Wahhaj, Zaki

Working Paper

Social Norms, Higher-Order Beliefs and the Emperor's New Clothes

School of Economics Discussion Papers, No. 1210

Provided in Cooperation with:
University of Kent, School of Economics

Suggested Citation: Wahhaj, Zaki (2012) : Social Norms, Higher-Order Beliefs and the Emperor's New Clothes, School of Economics Discussion Papers, No. 1210, University of Kent, School of Economics, Canterbury

This Version is available at:
https://hdl.handle.net/10419/105565

University of Kent

School of Economics Discussion Papers

Social Norms, Higher-Order Beliefs and the Emperor's New Clothes

Zaki Wahhaj

February 2012

KDPE 1210

SOCIAL NORMS, HIGHER-ORDER BELIEFS AND THE EMPEROR'S NEW CLOTHES

ZAKI WAHHAJ

Abstract. The use of social sanctions against behaviour which contradicts a set of informal rules is often an important element in the functioning of informal institutions in traditional societies. In the social sciences, sanctioning behaviour has often been explained in terms of the internalisation of norms that prescribe the sanctions (e.g. Parsons 1951) or the threat of new sanctions against those who do not follow sanctioning behaviour (e.g. Akerlof 1976). We propose an alternative mechanism for maintaining a credible threat of social sanctions, showing that even in a population where individuals have not internalised a set of social norms, do not believe that others have internalised them, do not believe that others believe that others have internalised these norms, etc., up to a finite nth order, collective participation in social sanctions against behaviour which contradict the norms is an equilibrium if such beliefs exist at higher orders. The equilibrium can persist even if beliefs change over time, as long as the norms are believed to have been internalised at some finite nth order. The framework shows how precisely beliefs must change for the equilibrium to unravel and social norms to evolve.

JEL Codes: D01, D02, D83, Z10

Keywords: social norms, higher-order belief, social sanctions, community enforcement, dynamics of norms, institutional change.

Corresponding Address: School of Economics, Keynes College, University of Kent, Canterbury CT2 7NP, United Kingdom

Date: February 2012.

1. Introduction

The use of social sanctions against behaviour which contradicts a set of informal rules is often an important element in the functioning of informal institutions. It appears, for example, in theoretical explanations of informal risk-sharing in village societies (Kimball 1988; Fafchamps 1992; Coate and Ravallion 1994), the effectiveness of joint liability credit contracts in eliciting high repayment rates (Besley and Coate 1994), the endurance of the caste system in India (Akerlof 1978), and contract enforcement in the context of medieval trade (Greif 1993).

Sanctioning behaviour may be costly for individuals who are required to impose the sanctions, for they need to break, at least temporarily, a profitable social connection; how sanctions may be sustained in spite of this cost have, broadly, two sorts of explanations in the literature, encapsulated in the terms *homo economicus* and *homo sociologicus* (Elster, 1989).

The homo economicus is a person who has no instrinsic views on the behaviour that is contrary to the social norm. He weighs the cost and benefit of participating in a social sanction; in particular, the direct cost to himself from ostracising a person in the community, and the communal punishment he may face himself if he refuses to participate in the sanctions. Social sanctions against certain types of behaviour may be sustained because each fears being subject to similar sanctions if he refuses to engage in the collective punishment of another.

For example, in Akerlof's explanation for the endurance of the Indian caste system, individuals adhere to caste rules, which includes sanctioning those who have not adhered to them, because they fear being subject to the same punishment otherwise. Thus, the caste system is sustained although individuals (within this theoretical framework) have no intrinsic views on the validity of the caste rules (Akerlof 1976).

The homo sociologicus, by contrast, is socially conditioned to disapprove of behaviour that contradicts the social norm; in other words, he has internalised these norms. His disapproval may be sustained partly by the fact that this response is supported by others, but there is no cost-benefit calculation behind the response. (Elster, 1989) Rather, he is driven by emotion and instinct, and he may go to some length to express his abhorrence of the behaviour that violates the social norm, even at a personal cost to himself. The

internalisation of norms play an important in, for example, Talcott Parsons' theory of socialisation (Parsons 1951).

An important characteristic of the economic approach to modelling social sanctions is that each person follows — or, at least, is expected to follow — the sanctioning rules because doing so is optimal given the strategies of the other agents. But a coordinated change in strategies, if this were feasible, could lead to a change in collective behaviour, and perhaps an improvement in welfare. In game-theoretic terms, the proposed equilibria are not necessarily *renegotiation-proof* (Farrell and Maskin, 1989).

By contrast, the sociological approach posits a close correspondence between the preferences of individuals and the equilibrium in which sanctioning behaviour occurs. But, if preferences are slow to evolve, then it offers limited scope for explaning why, as documented widely in the literature, social norms can remain stationary over long periods and unravel suddenly (Bicchieri, 2011).[1]

In this paper, we propose an alternative mechanism for maintaining a credible threat of social sanctions. We show that even in a population where individuals have not internalised a set of social norms, do not believe that others have internalised them, do not believe that others believe that others have internalised these norms, etc., up to a finite nth order, collective participation in social sanctions against behaviour which contradict the norms is an equilibrium if such beliefs exist at higher orders.

Given first-order and higher-order beliefs, the equilibrium is renegotiation-proof: in the subgame where sanctioning behaviour occurs, there is no alternative equilibrium path in which all individuals are better-off. The equilibrium can persist even if beliefs change over time, as long as the norms are believed to have been internalised at some finite nth order. The framework shows how precisely beliefs must change for the equilibrium to unravel.

The main technical result in this work is anticipated in Ariel Rubinstein's seminal paper on the 'Electronic Mail Game' (Rubinstein 1989). The important insight to emerge from the 'Electronic Mail Game' is that 'almost common knowledge', referring to a situation where players have very high-order knowledge about a particular event, will not necessarily lead to the same behaviour as common knowledge.

[1]In this context, two important examples are the abolition of footbinding in China during the 20th century and the shift in norms regarding female circumcision in Senegal at the start of the 21st, documented by Mackie (1996, 2000). Bicchieri (2011) provides further examples.

