intervention.

References

Abreu, D., D. Pearce and E. Stacchetti (1990): "Toward a theory of discounted repeated games with imperfect monitoring," *Econometrica*: 1041-1063.

Bano, Masooda (1973): *Breakdown in Pakistan : how aid is eroding institutions for collective action,* Stanford University Press

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

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.

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

Boyer, Pierre C., Thomas Delemotte, Germain Gauthier, Vincent Rollet and Benoît Schmutz (2019): "Les déterminants de la mobilisation des "gilets jaunes", Working Papers 2019-06, Center for Research in Economics and Statistics.

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

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.

Chari, V. V. and L. E Jones (2000): "A reconsideration of the problem of social cost: Free riders and monopolists," *Economic Theory* 16: 1-22.

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

Dell, Melissa, Nathan Lane, and Pablo Querubin (2918): "The historical state, local collective action, and economic development in Vietnam," *Econometrica* 86: 2083-2121.

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

Ergin, H., and T. Sarver (2010): "A unique costly contemplation representation," *Econometrica* 78: 1285-1339.

Feddersen, T. , and A. Sandroni (2006): "A theory of participation in elections," A*merican Economic Review* 96: 1271-1282.

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

Joseph E. Harrington, Joseph E. and Andrzej Skrzypacz (2011): "Private Monitoring and Communication in Cartels: Explaining Recent Collusive Practices", A*merican Economic Review* 101: 2425–2449

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.

Genesove, D. and Mullin, W. P. (2001): "Rules, communication, and collusion: Narrative evidence from the Sugar Institute case," *American Economic Review* 91: 379-398.

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.

Green, E. J. and Porter, R. H. (1984): "Noncooperative collusion under imperfect price information," *Econometrica*: 87-100.

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

Levenstein, M. C. and Suslow, V. Y. (2006): "What determines cartel success?" *Journal of Economic Literature* 44: 43-95.

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

Levine, David and Andrea Mattozzi (2019): "Success in Contests," mimeo 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

Modica, S. and A. Rustichini (1994): "Awareness and partitional information structures," *Theory and Decision* 37: 107-124.

Mufson, S. and W. Englund (2020): "Oil Price War Threatens Widespread Collateral Damage," *Washington Post*, March 9.

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.

Rotemberg, J. and G. Saloner (1986): "A supergame-theoretic model of price wars during booms," *American Economic Review* 76: 390-407.

Sims, C. A. (2003): "Implications of Rational Inattention," *Journal of Monetary Economics* 50: 665-690.

Tirole, J. (2009): "Cognition and incomplete contracts." *American Economic Review* 99(1): 265-94.

Townsend, Robert M. (1994): "Risk and Insurance in Village India," *Econometrica* 62: 539-539.

Appendix: The General Model

Regularity of utility and monitoring technology is assumed throughout this appendix.

Cost Minimization and the Reoptimal Social Norm

Lemma 2. *For fixed ω_t,$x_t \leq x_t^N$ the function $\psi\Pi(x_t - y_t)P_t$ has a minimum M_t over incentive compatible (x_t, y_t, P_t) and M_t is lower semi-continuous in ω_t, x_t.*

Proof. Recall that incentive compatibility is given by the constraint

$$u(\omega_t, x_t, x_t) - P\Pi(x_t - y_t) \geq u(\omega_t, x_t, x_t^i) - P\Pi(x_t^i - y_t).$$

First we show that for fixed $x_t \leq x_t^N$ the set of incentive compatible (x_t, y_t, P_t) is not empty. To this end define $H \equiv \sup\{h \leq 0 | \Pi'(h) = 0\}$ (possibly $-\infty$). We have three cases depending on H. Note that the assumption of regular utility implies that for $x_t \leq x_t^N$ we have $D_3u(\omega_t, x_t, x_t) \geq 0$.

First $H = -\infty$. In this case take $y_t = X$. The objective $u(\omega_t, x_t, x_t^i) - P\Pi(x_t^i - y_t)$ is then smooth and concave in x_t^i so it is sufficient for a feasible solution that $D_3u(\omega_1, x_t, x_t) = P\Pi'(x_t - y_t)$. As by construction $\Pi'(x_t - y_t) \neq 0$ take $P = D_3u(\omega_t, x_t, x_t)/\Pi'(x_t - y_t)$.

Second $H = 0$. Take $y_t = x_t$; then we only have to check incentive compatibility for $x_t^i > x_t$. If Π is continuous at zero we must have $\Pi^{+\prime}(0) > 0$ because Π is constant for $h \leq 0$, $\Pi(h) > \Pi(0)$ and is concave for $h > 0$. The incentive constraint for rightward deviations may be written as

$$P \geq \frac{u(\omega_t, x_t, x_t^i) - u(\omega_t, x_t, x_t)}{\Pi(x_t^i - y_t) - \Pi(0)}.$$

In the discontinuous case the right hand side is clearly bounded. In the continuous case this is also true as there is a finite limit at $x_t^i \downarrow y_t$ of $D_3u(\omega_t, x_t, x_t)/\Pi^{+\prime}(0)$. Hence it is possible to choose P sufficiently large that the constraint is satisfied for all x_t^i.

Third $-\infty < H < 0$. Choose y_t such that $H < x_t - y_t < 0$. As in the second part, choose P sufficiently large that the rightward deviation constraint is satisfied, and we can do so such that this is true for all $H \leq x_t - y_t \leq 0$. The leftward constraint is

$$u(\omega_t, x_t, x_t) - u(\omega_t, x_t, x_t^i) \geq P\left(\Pi(x_t - y_t) - \Pi(x_t^i - y_t)\right)$$

and dividing both sides by $x_t - x_t^i$ we get $D_3u(\omega_t, x_t, x_t) \geq P\Pi'(x_t - y_t)$ in the limit. Consider then $D_3u(\omega_t, x_t, x_t)/\Pi'(x_t - y_t)$. As $y_t \uparrow x_t - H$ we have $\Pi'(x_t - y_t) \to 0$ continuously. Hence for some such y_t we have $D_3u(\omega_t, x_t, x_t)/\Pi'(x_t - y_t) \geq P$ so leftward deviation is unprofitable.

Next we show that the set of incentive compatible $(\omega_t, x_t, y_t, P_t)$ is closed, or equivalently that the correspondence $(\omega_t, x_t) \mapsto (y_t, P_t)$ has the closed graph property. This directly implies existence of $M_t(x_t)$ and since Π is itself lower semi-continuous that M_t is lower semi-continuous in ω_t, x_t.

To show the set is closed observe that if Π were continuous this would be immediate. However Π may be discontinuous at 0. Let $(\omega_t^n, x_t^n, y_t^n, P_t^n)$ be an incentive compatible sequence converging

to $(\omega_t, x_t, y_t, P_t)$. Suppose first that $x_t = y_t$. Since the only discontinuity is at 0 fixing x_t^i it follows from incentive compatibility that $u(\omega_t, x_t, x_t) - P_t \limsup \Pi(x_t^n - y_t^n) \geq u(\omega_t, x_t, x_t^i) - P_t\Pi(x_t^i - y_t)$. Since Π can jump down but not up (lower semi-continuity) we also have $\limsup \Pi(x_t^n - y_t^n) \geq \Pi(x_t - y_t) = \Pi(0)$. Hence also $u(\omega_t, x_t, x_t^i) - P_t\Pi(x_t - y_t) \geq u(\omega_t, x_t, x_t^i) - P_t\Pi(x_t^i - y_t)$, the desired result.