In the recent game-theoretic literature on higher-order beliefs, Weinstein and Yildiz (2007) have shown that there is a strong correspondence between beliefs (including higher-order beliefs) and the set of rationalisable outcomes in a normal-form game. In particular, given any rationalisable outcome of the game, players' beliefs may be perturbed in such a way that the outcome is uniquely rationalisable. Chen (2008) and Weinstein and Yildiz (2010) obtain similar results for dynamic games.

From the perspective of this literature, we propose a mechanism, for the functioning of social sanctions, for which the belief structure regarding the internalisation of a particular social norm determines whether contrary behaviour will be subject to social sanctions in equilibrium. Thus, it provides a link between the game-theoretic literature on the role of higher-order beliefs in equilibrium selection and the question of how social sanctions operate in traditional societies.

There are important parallels between Timur Kuran's concept of 'preference falsification' and the role of higher order beliefs in sustaining social taboos explored in this paper. Kuran (1995) considers a variety of social situations where individuals refrain from actions that express their true beliefs or preferences for fear of the repurcussions that such a revelation would bring. Within this framework, people may go along with a particular type of sanctioning behaviour not because they have internalised the social norms that prescribe the sanctions, but because they would rather not reveal to anyone that they have not internalised these norms. This may give rise to situations where nobody gives public expression to their true beliefs, people harbour false notions of each other's true beliefs, and a social taboo is maintained although everyone's true preferences are contrary to the social norm that prescribe the taboo.

Our results imply that 'preference falsification' (whereby individuals punish certain behaviour although they have not internalised the norms that forbid it) can provide a basis for maintaining social taboos even when individuals have accurate beliefs about each others' true beliefs up to any finite nth order.

The remainder of this paper is organised as follows. Section 2 revisits Han Christian Andersen's famous story of "The Emperor's New Clothes", which provides an elegant way to illustrate the mechanism by which social taboos are maintained in our theoretical framework. Section 3 presents the formal model, and the standard economic theory as to how a social taboo may be sustained within this model. An epistemic game based on this formal model is developed in Sections 3.1–3.3 to illustrate the role of higher-order

beliefs on the maintenance of social taboos. Section 3.4 discusses some properties of the equilibrium of interest while the dynamic implications of the model are discussed in Sections 4 and 5.

2. An Interpretation of Hans Christian Andersen's "The Emperor's New Clothes"

The fundamental insight that is being proposed in this paper may be illustrated through a particular interpretation of Hans Christian Andersen's story, "The Emperor's New Clothes"[2]. In the story, two swindlers appear before an emperor pretending to be tailors and propose to make him a costume from the finest possible cloth. They add that this cloth is 'invisible to those who are unpardonably stupid or unfit for their office'. Of course, no such cloth exist. But each person sent by the emperor to observe the swindlers at work pretends to see the cloth, and the emperor, in his turn, pretends to see it as well.

Everyone keeps up this pretence because they fear being called 'unpardonably stupid or unfit for their office' if they admit that they cannot actually see the cloth. We can argue, quite reasonably, that even if one of the emperor's ministers were quite sure that the cloth did not really exist, he would keep silent as long as he believed that others believed in its existence and the swindlers' declaration about it; since they would think him 'unpardonably stupid or unfit for his office' otherwise.

But if a person – let us call him B1 – who does not believe in the existence of the cloth, and merely believes that others do, has reason to keep silent, then so does a person, let us call him B2, who does not believe that the cloth exists or that others believe that it does, but does believe that everyone else is like B1. This point is critical, for it shows how beliefs can interact with each other to produce very strange situations. We can apply this reasoning iteratively to show that any higher order belief in the existence of the cloth may be sufficient to sustain an equilibrium where no one admits that they cannot see anything.

At the end of the story, during a regal procession in which the emperor marches adorned in his new 'garments', a little child points to the obvious – that the emperor is not wearing

[2]Jean Hersholt's *The Complete Andersen* (The Limited Editions Club, New York 1949), which includes an English translation of "The Emperor's New Clothes" may be accessed at this website: http://www.andersen.sdu.dk/vaerk/hersholt/index_e.html

anything. Immediately, everyone gives up the pretence. But if everyone already knew that the emperor was naked, should the child's declaration make any difference in people's behaviour? One possible explanation is that no one, not even a hypothetical person who only exists in someone's higher order beliefs can continue to believe that the cloth really exits after the child has made his declaration, because a child cannot be 'unpardonably stupid' or 'unfit for his office'. Thus, we see that a statement of the obvious by the person with the 'right' credentials can dramatically change social behaviour in certain contexts.

We discussed "The Emperor's New Clothes" here to illustrate that by focusing on beliefs of individuals, and particularly what they believe about what others believe, etc. can produce a rich theorecal framework for the analysis of social sanctions and social taboos. Much of this richness is lost within a framework where one holds beliefs only about how others are going to behave. The next section formalises the argument made here in the context of Hans Christian Andersen's story.

3. Formal Model

Imagine a population of individuals indexed $i = 1, 2, .., n$. We denote by $\mathcal{I} = \{1, 2, .., n\}$ the set of individuals. We define a stage game $\mathcal{G}$ in which two types of random events may occur:

(i) Let e_o^{ij} be the event that person i is in a position to 'engage in social ostracism against' person j. If event e_o^{ij} occurs, then person i has a choice of action α_o^{ij} which can take a value of 0 or 1, where $\alpha_o^{ij} = 1$ represents the action that person i 'opts to ostracize j', and $\alpha_o^{ij} = 0$ represents the action that he does not.

(ii) Let e_w^i be the event that person is in a position to 'engage a certain public act with welfare implications for the entire community'. If event e_w^i occurs, then person i has a choice of action α_w^i which can take a value of 0 or 1, where $\alpha_w^i = 1$ represents the action that 'person i engages in the public act in question', and $\alpha_w^i = 0$ represents the action that he 'desists from it.'

We assume that $\Pr(e_w^i) = \delta_w$ for each $i \in \mathcal{I}$ and $\Pr(e_o^{ij}) = \delta_o$ for $i, j \in \mathcal{I}$. Furthermore, we assume that these events are mutually exclusive. Therefore, we require $n\delta_w + n(n-1)\delta_o \leq 1$.