Now suppose that $x_t \neq y_t$. A deviation $x_t^i \neq y_t$ cannot be profitable by continuity. For $x_t^i = y_t$ there are two cases depending on whether we choose a subsequence converging from the right or from the left. If $y_t^n \downarrow x_t^i$ the result is implied by left continuity of Π. Finally, if $y_t^n \uparrow x_t^i$ and $u(\omega_t, x_t, x_t) - P_t\Pi(x_t - y_t) < u(\omega_t, x_t, x_t^i) - P_t\Pi(x_t^i - y_t)$ then with $\tilde{x}_t^{in} \equiv x_t^i - 2(x_t^i - y_t^n)$ we have $\tilde{x}_t^{in} - y_t^n = y_t^n - x_t^i < 0$ so by left continuity of Π and sufficiently large n we have $u(\omega_t^n, x_t^n, x_t^n) - P_t\Pi(x_t^n - y_t^n) < u(\omega_t^n, x_t^n, \tilde{x}_t^{in}) - P_t\Pi(\tilde{x}_t^{in} - y_t^n)$ contradicting the fact that we assumed $u(\omega_t^n, x_t^n, x_t^n) - P_t\Pi(x_t^n - y_t^n) \geq u(\omega_t^n, x_t^n, x_t^i) - P_t\Pi(x_t^i - y_t^n)$ for any x_t^i. □

Proposition 1. *For all $\omega_t \geq \omega_1$ a reoptimal social norm (x_t, y_t, P_t) exists and the set of all such norms is closed.*

Proof. Follows directly from objective function $u(\omega_t, x_t, x_t) - M_t(x_t)$ being upper semi-continuous given in Lemma 2. □

Lemma (Lemma 1 in the text)**.** *Any cost minimizing y_t satisfies $y_t \geq x_t$. Moreover, in the left insensitive case the monitoring cost $M(x_t)$ is non-increasing.*

Proof. Suppose that $y_t < x_t$ is cost minimizing. From the incentive compatibility of social norms it follows that $u(\omega_t, x_t, x_t) - P\Pi(x_t - y_t) = \max_{x_t^i} \left[u(\omega_t, x_t, x_t^i) - P\Pi(x_t^i - y_t)\right]$ so that the necessary first order condition $D_3u(\omega_t, x_t, x_t) = P_t\Pi'(x_t - y_t)$ must be satisfied. It is tempting to observe that for x_t to the right of y_t the objective function $\psi\Pi(x_t - y_t)D_3u(\omega_t, x_t, x_t)/\Pi'(x_t - y_t)$ is decreasing in y_t but this is not helpful since all values of y_t may not be feasible. We show how to construct a $\hat{y}_t \geq x_t$ that satisfies incentive compatibility and has strictly lower cost than y_t. Specifically, if $\Pi'(0-) > \Pi'(x_t - y_t)$ we can find a $\hat{y}_t > x_t$ with $D_3u(\omega_t, x_t, x_t) = P_t\Pi'(x_t - \hat{y}_t)$ (since Π is smooth and bounded below at 0) and if $\Pi'(0-) \leq \Pi'(x_t - y_t)$, take $\hat{y}_t = x_t$. We keep the punishment fixed at P_t. Since necessarily $\Pi(x_t - \hat{y}_t) < \Pi(x_t - y_t)$, monitoring cost $\psi P_t\Pi(x_t - \hat{y}_t) < \psi P_t\Pi(x_t - y_t)$ is strictly lower. It remains to show that $\hat{y}_t$ is in fact incentive compatible. Consider $x_t^i \leq x_t$, so in particular $x_t^i \leq \hat{y}_t$. Then the objective function $u(\omega_t, x_t, x_t^i) - P_t\Pi(x_t^i - \hat{y}_t)$ is concave for $x_t^i - \hat{y}_t \leq 0$ and the first order condition is satisfied at $x_t^i = x_t$ so there can be no profitable deviation to the left.

Consider a deviation to the right $x_t^i > x_t$. Since y_t was incentive compatible we have $u(\omega_t, x_t, x_t) - P_t\Pi(x_t - y_t) \geq u(\omega_t, x_t, x_t^i) - P_t\Pi(x_t^i - y_t)$. We would like to show that the same holds for $\hat{y}_t$. A sufficient condition is $\Pi(x_t^i - \hat{y}_t) - \Pi(x_t - \hat{y}_t) \geq \Pi(x_t^i - y_t) - \Pi(x_t - y_t)$. Write $\xi_1 = \min\{x_t^i - x_t, \hat{y}_t - y_t\}$ and $\xi_2 = \max\{0, x_t^i - x_t - (\hat{y}_t - y_t)\}$ observing that $x_t^i - x_t = \xi_1 + \xi_2$. For $\tilde{y}_t \in \{y_t, \hat{y}_t\}$ we have

$$\Pi(x_t^i - \tilde{y}_t) - \Pi(x_t - \tilde{y}_t) \geq \int_0^{\xi_1 + \xi_2} \Pi'(h + x_t - \tilde{y}_t)dh.$$

For $\tilde{y}_t = \hat{y}_t$ the inequality may be strict if Π jumps up at zero; for $\tilde{y}_t = y_t$ this must hold with equality. Hence the sufficient condition will follow from $\Pi'(\xi + x_t - \hat{y}_t) \geq \Pi'(\xi + x_t - y_t)$ for almost all $0 \leq \xi \leq \xi_1 + \xi_2$.

If $\hat{y}_t > x_t$ and $\xi < \hat{y}_t - x_t$ we have $\Pi'(\xi + x_t - \hat{y}_t) \geq \Pi'(x_t - \hat{y}_t) = \Pi'(x_t - y_t)$ because $\Pi(h)$ is convex for $h < 0$. For $0 \leq \hat{y}_t - x_t < \xi \leq \xi_1$ we have $\Pi'(\xi + x_t - \hat{y}_t) \geq \Pi'(x_t - y_t)$ because $\Pi(h)$ is concave for $h > 0$. Finally, $\Pi'(x_t - y_t) \geq \Pi'(\xi + x_t - y_t)$ because $x_t - y_t > 0$. This shows that $\Pi'(\xi + x_t - \hat{y}_t) \geq \Pi'(\xi + x_t - y_t)$ for $0 \leq \xi \leq \xi_1$. If $\xi_2 > 0$, then for $\xi > \xi_1$ we have $\xi \geq \hat{y}_t - y_t$ so that $\xi + x_t - \hat{y}_t \geq x_t - y_t > 0$. Hence Π is concave between $\xi + x_t - \hat{y}_t$ and $\xi + x_t - y_t$ giving $\Pi'(\xi + x_t - \hat{y}_t) \geq \Pi'(\xi + x_t - y_t)$.

To show that left insensitivity implies $M(x_t)$ is non-increasing we can apply the first result together with left insensitivity to see that optimal $x_t = y_t$ and the only incentive constraint is $P\left[\Pi(h_t^i) - \Pi(0)\right] \geq u(\omega_t, x_t, x_t + h_t^i) - u(\omega_t, x_t, x_t)$ for $h_t^i \geq 0$. Moreover, $M(x_t) = P\Pi(0)$ so monitoring cost minimization is punishment minimization. By assumption $u(\omega_t, x_t, x_t + h_t^i) - u(\omega_t, x_t, x_t)$ is decreasing in x_t so it follows that the minimal punishment is non-increasing in x_t which gives the second result. □

The Status Quo Social Norm

Lemma 3. *If monitoring is left insensitive or utility separable then for any (x_1, y_1, P_1) optimal with respect to ω_1 and for any $\omega_2 \geq \omega_1$, in $0 \leq x_2 \leq x_1$ there is a unique incentive compatible social norm $\left(x_2^L, y_1, P_1\right)$. The* left status quo, *x_2^L is either x_1 with $D_3u(\omega_2, x_1, x_1) - P_1\Pi'(x_1 - y_1) > 0$* [22] *or it is the unique solution in $0 \leq x_2 \leq x_1$ of $D_3u(\omega_2, x_2, x_2) - P_1\Pi'(x_2 - y_1) = 0$. $x_2^L \leq x_2^N$ and is decreasing and continuous in ω_2. If Π is smooth and $\omega_2 > \omega_1$ then $x_2^L < x_1$.*

Proof. By Lemma 1 $x_1 \leq y_1$. Hence in $0 \leq x_2 \leq x_1$ the function $\Pi(x_2 - y_1)$ is smooth so any incentive compatible social norm (x_2, y_1, P_1) must satisfy the first order condition. Hence we need only show that a solution exists and is incentive compatible. We have $D_3u(\omega_1, x_1, x_1) - P_1\Pi'(x_1 - y_1) \geq 0$ because it cannot be profitable to deviate to the left at ω_1. If this holds with equality define $\tilde{\omega}_2 = \omega_1$. Otherwise there is a unique value $\tilde{\omega}_2 > \omega_1$ such that $D_3u(\tilde{\omega}_2, x_1, x_1) - P_1\Pi'(x_1 - y_1) = 0$. Hence for $\omega_2 > \tilde{\omega}_2$ we have $D_3u(\omega_2, x_1, x_1) - P_1\Pi'(x_1 - y_1) < 0$, while for $\tilde{\omega}_2 > \omega_2 > \omega_1$ we have $D_3u(\omega_2, x_1, x_1) - P_1\Pi'(x_1 - y_1) > 0$.

Consider $g(\omega_2, x_2) \equiv D_3u(\omega_2, x_2, x_2) - P_1\Pi'(x_2 - y_1)$. We have $D_1g(\omega_2, x_2) = D_{31}u(\omega_2, x_2, x_2) < 0$ and $D_2g(\omega_2, x_2) = D_{32}u(\omega_2, x_2, x_2) + D_{33}u(\omega_2, x_2, x_2) - P_1\Pi''(x_2 - y_1) < 0$. For $\tilde{\omega}_2 \geq \omega_2 \geq \omega_1$ we have $g(\omega_2, x_1) \geq 0$ so the unique solution of the first order condition in $[0, x_1]$ is at x_1. If $\omega_2 > \tilde{\omega}_2$ then $g(\omega_2, x_1) < 0$ so the first order condition is that either $g(\omega_2, x_2) = 0$ or that the derivative should be negative on the left boundary $g(\omega_2, 0) < 0$. It follows from the implicit function theorem that there is a unique solution x_2^L of this first order condition and that this solution is decreasing in ω_2. It follows also that it is weakly smaller than x_2^N since that solves the same problem with $\Pi' \equiv 0$.

[22] We always take $\Pi'(0)$ to be the left derivative.

It remains to establish that $\left(x_2^L, y_1, P_1\right)$ is incentive compatible. It is incentive compatible for deviations x_2^i to the left because of concavity and the first order condition being satisfied. For $\tilde{\omega}_2 \geq \omega_2 \geq \omega_1$ it is incentive compatible to the right because the left status quo is at x_1 and $D_{31}u(\omega_t, x_t, x_t^i) < 0$ implies that the utility gain from deviating to the right is reduced. For $\omega_2 > \tilde{\omega}_2$ there are two cases. In the left insensitive case the first order condition requires that $x_2^L = x_2^N$ which is certainly incentive compatible to the right.

For the case where $\omega_2 > \tilde{\omega}_2$, $x_2^L > 0$ and utility is separable observe that for $\omega_2 \geq \tilde{\omega}_2$ the first order condition holds with equality $D_3u(\omega_2, x_2^L, x_2^L) = P_1\Pi'(x_2^L - y_1)$. From the inverse function theorem$dx_2^L/d\omega_2 < 0$ and by the convexity of $\Pi(h)$ for $h < 0$ we have

$$\begin{aligned} D_3u(\omega_2, x_2^L, x_1) < D_3u(\omega_2, x_2^L, x_2^L) = P_1\Pi'(x_2^L - y_1) \leq P_1\Pi'(x_1 - y_1) = D_3u(\tilde{\omega}_2, x_1, x_1) \\ \leq D_3u(\omega_1, x_1, x_1). \end{aligned}$$

Observe that deviations $x_2^i \leq x_1$ yield no utility gain at ω_2, x_2^L because the objective function is concave and the first order condition is satisfied. Consider $x_2^i > x_1$. We have $u(\omega_2, x_2^L, x_2^i) - u(\omega_2, x_2^L, x_1) = \int_{x_1}^{x_2^i} D_3u(\omega_2, x_2^L, \xi)d\xi$. Since $D_3u(\omega_2, x_2^L, x_1) < D_3u(\omega_1, x_1, x_1)$ separability enables us to conclude that $D_3u(\omega_2, x_2^L, \xi) < D_3u(\omega_1, x_1, \xi)$. Hence

$$\begin{aligned} u(\omega_2, x_2^L, x_2^i) - u(\omega_2, x_2^L, x_1) < \int_{x_1}^{x_2^i} D_3u(\omega_1, x_{1,}\xi)d\xi = u(\omega_1, x_1, x_2^i) - u(\omega_1, x_1, x_1) \\ \leq P_1\left[\Pi(x_2^i - y_1) - \Pi(x_1 - y_1)\right]. \end{aligned}$$

Moreover, $u(\omega_2, x_2^L, x_1) - u(\omega_2, x_2^L, x_2^L) \leq P_1\left[\Pi(x_1 - y_1) - \Pi(x_2^L - y_1)\right]$ so adding up we have $u(\omega_2, x_2^L, x_2^i) - u(\omega_2, x_2^L, x_2^L) \leq P_1\left[\Pi(x_2^i - y_1) - \Pi(x_2^L - y_1)\right]$ which is the required incentive compatibility condition.

In case x_2^L falls to zero at $\omega^* > \tilde{\omega}_2$ then note that the argument above is valid for ω^* and once x_2^L is fixed at zero further increases in ω_2 simply increase incentive compatibility to the right.

Finally, if Π is smooth then the first order condition $D_3u(\omega_1, x_1, x_1) - P_1\Pi'(x_1 - y_1) = 0$ must be satisfied with equality as must $D_3u(\omega_2, x_2^L, x_2^L) - P_1\Pi'(x_2^L - y_1) = 0$, and since $D_3u(\omega_t, x_t, x_t) - P_1\Pi'(x_t - y_1)$ is decreasing in ω_t and in x_t for $x_t \leq y_1$ it follows that $x_2^L < x_1$. □

Proposition 2. *If monitoring is left insensitive or utility separable, for any (x_1, y_1, P_1) optimal with respect to ω_1 and for any $\omega_2 \geq \omega_1$ an incentive compatible social norm (x_2, y_1, P_1) exists and the set of all such social norms is closed so that there is a status quo social norm. For $x_2^L \geq x_2^f$ it is uniquely given by $\left(x_2^L, y_1, P_1\right)$.*

Proof. Lemma 3 establishes existence of an incentive compatible social norm. Any incentive compatible social norm to the right of x_1 and less than y_1 must satisfy the first order condition with equality, hence the set is closed. That in turn implies existence of a status quo social norm (that is optimal within the class of incentive compatible norms). For $x_2^L \geq x_2^f$ observe that any solution to the right of x_2^L has weakly lower social utility since both x_2^L and that solution lie to the right

of x_2^f and the social objective function $u(\omega_2, x_2, x_2)$ is concave, and also has strictly higher monitoring costs since P_1 is fixed and $\Pi(x_2 - y_1)$ must strictly increase in moving from non-negative to positive. □

In the proof of Theorem 4 in the text we assert that the last two propositions imply that if Π is left insensitive and $D_3u(\omega_2, x_1, x_1) \geq 0$ we must have $x_2^S = x_1$. To see this observe that in this case 3 implies that $x_2^L = x_1$, since for any $x_2 < x_1$ the equation $D_3u(\omega_2, x_2, x_2) - P_1\Pi'(x_2 - y_1) = 0$ has no solution ($D_3u(\omega_2, x_2, x_2) > 0$ and $\Pi'(x_2 - y_1) = 0$). Since $x_2^f < x_1$ Proposition 2 then implies $x_2^S = x_2^L$.