We introduce to this environment the notion of a personal characteristic called 'being immoral'. A community member will receive some psychological reward from ostracising a person who is 'immoral', and, therefore, would willingly engage in such an act of ostracism in the absence of any other incentives or disincentives.

What 'being immoral' may actually mean is unimportant for our purpose. Its significance lies in the notion that it is a characteristic that is generally found to be abhorrent, such that people would not wish to associate with those who are believed to possess this quality. There may be no scientific method of detecting, or even defining, what it means to 'be immoral'. Nevertheless, as we shall see, the notion will play a critical role in sustaining a social taboo, and a credible threat of social ostracism in the mechanism proposed in this paper.

To each person i, we assign a variable M_i describes his or her 'moral character': $\mathrm{M}_i = 1$ if person i is 'immoral' and $\mathrm{M}_i = 0$ otherwise. We assume that M_i is unobservable to any community member, including person i; and prior beliefs are given by $\Pr(\mathrm{M}_i = 1) = \varepsilon$ where ε is positive but negligibly small. The payoffs in the stage-game are given by

$$u^i = -\sum_{j \neq i} \left[\mathbf{I}\left(e_o^{ji}\right) \alpha_o^{ji} P + \mathbf{I}\left(e_o^{ij}\right) \alpha_o^{ij} \left\{ C - \mathrm{M}_j R \right\} \right] + \sum_{j \in \mathcal{I}} \mathbf{I}\left(e_w^j\right) \alpha_w^j W \tag{1}$$

where $\mathbf{I}(e)$ is an indicator function which takes a value of 0 or 1 depending on whether or not event e has occurred; C represents the cost of engaging in an act of social ostracism, and R is the psychological reward from ostracizing an 'immoral' person; P is the disutility that such an action would inflict on the person being ostracized; W represents the payoff to each community member from any one person engaging in the public act in question. We allow for the possibility that this act may be either a public good or a public bad; i.e. $W \lesseqgtr 0$. On the other hand, since the negative of P and C represent costs and R is a reward, we have $P, C, R > 0$.

We analyse the game $\mathcal{G}(\infty)$ in which the stage game $\mathcal{G}$ is repeated infinitely many times and future payoffs are discounted at a constant rate $\beta \in (0, 1)$ per period. The infinite repetition ensures that there is, in particular, always a future period in which one may be subject to social ostracism by others. Suppose, first, that past behaviour regarding the public act do not affect players' beliefs regarding the variables M_i, $i \in \mathcal{I}$. This can be interpreted as meaning that they do not have any intrinsic views about the 'morality' of the public act. Nevertheless, the variety of norms regarding the public act can be sustained in a (subgame-perfect) equilibrium. Below we illustrate two possibilities.

To describe the first of these equilibria, we shall make use of the following definition.

Definition 3.1. $(\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2..)$ *is a sequence of subsets of* $\mathcal{I}$ *defined as follows:*

$\mathcal{B}_0 = \emptyset$

For $t = 1, 2, ..,$

$$\mathcal{B}_t = \left\{ i \in \mathcal{I} : \begin{array}{c} i \in \mathcal{B}_{t-1} \text{ or } \alpha^i_{w,t} = 1 \\ \text{or } (\alpha^{ij}_{o,t} = 0 \text{ and } j \in \mathcal{B}_{t-1} \text{ for some } j \in \mathcal{I}) \end{array} \right\}$$

The set $\mathcal{B}_t$ is a time-specific 'blacklist' which includes all individuals who have previously engaged in the public act, or has failed to ostracise someone on the 'blacklist'. Consider the following strategy of the stage game $\mathcal{G}$ which makes use of this 'blacklist':

$\bar{s}^i_1$: If $e^i_{w,t} = 1$, choose $\alpha^i_{w,t} = 0$; if $e^{ij}_{o,t} = 1$ and $j \in \mathcal{B}_{t-1}$, choose $\alpha^{ij}_{o,t} = 1$; if $e^{ij}_{o,t} = 1$ and $j \notin \mathcal{B}_{t-1}$, choose $\alpha^{ij}_{o,t} = 0$.

Consider also an alternative stage-game strategy defined as follows:

$\bar{s}^i_2$: If $e^i_{w,t} = 1$, choose $\alpha^i_{w,t} = \arg\max_{\alpha \in \{0,1\}} \alpha W$; if $e^{ij}_{o,t} = 1$, choose $\alpha^{ij}_{o,t} = 0$.

The strategy $\bar{s}^i_1$ says that one should not engage in the public act and ostracise only those who are on the blacklist. The strategy $\bar{s}^i_2$ simply instructs the player to take the lowest cost action in the stage game. Suppose that, in each period t, each person $i \in \mathcal{I}$ adopts the stage-game strategy $\bar{s}^i_1$ while $i \notin \mathcal{B}_t$ and the alternate strategy $\bar{s}^i_2$ if $i \in \mathcal{B}_t$. This constitutes a subgame perfect equilibrium of the repeated game $\mathcal{G}(\infty)$ if

$$W < \frac{\beta(n-1)\delta_o}{1-\beta} P \tag{2}$$

$$C - \varepsilon R < \frac{\beta(n-1)\delta_o}{1-\beta} P \tag{3}$$

The first condition (2) ensures that it never pays to engage in the public act when doing so would cause one to be 'blacklisted' and lead to perpetual ostracism within the community. The second condition (3) ensures that one is better-off following the rules of ostracism rather than ignoring them.

Thus, we have an equilibrium in which no one engages in the public act for fear of being ostracised. This is regardless of whether commiting this act is a public bad – such as damaging a public property – or a public good, such as accomplishing a task which is beneficial to the entire community.