Partial Monotonicity

Proposition 3. *If monitoring is left insensitive or utility separable and x_1 is optimal first period output there is $\omega_1 < \omega_a \leq \omega_b < \overline{\omega}$ such that when $\mathring{x}_2$ is a corresponding optimal second period output if $\omega_2 > \omega_b$ then $\mathring{x}_2 < x_1$ and for generic ψ_0 if $\omega_1 \leq \omega_2 \leq \omega_a$ then $\mathring{x}_2 \leq x_1$.*

Proof. Observe that in the second period we must have $\mathring{x}_2 \leq x_2^N$. As $x_2^N \geq x_2^f$ and the social objective function $u(\omega_2, x_2, x_2)$ is concave any $x_2 > x_2^N$ would have no greater social utility and so would be strictly worse than x_2^N if it had any monitoring costs, that is, if it was reoptimal or status quo.

If ψ_0 is so large that the first period solution is non-cooperative simply observe that $\omega_2 \geq \omega_1$ implies $x_2^N \leq x_1^N$ so in fact $\mathring{x}_2 < x_1$ for all $\omega_2 > \omega_1$. The same applies as soon as ω_2 is sufficiently large that $x_2^N < x_1$, giving the ω_b result.

Finally, for ω_2 sufficiently close to ω_1 the fact that $F > 0$ and reoptimal utility is upper semi-continuous in ω_2 implies that it is not optimal to reoptimize. Moreover, as $x_1 > x_1^f$ it follows that, again for ω_2 sufficiently close to ω_1, that $x_2^L > x_2^f$ as both are continuous in ω_2 so the status quo social norm involves no higher output than x_1 by Proposition 2 and Lemma 3. It remains then only to rule out the case where at ω_1 there was indifference between the reoptimal social norm and the non-cooperative social norm. As that can happen for only one value of ψ_0 this is indeed non-generic. □

Limit Monitoring

Lemma 4. *Let v^n, w^n be sequences with $v^n \to v > 0$ and $\liminf w^n > 0$. If $\Pi^n \to \Pi$ and for a sequence $h^n < 0$ it is the case that $(\Pi^n)'(h^n) = w^n/v^n$ then we have $h^n \to 0$ and $\Pi^n(h^n) \to \pi$.*

Proof. First we establish that for all $\epsilon > 0$ the functions $(\Pi^n)'$ converge uniformly to 0 on the set $h < -\epsilon$. This directly implies that $h^n \to 0$. Since $(\Pi^n)'$ is nonnegative and increasing it suffices to show for any $\epsilon > 0$ we have $(\Pi^n)'(-\epsilon) \to 0$. To see this, since $\Pi^n(-\epsilon)$ is convex for $\epsilon > 0$ we may write $\Pi^n(-\epsilon/2) - \pi \geq \Pi^n(-\epsilon) - \pi + (\Pi^n)'(-\epsilon)\epsilon/2$. Since the LHS goes to zero (recall that Π is left insensitive by Definition 5), $\Pi^n(-\epsilon) - \pi$ is non-negative, and $(\Pi^n)'(-\epsilon) \geq 0$ by assumption it follows that $(\Pi^n)'(-\epsilon) \to 0$.

To establish the second assertion, if not then there exists an $\epsilon > 0$ and a subsequence in which $\Pi^n(h^n) > \pi + \epsilon$. Draw the tangent line to Π^n at the point h^n, it has slope w^n/v^n and as Π^n is convex for negative h, it lies below Π^n there. The tangent line intersects the constant function π at $h^n - \gamma^n$ where $\Pi^n(h^n) - \gamma^n (w^n/v^n) = \pi$, which is to say at $\gamma^n = [\Pi^n(h^n) - \pi] v^n/w^n$. Consider then that $\Pi^n(h^n - \gamma^n/2) \geq \pi + \epsilon/2$, and $\Pi^n(-\gamma^n/2) \geq \Pi^n(h^n - \gamma^n/2) \geq \pi + \epsilon/2$. Unfortunately $\lim \gamma^n/2 > 0$ so $\Pi^n(-\gamma^n/2) \to \pi$ which is a contradiction. □

Lemma 5. *If* $\Pi^n \to \Pi$ *then* $(\Pi^n)'(0) \to \infty$.

Proof. This result is driven by $\Pi^+(0) > \pi$, which itself follows from definition 5 part 1. Suppose that there is a subsequence along which $(\Pi^n)'(0)$ is bounded above by Q. We cannot reconcile that with the discontinuity in the limit. To see this, choose a further subsequence along which $\Pi^n(0) \to q$. There are two cases: if $q < \Pi^+(0)$ then pointwise convergence is violated to the right. Specifically, define

$$h = \frac{\Pi^+(0) - q}{3Q}.$$

Then for large enough n we have $\Pi^n(h) - q$ close to $\Pi^n(h) - \Pi^n(0)$ so $\Pi^n(h) - q \leq (1/2)(\Pi^+(0) - q)$, while $\liminf \Pi^n(h) \geq \Pi^+(0)$ a contradiction.

If $q = \Pi^+(0)$ then $q > \pi$ and pointwise convergence is violated to the left. Specifically, define

$$h = \frac{\pi - q}{3Q}.$$

Then for large enough n we have $q - \Pi^n(h) \leq (1/2)(q - \pi)$ and therefore $\pi < (1/2)(q + \pi) \leq \Pi^n(h)$ while $\limsup \Pi^n(h) = \pi$, a contradiction. □

We use these to show that

Theorem 7. *Suppose that* $\Pi^n \to \Pi$, $x_1^n \to x_1 < x_1^N$. *If* y_1^n *is cost minimizing then* $y_1^n \to x_1$, $\Pi^n(x_1^n - y_1^n) \to \pi$, $(\Pi^n)'(x_1^n - y_1^n) \to \overline{\Pi} > 0$ *and finite and the monitoring cost* $M^n(x_1^n) \to M(x_1)$.

Proof. Define $P = M(x_1)/(\psi\pi)$. Since $x_1 < x_1^N$ we have $D_3 u(\omega_1, x_1, x_1) > 0$ which implies in particular that $P > 0$ ($M(x_1)$ must be positive because with no punishment there is strict incentive to deviate), and for large enough n we have $x_1^n < x_1^N$ and $D_3 u(\omega_1, x_1^n, x_1^n) > 0$.

Take $\hat{P} > P$. By Lemma 5 for all n sufficiently large there exists an $h^n < 0$ with $(\Pi^n)'(h^n) > D_3 u(\omega_1, x_1^n, x_1^n)/(P/2)$. Hence we may find a solution $\hat{h}^n$ to $(\Pi^n)'(h^n) = D_3 u(\omega_1, x_1^n, x_1^n)/\hat{P}$ with $\hat{h}^n < 0$. By Lemma 4 (applied to the sequences $v^n = \hat{P}$ and $w^n = D_3 u(\omega_1, x_1^n, x_1^n))$ $\hat{h}^n \to 0$ and $\Pi^n(\hat{h}^n) \to \pi$. We claim in fact that for n sufficiently large $x_1^n, y_1^n \equiv x_1^n - \hat{h}^n, \hat{P}$ is incentive compatible for the monitoring technology Π^n. This norm is not necessarily cost minimizing but will serve the purpose of showing that the minimum cost $M^n(x_1^n) \to M(x_1)$.

Observe first that it cannot be optimal to deviate to $x_1^i \leq y_1^n$ since the objective function is concave and the first order condition is satisfied at x_1^n. Moreover, it cannot be optimal to deviate

to an $x_1^i > y_1^n$ with

$$D_3u(\omega_1, x_1^n, x_1^n)(x_1^i - x_1^n) \leq \hat{P}\left(\Pi^n(x_1^i - y_1^n) - \Pi^n(x_1^n - y_1^n)\right)$$

that is with

$$x_1^i - y_1^n \leq \hat{P}\left(\frac{\Pi^n(x_1^i - y_1^n) - \Pi^n(x_1^n - y_1^n)}{D_3u(\omega_1, x_1^n, x_1^n)}\right) + \hat{h}^n.$$

Now for n sufficiently large the RHS is bounded below by $\hat{\pi} \equiv (1/2)\hat{P}(\Pi^+(0) - \pi)/D_3u(\omega_1, x_1, x_1) > 0$. Hence if there is a profitable deviation it must be to $x_1^i > y_1^n + \hat{\pi}$.

Let x_1^{in} be an optimal deviation in the range $x_1^i \geq y_1^n + \hat{\pi}$. Since $\hat{P} > P$ and $x_1^i \geq y_1^n + \hat{\pi}$ we know - since $\Pi(0) = \pi$ - that

$$u(\omega_1, x_1, x_1^{in}) - u(\omega_1, x_1, x_1) - \hat{P}\left(\Pi(x_1^{in} - x_1) - \pi\right) \leq -\epsilon < 0.$$

Hence since u is uniformly continuous and Π is for $x_1^i \geq y_1^n + \hat{\pi}$ we have

$$\limsup\left[u(\omega_1, x_1^n, x_1^{in}) - u(\omega_1, x_1^n, x_1^n) - \hat{P}\left(\Pi(x_1^{in} - y_1^n) - \Pi(x_1^n - y_1^n)\right)\right] \leq -\epsilon.$$

For $x_1^i \geq y_1^n + \hat{\pi}$ we also know that Π^n converges uniformly to Π so

$$\limsup\left[u(\omega_1, x_1^n, x_1^{in}) - u(\omega_1, x_1^n, x_1^n) - \hat{P}\left(\Pi^n(x_1^{in} - y_1^n) - \Pi^n(x_1^n - y_1^n)\right)\right] \leq -\epsilon.$$

Hence for large enough n the deviation x_1^{in} is not profitable.

Observe that along this sequence monitoring cost is by construction $M_{\hat{P}}^n(x_1^n) = \psi\Pi^n(\hat{h}^n)\hat{P} \to \psi\pi\hat{P}$. Since the least monitoring cost $M^n(x_1^n) \leq M_{\hat{P}}^n(x_1^n)$ and $\hat{P} > P$ was arbitrary we must have $\limsup M^n(x_1^n) \leq \psi\pi P = M(x_1)$. Moreover since the Π^n technology is strictly inferior to the Π technology (part (2) of Definition 5), we have $M^n(x_1^n) \geq M(x_1^n)$ and since M is lower semi-continuous, the fact that for n large enough $M(x_1^n) \leq M^n(x_1^n) \leq M(x_1)$ implies in fact that $M^n(x_1^n) \to M(x_1)$ as asserted. The convergence of $(\Pi^n)'(x_1^n - y_1^n) \to \overline{\Pi} > 0$ follows directly from the convergence of monitoring cost and the first order condition.

For the rest, let P^n be a cost minimizing punishment corresponding to the optimal $\hat{y}_1^n$. Observe that $\Pi^n(x_1^n - \hat{y}_1^n) \geq \pi$. As we have just shown, monitoring cost must converge to $\psi\pi P$, and this implies that P^n is bounded above and away from zero. Hence we may extract a subsequence along which $P^n \to P > 0$. From Lemma 1 we have $x_1^n \leq \hat{y}_1^n$. Since Π is discontinuous at 0 the first order condition $(\Pi^n)'(x_1^n - \hat{y}_1^n) = D_3(\omega_1, x_1^n, x_1^n)/P^n$ implies that in fact $x_1^n < \hat{y}_1^n$ (since otherwise from Lemma 5 the slope would go to infinity). We may then apply Lemma 4 to reach the desired conclusion that $\hat{y}_1^n \to x_1$ and $\Pi^n(x_1^n - y_1^n) \to \pi$. □

Lemma 6. *Suppose that utility is separable and $\Pi^n \to \Pi$, $x_1^n \to x_1$ with $x_1^f \leq x_1 < x_1^N$ and that y_1^n is cost minimizing with corresponding punishment P_1^n. Fix $\omega_2 > \omega_1$ such that $x_2^N > x_1$. Then the status quo $x_2^{Sn} \to x_1$ and $\psi P_1^n \Pi^n(x_2^{Sn} - y_1^n) \to M(x_1)$.*

Proof. It suffices to prove the result for x_2^{Ln} the left status quo: if $x_2^{Ln} \to x_1$ since $x_1 \geq x_1^f > x_2^f$ then Proposition 2 implies that for large enough n the left status quo is the unique status quo.

From the first order condition $P_1^n = D_3u(\omega_1, x_1^n, x_1^n)/\left(\Pi^n\right)'(x_1^n - y_1^n) \to D_3u(\omega_1, x_1, x_1)/\overline{\Pi}$ by Theorem 7. By definition $x_2^{Ln} \leq x_1^n$ so $D_3u(\omega_2, x_2^{Ln}, x_2^{Ln}) \geq D_3u(\omega_2, x_1^n, x_1^n)$. Moreover $D_3u(\omega_2, x_1, x_1) > 0$ (from$x_1 < x_2^N$) and $D_3u(\omega_2, x_1^n, x_1^n) \to D_3u(\omega_2, x_1, x_1)$ so $D_3u(\omega_2, x_2^{Ln}, x_2^{Ln})$ is bounded away from zero. From Lemma 3 we have $x_2^{Ln} < x_1^n \leq y_1^n$ and since $\left(\Pi^n\right)'(x_2^{Ln} - y_1^n) = D_3u(\omega_2, x_2^{Ln}, x_2^{Ln})/P_1^n$ we may apply Lemma 4 to get $x_2^{Ln} \to x_1$ and $\Pi^n(x_2^{Ln} - y_1^n) \to \pi$.

By Lemma 7 we also have $\Pi^n(x_1^n - y_1^n) \to \pi$ and $\psi P_1^n \Pi^n(x_1^n - y_1^n) \to M(x_1)$. Hence $\psi P_1^n \Pi^n(x_2^{Ln} - y_1^n) \to M(x_1)$. □

Definition 6. Property (V) is said to hold at (x_1, y_1, P_1) incentive compatible with respect to ω_1 if there exists an $\omega_2 > \omega_1$ such that $x_2^N > x_1$ and the non-cooperative social norm is strictly better than any status quo social norm (x_2, y_1, P_1).