Our second example of an equilibrium will be exactly the inverse of the first and is just as simple to construct. First we define an alternative 'blacklist' as follows:

Definition 3.2. $\left(\tilde{\mathcal{B}}_0, \tilde{\mathcal{B}}_1, \tilde{\mathcal{B}}_2..\right)$ *is a sequence of subsets of* $\mathcal{I}$ *defined as follows:*

$\tilde{\mathcal{B}}_0 = \emptyset$

For $t = 1, 2, ..,$

$$\tilde{\mathcal{B}}_t = \left\{ i \in \mathcal{I} : \begin{array}{c} i \in \tilde{\mathcal{B}}_{t-1} \text{ or } \alpha^j_{w,t} = 0 \\ \text{or } (\alpha^{ji}_{o,t} = 0 \text{ and } j \in \mathcal{B}_{t-1} \text{ for some } j \in \mathcal{I}) \end{array} \right\}$$

The 'blacklist' $\tilde{\mathcal{B}}_t$ is the opposite of $\mathcal{B}_t$. One finds oneself on the blacklist by *failing* to engage in the public act in question when one has the opportunity to do, or failing to ostracise a blacklisted person.

As before, we define a stage-game strategy which is based on this blacklist:

$\bar{s}^i_3$: If $e^i_{w,t} = 1$, choose $\alpha^i_{w,t} = 1$; if $e^{ij}_{o,t} = 1$ and $j \in \mathcal{B}_{t-1}$, choose $\alpha^{ij}_{o,t} = 1$; if $e^{ij}_{o,t} = 1$ and $j \notin \mathcal{B}_{t-1}$, choose $\alpha^{ij}_{o,t} = 0$.

The stage game strategy $\bar{s}^i_3$ says that one should engage in the public act and ostracise those who are on the blacklist. If, in each period t, each person $i \in \mathcal{I}$ adopts the stage-game strategy $\bar{s}^i_3$ while $i \notin \mathcal{B}_t$ and the strategy $\bar{s}^i_2$ if $i \in \mathcal{B}_t$, then this also constitutes a subgame perfect equilibrium of the repeated game if

$$-W < \frac{\beta (n-1) \delta_o}{1-\beta} P \tag{4}$$

and the condition in (3) holds. Thus, we have an equilibrium in which everyone engages in the public act in question for fear of being ostracised.

The theory developed thus far offers a mechanism whereby social taboos may be sustained, and provides conditions under which a particular taboo can be sustained. But it is unsatisfactory in a number of respects. It does not explain why a taboo exists with

respect to one type of behaviour and not another: we see above that is only slightly more difficult to maintain a taboo against a behaviour which is a public good as against a public bad. And it does not explain when a social taboo may emerge or how it may unravel. And perhaps most unsatisfactorily, a social taboo, if it exists, need bear no relationship with any sort of moral beliefs shared by the community: the problem of maintaining or breaking a taboo is merely a problem of social organisation.

In the following section, we develop an alternative theory of social taboos which addresses some of the concerns raised here.

3.1. **A Syntactic Language to Model Beliefs.** To formally introduce beliefs into this framework, we shall make use of a syntactic language based on Aumann (1999) to describe the structure of knowledge and beliefs at each stage of the game. The building blocks of the language consists of the letters of an 'alphabet' $\mathrm{X} = \{x, y, z, ...\}$ and the symbols $\neg, \vee, (,)$, and k_i for each $i \in \mathcal{I}$.

Aumann defines a *formula* as a finite string of symbols constructed according to the following three rules:

(i) Every letter in the alphabet is a formula.
(ii) If f and g are formulae, so is $(f) \vee (g)$.
(iii) If f is a formula, so are $\neg (f)$ and $k_i (f)$ for each i.

Parantheses may be omitted if doing so does not result in any ambiguity. The symbol $\Longrightarrow$, used as in $f \Longrightarrow g$ is used as an abbreviation of the formula $\neg (f) \vee (g)$. The symbol $\wedge$, used as in $(f) \wedge (g)$ is used as an abbreviation of the formula $\neg ((\neg f) \vee (\neg g))$.

A *list* is defined as a set of formulae. A list $\mathcal{L}$ is called *logically closed* if

$$f \in \mathcal{L} \text{ and } (f \Longrightarrow g) \in \mathcal{L} \text{ implies } g \in \mathcal{L}$$

A list $\mathcal{L}$ is called *epistemically closed* if

$$f \in \mathcal{L} \text{ implies } k_i f \in \mathcal{L} \text{ for each } i \in \mathcal{I}$$

A list $\mathcal{L}$ is called *strongly closed* if it is both logically closed and epistemically closed.

The *strong closure* of a list $\mathcal{L}$ is the smallest strongly closed list that includes $\mathcal{L}$ (i.e. the intersection of all strongly closed lists including $\mathcal{L}$). Aumann then defines a *tautology* as a formula in the strong closure of the list of all formulae having one of the following seven forms (for some formulae f, g, h, and person i).

(a) $(f \vee f) \Longrightarrow f$
(b) $f \Longrightarrow (f \vee g)$
(c) $(f \vee g) \Longrightarrow (g \vee f)$
(d) $(f \Longrightarrow g) \Longrightarrow ((h \vee f) \Longrightarrow (h \vee g))$
(e) $k_i f \Longrightarrow f$
(f) $k_i\,(f \Longrightarrow g) \Longrightarrow ((k_i f) \Longrightarrow (k_i g))$
(g) $\neg k_i f \Longrightarrow k_i \neg k_i f$

Aumann calls the set of all formulae specific to a population $\mathcal{I}$ a *syntax*, and provides the following interpretation of the syntactic formalism. The letters of the alphabet are 'natural occurrences'; i.e. happenings in the physical world, that exclude logical statements, and statements involving knowledge. The symbol k_i means that 'person i knows that ...'. The symbol $\neg$ stands for 'not', and $\vee$ stands for 'or'. From the definition of $\Longrightarrow$, one can verify that the symbol retains its standard meaning in mathematics, and stands for 'implies that...'; while $\vee$ stands for 'and'.

A *tautology* has been defined to capture its standard meaning in the English language: a statement whose truth is inherent in the meanings of the terms involved. Furthermore, Aumann's requirement that the set of tautologies be epistemically closed means that tautologies are common knowledge.

We shall extend the language developed by Aumann by introducing the operator b_i which will mean that 'person i believes that...'. If f is a formula, then $b_i f$ is also a formula.