Lemma 7. *Suppose that utility is separable and* $\Pi^n \to \Pi$, $x_1^n \to x_1$ *with* $x_1^f \leq x_1 < x_1^N$ *and that* y_1^n *is cost minimizing with corresponding punishment* P_1^n. *Then for all sufficiently large n property (V) is satisfied at* (x_1^n, y_1^n, P_1^n).

Proof. Let ω_2^{SN} be the unique solution of $D_3u(\omega_2^{SN}, x_1, x_1) = 0$. Then there exists a $\omega_2^{SN} > \omega_2 > \omega_1$ such that property (V) holds for Π at (x_1, y_1, P_1): the proof is identical to that of Theorem 4. This is to say that $u(\omega_2^{SN}, x_1, x_1) - M(x_1) < u(\omega_2^{SN}, x_2^N, x_2^N)$. From Lemma 6 and the continuity of u the utility from the status quo at n given by $u(\omega_2^{SN}, x_2^{Sn}, x_2^{Sn}) - \psi P_1^n \Pi^n(x_2^{Sn} - y_1^n) \to u(\omega_2^{SN}, x_1, x_1) - M(x_1)$ giving the desired result. □

The next two lemmas show that $x_1^N > \limsup x_1^n \geq \liminf x_1^n \geq x_1^f$. The argument is that this is true in the limit for Π so by Theorem 7 holds for sufficiently large n.

Lemma 8. *Suppose that* $\Pi^n \to \Pi$. *Then for every* ω_1 *and any* (x_1^n, y_1^n, P_1^n) *optimal with respect to* ω_1 *we have* $\liminf x_1^n \geq x_1^f$.

Proof. If not we can find a sequence (x_1^n, y_1^n, P_1^n) optimal with respect to ω_1 with $\lim x_1^n = x_1 < x_1^f$. By Theorem 7 this implies $M^n(x_1^n) \to M(x_1)$ and $M^n(x_1^f) \to M(x_1^f)$. Hence it must be that $u(\omega_1, x_1, x_1) - M(x_1) \geq u(\omega_1, x_1^f, x_1^f) - M(x_1^f)$, and since $u(\omega_1, x_1^f, x_1^f) > u(\omega_1, x_1, x_1)$ that $M(x_1^f) > M(x_1)$. Consider a cost minimizing y_1, P_1 at x_1. Then $y_1 = x_1$ since Π is left insensitive, and for $x_1^i > x_1^f$we have

$$\begin{aligned} P_1\left(\Pi(x_1^i - x_1^f) - \pi\right) &\geq u(\omega_1, x_1, x_1^i - (x_1^f - x_1)) - u(\omega_1, x_1, x_1) \\ &\geq u(\omega_1, x_1^f, x_1^i - (x_1^f - x_1) + (x_1^f - x_1)) - u(\omega_1, x_1^f, x_1^f) \\ &= u(\omega_1, x_1^f, x_1^i) - u(\omega_1, x_1^f, x_1^f) \end{aligned}$$

(the second inequality from $D_{33}u(\omega_t, x_t, x_t) + D_{23}u(\omega_t, x_t, x_t) < 0$) we see that in fact x_1^f, x_1^f, P_1 is incentive compatible. Then $M(x_1^f) > M(x_1) = \psi\pi P_1 \geq M(x_1^f)$ a contradiction. □

Lemma 9. *Suppose that* $\Pi^n \to \Pi$. *Then for every* ω_1 *and any* (x_1^n, y_1^n, P_1^n) *optimal with respect to* ω_1 *we have* $\limsup x_1^n < x_1^N$.

Proof. Certainly $x_1^n \leq x_1^N$ for if $x_1^n > x_1^N$ social welfare would strictly increase and monitoring cost would be at the minimum of zero by moving to x_1^N. Suppose in fact that for some subsequence $\lim x_1^n = x_1^N$. By Lemma 7 this implies $M^n(x_1^n) \to 0$. Letting u_1^{Rn} be the optimal utility we have so $u_1^{Rn} \to u_1^N$.

Now consider Π. Set

$$P_1 = \frac{1}{\Pi^+(0) - \pi} \left(\max_{X \geq x_1^i \geq x_1} u(\omega_1, x_1, x_1^i) - u(\omega_1, x_1, x_1) \right)$$

so that $x_1, y_1 = x_1, P_1$ is incentive compatible. Then the maximum social utility is bounded below by $v(x_1) = u(\omega_1, x_1, x_1) - \psi P_1 \pi$ where $v(x_1^N) = u_1^N$ (at $x_1 = x_1^N$ we have $P_1 = 0$). Moreover $v'(x_1^N) = D_2 u(\omega_1, x_1^N, x_1^N) + D_3 u(\omega_1, x_1^N, x_1^N) < 0$ since $x_1^N > x_1^f$. Hence there is $x_1 < x_1^N$ and $\epsilon > 0$ with $u(\omega_1, x_1, x_1) - M(x_1) > u_1^N + \epsilon$. However by Theorem 7 we have $M^n(x_1) \to M(x_1)$ so for all large enough n it must be that $u(\omega_1, x_1, x_1) - M^n(x_1) > u_1^N + \epsilon/2$ contradicting $u_1^{Rn} \to u_1^N$. □

Theorem 8. *If* F *is large,* ψ_0 *not too large, utility is separable and* $\Pi^n \to \Pi$ *then for sufficiently large* n *property (U) holds.*

Proof. We need only show that property (V) holds for all sufficiently large n and all corresponding optimal first period optimal norms (for F large enough reoptimizing is out of the question): continuity of utility with respect to the status quo and non-cooperative social norms yields the desired range. Hence if the result fails there must be a sequence along which property (V) fails. Then extract a subsequence with x_1^n converging to x_1 and apply Lemmas 8, 9 and 7 to get a contradiction. □

Online Appendix: Pigou

The monitoring technology is $\Pi(x_t^i - y_t) = \pi > 0$ if $x_t^i \leq y_t$ and $\Pi(x_t^i - y_t) = \pi_B > \pi$ if $x_t^i > y_t$. As in the text $\theta = \pi/(\pi_B - \pi)$, and the punishment cost parameters are $\psi_0 = 0$ and $\psi = 1$.

We shall repeatedly use the fact that the solution of a quadratic optimization problem $Ax_t - (B/2)(x_t)^2 = x_t\left[A - (B/2)x_t\right]$ is given by $x_t = A/B$ and the resulting optimum is $A^2/(2B)$.

Individual direct utility is $U(x_t^i) = (V+1)x_t^i - (V/2)(x_t^i)^2$ up to the satiation point $X = (V+1)/V$. Overall individual utility is

$$\begin{aligned} u(\omega_t, x_t, x_t^i) &= U(x_t^i) - \omega_t x_t^i - (1 - \alpha\omega_t)x_t \\ &= (V + 1 - \omega_t)x_t^i - (V/2)(x_t^i)^2 - (1 - \alpha\omega_t)x_t. \end{aligned}$$

The first best x^f is defined as the maximum of

$$\begin{aligned} u(\omega_t, x_t, x_t) &= (V + 1 - \omega_t)x_t - (V/2)(x_t)^2 - (1 - \alpha\omega_t)x_t \\ &= x_t\left[(V - (1-\alpha)\omega_t) - (V/2)x_t\right]. \end{aligned}$$

We always assume $\omega_t \leq 1/\alpha$ which will be seen to imply that $x_t^N \geq x_t^f$ and $\omega_t \leq V/(1-\alpha)$ which will imply that $x_t^f \geq 0$.

Proposition 4. *The first best is $x_t^f = (V - (1-\alpha)\omega_t)/V$ with corresponding welfare $u_t^f = (V - (1-\alpha)\omega_t)^2/(2V)$.*

When $\alpha = 1$ this is the called the *Pigouvian solution* and is $x^P = 1$ with corresponding welfare $u^P = V/2$. The Pigouvian tax is $U'(x^P) = 1$. Note that as indicated for $\omega_t \leq V/(1-\alpha)$ this is non-negative.

Proposition 5. *The individual optimum in the absence of penalty - that is the maximum of $u(\omega_t, x_t, x_t^i)$ with respect to x_t^i - is $x_t^B = (V + 1 - \omega_t)/V$ with utility $u_t^B(\omega_t, x_t, x_t^B) = (V + 1 - \omega_t)^2/(2V) - (1 - \alpha\omega_t)x_t$.*

As the optimum is independent of x_t this is also the noncooperative social norm: $x_t^N = x_t^B$.

Proposition 6. *The noncooperative social norm has $x_t^N = (V + 1 - \omega_t)/V$ with corresponding welfare*

$$u_t^N = u(\omega_t, x_t^N, x_t^N) = \left[\frac{V}{2} - \frac{1}{2V}\right] + \left[\frac{\alpha(V+1) - V}{V}\right]\omega_t + \left[\frac{1 - 2\alpha}{2V}\right]\omega_t^2.$$

For $\omega_t \leq \overline{\omega} \equiv 1/\alpha$ it is $x_t^N \geq x_t^f$.

Proof. The value of u^N is given by direct computation of

$$u(\omega_t, x^N, x^N) = (V + 1 - \omega_t)^2/(2V) - (1 - \alpha\omega_t)(V + 1 - \omega_t)/V.$$

Similar elementary computation gives $x_t^N \geq x_t^f$ if and only if $\omega_t \leq 1/\alpha$. □

Given ω_t and a quota $y_t = x_t$ made incentive compatible by punishment

$$P_t = \left[u(\omega_t, x_t, x_t^B) - u(\omega_t, x_t, x_t)\right]/(\pi_B - \pi)$$

yields monitoring cost πP_t hence social utility

$$u(\omega_t, x_t, x_t) - \theta\left[u(\omega_t, x_t, x_t^B) - u(\omega_t, x_t, x_t)\right].$$

We have denoted by "re-optimal" the norm maximizing this function.

Proposition 7. *The re-optimal social norm has*

$$x_t^R = \frac{(1+\theta)V + \theta - (1+\theta-\alpha)\,\omega_t}{V(1+\theta)}$$

when this is non-negative and the corresponding welfare is

$$\begin{aligned} u_t^R &= \left[\frac{V}{2} - \frac{1}{2V}\frac{\theta}{1+\theta}\right] + \left[\frac{\alpha}{V}\frac{\theta}{1+\theta} - (1-\alpha)\right]\omega_t + \frac{1}{2V}\frac{1}{1+\theta}\left[(1-\alpha)^2 + \theta(1-2\alpha)\right]\omega_t^2 \\ &= u_t^N + \frac{1}{2V}\frac{1}{1+\theta}\left(1-\alpha\omega_t\right)^2. \end{aligned}$$

Proof. The objective function is

$$\begin{aligned} &u(\omega_t, x_t, x_t) - \theta\left[u(\omega_t, x_t, x^B) - u(\omega_t, x_t, x_t)\right] = (1+\theta)u(\omega_t, x_t, x_t) - \theta u(\omega_t, x_t, x^B) \\ =&(1+\theta)x_t\left[(V-(1-\alpha)\omega_t) - (V/2)x_t\right] - \theta\left[(V+1-\omega_t)^2/(2V) - (1-\alpha\omega_t)x_t\right] \\ =&x_t\left[(1+\theta)(V-(1-\alpha)\omega_t) - (1+\theta)(V/2)x_t + \theta(1-\alpha\omega_t)\right] - \theta(V+1-\omega_t)^2/(2V) \\ =&x_t\left[(1+\theta)(V-(1-\alpha)\omega_t) + \theta(1-\alpha\omega_t) - (1+\theta)(V/2)x_t\right] - \theta(V+1-\omega_t)^2/(2V) \\ =&x_t\left[(1+\theta)V + \theta - (1-\alpha+\theta)\,\omega_t - (1+\theta)(V/2)x_t\right] - \theta(V+1-\omega_t)^2/(2V) \end{aligned}$$

whence x_t^R. As to u_t^R we have

$$\begin{aligned} u_t^R &= \frac{\left((1+\theta)V + \theta - (1-\alpha+\theta)\omega_t\right)^2}{2(1+\theta)V} - \frac{\theta(V+1-\omega_t)^2}{2V} \\ &= \frac{1}{2(1+\theta)V}\left[\left((1+\theta)V + \theta - (1-\alpha+\theta)\omega_t\right)^2 - \theta(1+\theta)(V+1-\omega_t)^2\right] \\ &= \frac{1}{2(1+\theta)V}\Big[\left((1+\theta)V+\theta\right)^2 - 2\left((1+\theta)V+\theta\right)(1+\theta-\alpha)\,\omega_t \\ &\qquad + (1+\theta-\alpha)^2\omega_t^2 - \theta(1+\theta)\left((V+1)^2 - 2(V+1)\omega_t + \omega_t^2\right)\Big] \\ &= \frac{1}{2V}\frac{1}{1+\theta}\Big[\left((1+\theta)V+\theta\right)^2 - 2\left((1+\theta)V+\theta\right)(1+\theta-\alpha)\,\omega_t \\ &\qquad + (1+\theta-\alpha)^2\omega_t^2 - \theta(1+\theta)\left((V+1)^2 - 2(V+1)\omega_t + \omega_t^2\right)\Big]. \end{aligned}$$

We examine the constant, linear, and quadratic coefficients separately to get the expression in the

Theorem. The constant term is

$$\begin{aligned}
&\frac{1}{2(1+\theta)V}\left[((1+\theta)V+\theta)^2-\theta(1+\theta)\left(V+1\right)^2\right]\\
=&\frac{1}{2(1+\theta)V}\left[(1+\theta)^2V^2+2\theta(1+\theta)V+\theta^2-(1+\theta)\theta\left(V^2+2V+1\right)\right]\\
=&\frac{1}{2(1+\theta)V}\left[\left((1+\theta)^2-(1+\theta)\theta\right)V^2+\theta^2-(1+\theta)\theta\right]=\frac{1}{2(1+\theta)V}\left[(1+\theta)V^2-\theta\right]\\
=&\frac{V}{2}-\frac{1}{2V}\frac{\theta}{1+\theta}
\end{aligned}$$

The linear coefficient is

$$\begin{aligned}
&\frac{1}{2V}\frac{1}{1+\theta}\left[-2\left((1+\theta)V+\theta\right)(1+\theta-\alpha)+2\theta(1+\theta)(V+1)\right]\\
=&\frac{1}{V}\frac{1}{1+\theta}\left[\left((1+\theta)V+\theta\right)(\alpha-(1+\theta))+\theta(1+\theta)(V+1)\right]\\
=&\frac{1}{V}\frac{1}{1+\theta}\left[\alpha\left((1+\theta)V+\theta\right)-(1+\theta)V\right]=\frac{\alpha\left((1+\theta)V+\theta\right)}{(1+\theta)V}-1\\
=&\frac{\alpha}{V}\frac{\theta}{1+\theta}-(1-\alpha)
\end{aligned}$$