We also add five new forms to the seven forms of formulae which Aumann uses to construct his list of tautologies:

(h) $b_i\,(f \Longrightarrow g) \Longrightarrow ((b_i f) \Longrightarrow (b_i g))$
(i) $\neg b_i f \Longrightarrow b_i \neg b_i f$
(j) $b_i f \Longrightarrow k_i b_i f$
(k) $k_i f \Longrightarrow b_i f$
(l) $b_i f \Longrightarrow \neg b_i \neg f$

The forms (h) and (i) mirror the forms (f) and (g) relating to knowledge. Thus, the belief function is like the knowledge function, with one important exception. While $k_i f \implies f$ is a tautology, we have not specified that $b_i f \implies f$ is: one may 'believe' in something without it being necessarily true. Form (j) says that if one believes in something, then one has knowledge of that belief. This knowledge may or may not be shared with others. Therefore, it may or may not constitute private information. Form (k) says that knowledge implies belief; but the contrary is not true. Finally, form (l) says that if one blieves in something, he does not also believe in its opposite.

We define the notion of 'common belief' akin to the notion of 'common knowledge'. Within the syntatic language, a formula f is common knowledge if all formulae of the form $k_i k_j..k_l f$ hold for each $i, j, m \in \mathcal{I}$ (including $k_i f$ and $k_i k_j f$). Likewise, we say that a formula f is 'common belief' if all formulae of the form $b_i b_j..b_l f$ hold for each $i, j, l \in \mathcal{I}$.

The belief operator, as defined here, is akin to the formulation of beliefs by Battigali and Bonanno (1999). Specifically, the tautological forms (i)-(l) ensure that the belief operator satisfies *seriality*, *transitivity* and *euclideanness* which, as Battigali and Bonanno note, can be regarded as an expression of 'rational' belief.

Following Aumann (1999), we define a *state of the world* ω as a *logically closed, coherent,* and *complete* list of formulae that contains all tautologies. In this context, a list $\mathcal{L}$ is said to be *coherent* if

$$\neg f \in \mathcal{L} \text{ implies } f \notin \mathcal{L}$$

and *complete* if

$$f \notin \mathcal{L} \text{ implies } \neg f \in \mathcal{L}$$

Finally, the set of all states will be denoted by Ω.

Aumann (1999) shows that there is a direct correspondence between the syntactic language developed here and the better-known semantic language of knowledge, in which knowledge is represented using partitions of Ω. The link between the two systems is provided by the assumption that if two states of the world are indistinguishable from each other on the basis of what some person i 'knows' (i.e. on the basis of formulae which begin with k_i) then they must belong to the same *information set* for person i. It is reasonable to impose the restriction that two states are indistinguishable in terms of person i's knowledge are also indistinguishable in terms of his beliefs. Otherwise, he would be able to tell these

states apart on the basis of what he believes in each state. This assumption is formalised below.

Let $K_i(\omega)$ be the set of formulae in ω which begin with k_i. Let B_i be the set of formulae in ω which begin with b_i.

Assumption 1. *If* $K_i(\omega) = K_i(\omega')$, *then* $B_i(\omega) = B_i(\omega')$.

Note that Assumption 1 does not preclude the possibility that what person i believes in state ω and ω' is, in fact, false.

3.2. **Applying the Syntactic Language to the Model.** Using the epistemic language developed in the preceding section, we shall now describe an alternative equilibrium of the model introduced at the beginning of section 3, in which the players' beliefs come into play.

The following will constitute the 'alphabet' of the epistemic game. Let $c_{i,\tau}$ be the occurrence that 'person i committed the public act in question in period τ'. Let $o_{ij,\tau}$ be the occurrence that 'person i ostracised person j in period τ'. Let m_i be the occurrence that person i 'is immoral' (i.e. $\mathrm{M}_i = i$). Thus, the 'alphabet' of the epistemic game is given by $\mathrm{X} = \left\{ \{o_{ij,\tau}\}_{j \neq i, \tau \in \mathbb{N}}, \{c_{i,t}\}_{\tau \in \mathbb{N}}, m_i \right\}_{i \in \mathcal{I}}$.[3]

We use $\omega_t \in \Omega_t$ to denote the state of the world at the beginning of period t following some history. Therefore, if person i has committed the public act in some period τ prior to period t, then $c_{i,\tau}$ 'holds true' at ω_t or, in mathematical terms, $c_{i,\tau} \in \omega_t$. For ease of notation, we may drop the time subscript when using the alphabet of the epistemic game when the exact time period of the occurrence is not relevant. Thus, c_i will stand for 'person i has committed the public act in the past' and, similarly, o_{ij} will stand for 'person i has previously ostracised person j.' We let h_t be the list of formulae in ω_t taking any of the following forms: $o_{ij,\tau}$, $\neg o_{ij,\tau}$, $c_{i,\tau}$ and $\neg c_{i,\tau}$. Thus, h_t represents the history of past actions at the beginning of period t, as summarised in ω_t.

We impose the condition that when a member commits the public act, or ostracises another individual, this action becomes common knowledge within the community. We also assume that no individual has any knowledge about whether anyone, including oneself, is

[3] $\mathbb{N} = \{1, 2, 3, ...\}$ stands for the set of positive integers.

moral although they may all have *beliefs* regarding their own moral integrity and that of others. Formally, these assumptions can be stated as follows:

Assumption 2. *If $c_{i,\tau} \in \omega_t$, then $k_i k_j..k_l c_{i,\tau} \in \omega_t$, for each $i, j, l \in \mathcal{I}$. If $o_{ij,\tau} \in \omega_t$, then $k_i k_j..k_l o_{ij,\tau} \in \omega_t$, for each $i, j, l \in \mathcal{I}$.*

Assumption 3. *For each $\omega_t \in \Omega_t$, $\neg k_i m_j \in \omega_t$ and $\neg k_i \neg m_j \in \omega_t$ for $i, j \in \mathcal{I}$.*

In the next step, we will specify the players' *beliefs* about whether someone else is 'immoral' or not as a function of past behaviour within the game. These beliefs, as will be seen, can be construed as a particular 'moral code' which can potentially restrict the number of possible equilibria within the original game. To construct these beliefs, we define a series of assertions, or formulae, using the letters of the alphabet o_{ij}, c_i and m_i:

(i) $T_0 = (c_i \implies m_i \text{ for each } i \in \mathcal{I})$
(ii) $T_n = (\neg b_i T_{n-1} \implies m_i \text{ for each } i \in \mathcal{I})$ for $n = 1, 2, ...$
(iii) $T = T_0 \wedge T_1 \wedge T_2...$

In words, the formula T_0 says that 'anyone who commits the public act is immoral'; T_n says that 'anyone who does not believe in T_{n-1} is immoral' where n is a positive integer. Finally, T can be interpreted as saying that 'anyone who commits the public act is immoral, and anyone who contradicts this proposition in any way is also immoral.' These statements do not have any significance as of yet since we have not specified any consequences of 'being immoral' or 'being perceived as being immoral.' We do this next.

Private Information: Following Harsanyi (1967, 1968), we use the notion of a person's 'type' to represent private information in the game. Which information is private? Under Assumption 2, all past actions in the repeated game are public information. Under Assumption 3, everyone is equally ignorant about who actually is or isn't 'immoral'. Therefore, the only information in the game that may be private relate to one's own beliefs, including higher order beliefs, about $T_0, T_1, T_2..$ and T. We allow for the following possible types:

Definition 3.3. *For each person $i \in \mathcal{I}$, the person is of 'type 0' if he/she believes T and believes that T is common belief; a person is of 'type θ' ($\theta = 1, 2, ..$) if he/she does not believe T and believes that all others are of type $\theta - 1$ or less.*

Note that the range of different types in Definition 3.3 imposes a particular structure on the states of the world. If each person in the community must belong to one of the types described in 3.3, this restricts the combination of belief-related formulae that must hold in some state ω_t. For example, if all players are of type 0, then T is common belief in each state of the world. Therefore, all statements of the form $b_i T$, $b_i b_j T$, $b_i b_j b_m T$, etc. must hold at each $\omega_t \in \Omega_t$. If all players are of type 1, we must have $b_i \neg T$, as well as $b_i b_j T$, $b_i b_j b_m T$, etc. at each $\omega_t \in \Omega_t$, and so on. A player of any one of the types defined in Definition 3.3 retains the same beliefs for each possible history and, therefore, Assumption 1 is satisfied.

Note that, unlike Harsanyi (1968), we do not define common priors on the probabilities of each type; moreover, prior beliefs are not used to evaluate each type's beliefs about the types of the other players. Instead, these beliefs are fully described in the definition of the types.

It remains for us to define what is a strategy and what is an equilibrium in the epistemic game:

Definition 3.4. *A strategy $s_i(\theta_i, t, h_t)$ for player i prescribes actions $\alpha_{o,t}^{ij}$ and $\alpha_{w,t}^{i}$ in each period t, (to be taken in the event that $e_o^{ij} = 1$ and $e_w^i = 1$ respectively) as a function of person i's type θ_i and the history of past actions h_t.*

Denote by $\mathcal{S}^i$ the set of possible strategies for player i.

Definition 3.5. *A strategy profile $\{s_i(.)\}_{i \in \mathcal{I}}$ constitutes an equilibrium iff*

$$s_i(\theta_i, t, h_t) \in \arg\max_{s \in \mathcal{S}^i} E_{\theta_i} \sum_{t=1}^{\infty} \beta^{t-1} u_i(s, s_{-i}(.); h_t) \text{ for each } i \in \mathcal{I} \tag{5}$$

In (5), the term $u_i(s, s_{-i}(.); h_t)$ should be interpreted as follows. It is the level of utility to person i, as defined in (1), when actions $\alpha_{o,t}^{ij}$ and $\alpha_{w,t}^{i}$ follow the prescriptions of s, the actions of all other players in the population follow the prescriptions of $\{s_j(.)\}_{j \in \mathcal{I}, j \neq i}$ and the prescriptions of s and $s_{-i}(.)$ may be contingent upon the history of past actions denoted by h_t. The operator E_{θ_i} evaluates utility on the basis of person i's beliefs, as represented by her type θ_i.

3.3. **Characterisation of Equilibria of the Epistemic Game.** In this section, we provide a characterisation of equilibria of the epistemic game. We begin by considering the

possible strategies for a type-0 individual. Such an individual, by definition, believes that a person who has committed the public act is immoral. Therefore, if the disutility from associating with an immoral person (represented by the variable R) is sufficiently high, a type-0 individual would ostracize one who has committed the public act, regardless of the strategies pursued by others.[4] Moreover, from the definition of statement T, it follows that the type-0 person believes that a person, say i, who has failed to ostracize someone who has committed the public act is immoral, someone who has failed to ostracize person i is immoral, etc. and should therefore also opt to ostracize all these individuals.

Furthermore, since a type-0 individual believes that everyone else in the community is of type-0, who, by definition, believe in the assertion T, she would expect to be ostracized by everyone were she to engage in the public act. Therefore, she would refrain from doing so. There remains only the question of whether a type-0 individual would ostracize someone whom she *does not* believe to be immoral. It is possible to construct equilibria where she does so out of fear of ostracism by others, but these equilibria are clearly inefficient and will involve a very complex set of rules. We shall discuss the possibility of such equilibria further in Appendix A. For the present analysis, we propose the simplest and most efficient choice: that a type-0 individual does not ostracize an individual whom she does not believe to be immoral. In summary, we propose the following strategy for the type-0 individual:

s_0^i : If $e_{w,t}^i = 1$, then choose $\alpha_{w,t}^i = 0$. If $e_{o,t}^{ij} = 1$ then, if $b^i m_j$, choose $\alpha_{o,t}^{ij} = 1$; otherwise, choose $\alpha_{o,t}^{ij} = 0$.

In words, strategy s_0^i says the following: 'Do not engage in the public act. Ostracize j if you have any reason to believe that he is immoral but not otherwise.' We shall see that if each type-0 individual adopts strategy s_0^i, then there is a unique optimal strategy for all higher types. The reasoning is straightforward and proceeds as follows.

[4]To make this argument more precisely, the largest punishment that a community can conceivably inflict on any one of its members is to subject him to perpetual ostracism and to engage in the public act, assuming it is a public bad (or desist from it if it is a public good) to punish the person in question even more. The expected disutility from such a collective punishment would equal $\frac{\beta(n-1)}{1-\beta}(\delta_o P + \delta_w \|W\|)$. Therefore, if

$$R - C > \frac{\beta(n-1)}{1-\beta}(\delta_o P + \delta_w \|W\|) \tag{6}$$

a type-0 individual should ostracise someone who has committed the public act regardless of the repercussions.