The quadratic coefficient is

$$\begin{aligned}
&\frac{1}{2V}\frac{1}{1+\theta}\left[(1+\theta-\alpha)^2-\theta(1+\theta)\right]\\
=&\frac{1}{2V}\frac{1}{1+\theta}\left[(1-\alpha)^2+\theta(1-2\alpha)\right].
\end{aligned}$$

These give u^R as in the statement. Lastly we compute u^R-u^N. Recall that $u^N=\left[\frac{V}{2}-\frac{1}{2V}\right]+\left[\frac{\alpha(V+1)-V}{V}\right]\omega_t+\left[\frac{1-2\alpha}{2V}\right]\omega_t^2$. The difference in the constants is

$$\frac{V}{2}-\frac{1}{2V}\frac{\theta}{1+\theta}-\left[\frac{V}{2}-\frac{1}{2V}\right]=-\frac{1}{2V}\frac{\theta}{1+\theta}+\frac{1}{2V}=\frac{1}{2V}\frac{1}{1+\theta}$$

For the linear coefficients we have

$$\frac{\alpha}{V}\frac{\theta}{1+\theta}-(1-\alpha)-\left[\frac{\alpha(V+1)-V}{V}\right]=\frac{\alpha}{V}\frac{\theta}{1+\theta}-\frac{\alpha}{V}=-\frac{\alpha}{V}\frac{1}{1+\theta}$$

and for the quadratic

$$\begin{aligned}
&\frac{1}{2V}\frac{1}{1+\theta}\left[(1-\alpha)^2+\theta(1-2\alpha)\right]-\left[\frac{1-2\alpha}{2V}\right]\\
=&\frac{1}{2V}\frac{1}{1+\theta}\left[(1-\alpha)^2+\theta(1-2\alpha)-(1-2\alpha)\left(1+\theta\right)\right]\\
=&\frac{1}{2V}\frac{1}{1+\theta}\cdot\alpha^2
\end{aligned}$$

Therefore

$$\begin{aligned} u^R - u^N &= \frac{1}{2V}\frac{1}{1+\theta} - \frac{\alpha}{V}\frac{1}{1+\theta}\omega_t + \frac{1}{2V}\frac{1}{1+\theta}\alpha^2\omega_t^2 \\ &= \frac{1}{2V}\frac{1}{1+\theta} - \frac{1}{2V}\frac{1}{1+\theta}2\alpha\omega_t + \frac{1}{2V}\frac{1}{1+\theta}\left(\alpha\omega_t\right)^2 \\ &= \frac{1}{2V}\frac{1}{1+\theta}\left(1-\alpha\omega\right)^2 \end{aligned}$$

as claimed. □

We verify a claim made in the text:

Proposition 8. *$x_1^N > x_1 > x_2^N(\omega_2)$ when $\omega_2 = 1/\alpha$*

Proof. The inequality $x_1^N > x_1$ reads

$$\frac{V+1-\omega_t}{V} > \frac{(1+\theta)V + \theta - (1+\theta-\alpha)\,\omega_1}{(1+\theta)V}$$

that is

$$1-\omega_t > \frac{\theta - (1+\theta-\alpha)\,\omega_1}{(1+\theta)}$$

At $\omega_2 = \omega_1$ this is

$$\begin{aligned} -(1+\theta)\omega_1 &> -1 - (1+\theta-\alpha)\,\omega_1 \\ 1 &> (1+\theta)\omega_1 - (1+\theta-\alpha)\,\omega_1 = \alpha\omega_1 \end{aligned}$$

which is true. At $\omega_2 = 1/\alpha$ the reverse inequality $x^N < x_1$ reads

$$\begin{aligned} -\frac{1-\alpha}{\alpha} = 1 - \frac{1}{\alpha} &< \frac{\theta - (1-\alpha+\theta)\,\omega_1}{(1+\theta)} \\ -\frac{1-\alpha}{\alpha} &< \frac{\theta - (1-\alpha+\theta)\,\omega_1}{(1+\theta)} \end{aligned}$$

where the right hand side is larger than

$$\frac{\alpha\theta - (1-\alpha+\theta)}{\alpha(1+\theta)} = \frac{\alpha\theta - (1-\alpha) - \theta}{\alpha(1+\theta)} = -\frac{1-\alpha}{\alpha}$$

so in fact $x_2^N < x_1$ when $\omega_2 = 1/\alpha$. □

Online Appendix: Separability

The sole use of the separability condition is to prove the existence of a status quo social norm when monitoring fails to be left insensitive. However, separability is much stronger than is required for this result, and the purpose of this Appendix is to discuss what is needed.

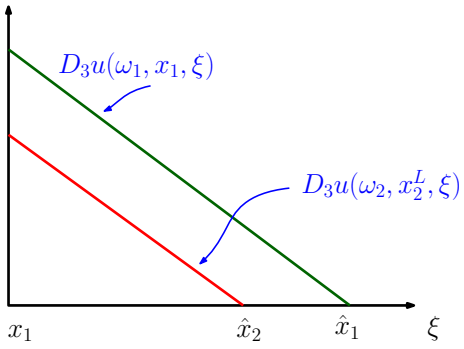
The crucial inequality in the proof (Lemma 3 in the text) is

$$u(\omega_2, x_2^L, x_2^i) - u(\omega_2, x_2^L, x_1) < u(\omega_1, x_1, x_2^i) - u(\omega_1, x_1, x_1)$$

for $x_2^i > x_1$, that is

$$\int_{x_1}^{x_2^i} D_3 u(\omega_2, x_2^L, \xi) d\xi < \int_{x_1}^{x_2^i} D_3 u(\omega_1, x_{1,} \xi) d\xi.$$

It is shown in the proof that $D_3 u(\omega_2, x_2^L, x_1) < D_3 u(\omega_1, x_1, x_1)$ so that the integrand on the left starts below the one on the right. Separability implies that the curves $D_3 u(\omega_1, x_1, \xi)$ and $D_3 u(\omega_2, x_2^L, \xi)$ are parallel, as in the figure below

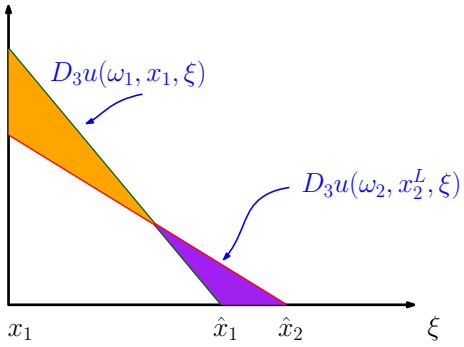


where

$$\hat{x}_1 = \arg\max_{\xi} u(\omega_1, x_1, \xi) \equiv x_1^B \qquad \hat{x}_2 = \arg\max_{\xi} u(\omega_2, x_2^L, \xi).$$

This makes the needed integral inequality clearly valid. On the other hand it is equally clear that this condition is not strictly needed. Indeed if for example $D_{331} u(\omega_t, x_t, x_t^i) \leq 0$ and $D_{332} u(\omega_t, x_t, x_t^i) \geq 0$ then $D_3 u(\omega_2, x_2^L, \xi)$ has a steeper downward slope than $D_3 u(\omega_1, x_1, \xi)$, so this also would suffice.

But even if not, all that is needed is that the yellow area is larger than the purple one in the figure below:



Overall, there is a substantial range and scope of utility functions for which the desired conclusion will hold.