Recall that the type-1 individuals believe that all other community members are of type-0. Therefore, in an equilibrium where a type-0 individual is playing strategy s_0^i, a type-1 individual must reason that everyone else is playing strategy s_0^i. Therefore, she expects to be ostracized by everyone else if she engages in the public act. Therefore, she would not do so if (2) holds. She also reasons that if she, through her actions, contradicts the logic of the assertion T, then a type-0 individual would conclude that she is immoral, and ostracize her thereafter. Therefore, it is optimal for her to ostracize anyone who has engaged in the public act, ostracize anyone who has failed to do the same, and so on if (3) holds. Moreover, she has no reason to ostracize anyone who has not engaged in the public act or who is not believed to be immoral by a type-0 individual. Therefore, if the type-0 individuals are playing strategy s_0^i, then under conditions (2) and (3), the following is the unique optimal strategy for a type-1 individual.

s_1^i : If $e_{w,t}^i = 1$, then choose $\alpha_{w,t}^i = 0$. If $e_{o,t}^{ij} = 1$ then, if $(h_t \wedge T) \implies m_j$, choose $\alpha_{o,t}^{ij} = 1$; otherwise, choose $\alpha_{o,t}^{ij} = 0$.

In words, strategy s_1^i prescribes the following: 'Do not engage in the public act. Ostracize person j if and only if belief in assertion T and the history of past events would lead one to conclude that this person is immoral.'

A type-2 individual believes that everyone else is either of type-0 or of type-1. According to the strategies s_0^i and s_1^i, both a type-0 and a type-1 individual would ostracize anyone who has engaged in the public act. Therefore, if the inequality (2) holds, then a type-2 individual would not engage in the public act. Recall also that both a type-0 and a type-1 individual would also ostracize another person when failing to do so contradicts the logic of assertion T and the inequality in (3) holds. Therefore, a type-2 individual will ostracize someone who has engaged in the public act, ostracize a person who has failed to ostracize someone previously engaged in the public act, etc. Lastly, a type-2 individual would not ostracize someone when none of the above conditions are fulfilled, because an act of ostracism is costly, and gains her nothing under these circumstances. Thus, we have established that if type-0 and type-1 individuals play strategies s_0^i and s_1^i respectively, then under conditions (2) and (3), the unique optimal strategy for a type-2 individual is the following:

s_2^i : If $e_{w,t}^i = 1$, then choose $\alpha_{w,t}^i = 0$. If $e_{o,t}^{ij} = 1$ then, if $(h_t \wedge T) \implies m_j$, choose $\alpha_{o,t}^{ij} = 1$; otherwise, choose $\alpha_{o,t}^{ij} = 0$.

Note that s_2^i is the same as s_1^i. By reasoning iteratively, we can show that s_1^i or s_2^i is also optimal for all higher types.

Finally, we show that if each individual i of type-0 is playing strategy s_0^i, then none of them have any incentive to deviate. We have already established that, for R sufficiently high, a type-0 individual would (i) ostracize another who has committed the public act, or has failed to ostracize someone who has committed the public act, etc; (ii) refrain from committing the public act herself. If all individuals $j \neq i$ play strategy s_0^j, then player i would receive no benefit for ostracizing a person who she does not believe to be immoral. Moreover, this is a costly action. Therefore, she would not ostracize such a person. Therefore, it is optimal for her to pursue strategy s_0^i herself. We have now established the following.

Proposition 3.1. *If the conditions in (2), (3) and (6) hold, and type-*0 *individuals play strategy* s_0^i*, the best response of all higher types is to play strategy* s_1^i*. Furthermore, this strategy profile constitutes an equilibrium.*

The reasoning behind Proposition 3.1 is, in many respects, similar to the main argument in Ariel Rubinstein's paper on 'The Electronic Mail Game' (Rubinstein, 1989). In Rubinstein's game, two players play a coordination game where payoffs depend on the true state of the world. Messages about the true state are communicated by an 'electronic mail' system which is such that the state may be known to both players but it is never common knowledge. If a player had no knowledge of the true state, he would prefer the action that involves 'less risk' (in the sense that, if he has chosen this action and they fail to coordinate, then he will not be penalised). Rubinstein shows, through iterative reasoning, that given the optimal choice for a player who has no knowledge about the state of the world, and the information structure implied by the electronic mail system, players with any finite level of higher-order knowledge about the true state would also opt for the less risky action.

3.4. **Characteristics of the Equilibrium in which the Social Taboo is sustained.** In this section, we discuss some important qualities of the equilibrium described in Proposition 3.1. The simplest type of equilibrium obtains if every member of the community is of type-0. Then they all believe in the association between the public act and the notion of 'immorality' embodied in the assertion T and behave accordingly. Thus we obtain a community of *homo sociologicus* who avoid the forbidden act, and spurn those who have

committed it, because they have internalised the social norm and are aware that those around them have internalised it too.

Preference Falsification under Increasingly Accurate Beliefs: In a community consisting entirely of type-1 individuals, we obtain the simplest possible example of a social taboo sustained by 'preference falsification', as defined by Kuran (1995): nobody believes in the association between the public act and the notion of 'immorality' but they all believe that everyone else does. They follow the behaviour implicitly prescribed by assertion T to hide their true beliefs, because they fear being accused of immorality otherwise.

In a community consisting entirely of type-2 individuals, everyone believes, accurately, that their neighbours may not believe in assertion T. This can be seen from the fact that if individuals i and j are of type-2, then we have, by construction, $b^i\left((b^j\neg T)\vee b^jT\right)$ (since i believes j to be either of type-0 or type-1; a type-0 individual believes in T but a type-1 individual does not) and $b^j\neg T$ (since a type-2 individual does not believe in T). However, they have inaccurate beliefs about what their neighbours believe about whether others believe in assertion T (since, by construction, $b^ib^jb^iT$ but $b^jb^i\neg T$). In other words, the second-order beliefs are inaccurate. And this causes everyone to follow the behaviour implied by assertion T to hide their true beliefs, because they fear being accused of immorality otherwise.

In a community consisting entirely of type-n individuals, for any positive integer n, everyone has accurate beliefs up to the n^{th} order. And *still* they hide their true beliefs, and behave in accordance with the social taboo, because they fear being accused of immorality otherwise.

Necessity of Common Knowledge of the 'Immorality' Norm: An important element of the equilibrium described in Proposition 3.1 is the psychological reward R that one obtains from ostracising an 'immoral' person. Without this reward, there is no reason why belief in the assertion T should affect a person's behaviour. Also, unless the reward R is common knowledge, the reasoning used in Proposition 3.1 would break down for some higher-order belief. In this sense, the social taboo requires that the community members have internalised *some* norms (e.g. one should ostracise an 'immoral' person, whatever 'immoral' may mean) and that this internalisation is common knowledge. The role of higher order beliefs regarding the psychological reward R here is akin to that in an elegant example by Gintis, called 'The Tactful Ladies' (Gintis 2009, page 153-156). In the example by Gintis, higher-order knowledge about certain social norms enable the ladies

in question to infer the state of their own appearance from very little information and the emotional response of others.

'*Renegotiation-Proofness' of the Social Taboo Equilibrium:* It is straightforward to show that the equilibrium in Proposition 3.1 satisfies the Farrell-Maskin criterion of 'renegotiation-proofness' (Farrell and Maskin, 1989). The criterion requires that the continuation payoffs following any history in the game cannot be Pareto dominated by the continuation payoffs following some other history (a formal and concise definition can be found in Fudenberg and Tirole, 1991, page 179). In other words, it cannot be that the community members follow a mode of behaviour following a particular history of events which makes them worse off, in the Pareto sense, than another mode of behaviour which they are supposed to practise following some other history. The idea behind such a restriction is that if the criterion were not satisfied, the players would have an interest to 'renegotiate' to the better equilibrium following the occurrence of the history of events referred to in the definition.

In the equilibrium described in Section 3.3, continuation strategies are contingent on the history of events only to the extent that beliefs about types depend on histories. Given beliefs about types following any history, a type-0 player would do worse in any other equilibrium, as we argued previously. It follows that the equilibrium is renegotiation-proof, as defined by Farrell and Maskin (1989).

The fact that the equilibrium is 'renegotiation-proof' has a significant meaning. It means that the person who has violated the social taboo cannot be 'forgiven'. Members of the community cannot 'let bygones be bygones': given existing beliefs, there is no other possible equilibrium where everyone is at least as well-off.

4. The Dynamics of Social Taboos

In sections 3.1-3.3, we developed a particular theory of how social taboos may be sustained in a community. An important characteristic of the proposed mechanism is that they depend on interactive beliefs of community members, and not on a particular coordination of actions or strategies. As shown by Proposition 3.1, the equilibrium outcome, and the existence of the social taboo, is almost fully determined by the assumed structure of interactive beliefs. By considering how such a belief structure may arise or unravel, we can obtain some insights about the dynamics of social taboos.

Consider, informally, what would happen if, in the repeated game described in section 3, a person of 'unquestionnable morality' engaged in the public act. If all community members believed that his morality is unquestionable, and this belief *supersedes* belief in the assertion T, then a type-0 individual would need to *abandon* his belief in the assertion T as soon as this person committed the public act. Moreover, as any engagement in the public act becomes common knowledge, a type-1 individual would know that a type-0 individual has abandoned his belief, and must therefore *abandon* his own beliefs about the existence of type-0 individuals in the community. Similarly, a type-2 individual would know that a type-1 individual has abandoned his own beliefs, and therefore would abandon her beliefs regarding the existence of type-1 individuals, and so on. If there is no belief in the assertion T in the community, and no higher order beliefs regarding T either, then the social taboo against the public act cannot be sustained.

The argument made in the previous paragraph is informal because we have not formally defined what it means for one belief to *supercede* another or for a person to *abandon* his beliefs. Indeed, to formalise this argument, we require a specific theory about the evolution of interactive beliefs. A theory for the evolution of beliefs, which would enable us to formalise the argument made above, is proposed in Section 5.

Consider, now, a very different case of a change in beliefs. If higher-order beliefs change over time because of reasons exogenous to the model, this would not affect the social taboo against the public act as long as each community member remained a θ-type. This follows from Proposition 3.1, where the equilibrium described requires only that all community members are θ-types, and not on any specific values of θ. This means that even if people grow more sophisticated in their beliefs over time – and realise, for instance, that the other community members do not actually believe in the association between the public act and immorality but are only pretending that they do – this is not sufficient to remove the social taboo. Thus, we have an instance where '... private variables ... undergo major changes without triggering changes in public opinion... they make it possible for profound transformations to occur, and much tension to build up, in a society that appears asleep.' (Kuran 1995; page 21) And the result suggests an explanation as to why cultural factors such as values, beliefs and social norms appear as 'slow-moving' institutions (Roland 2004).

Finally, the theoretical result offers some insights about how social taboos can emerge. In particular, we can ask ourselves what would we require for a community to be populated

with θ-type individuals? If a person with great moral authority makes a public statement equivalent to the assertion T, which becomes common knowledge, and he is believed by everyone, then we would have a situation where each community member is of type-0. It is obvious that no one would engage in the public act thereafter. Proposition 3.1 tells us that it is not, in fact, necessary for everyone to believe the public statement for the public act to become taboo. It would suffice that people believe that it is believed by others, or that they believe that others believe that it is believed by others, and so on. Thus, doubt becomes a potent tool for the enactment of self-enforcing norms.

5. Hierarchy of Beliefs

In this section, we propose a particular theory for the evolution of beliefs over time. This will allow us to formally describe the process whereby an equilibrium where a taboo is being practised may unravel, as in the first example discussed in Section 4.

Let $B^i(\omega)$ be the set of all formulae in ω which begin with b_i. We use the binary relation $\succ_{i,\omega}$to represent an ordering of elements of $B^i(\omega)$ satisfying the properties of

(1) *Completeness:* i.e. if $b_iA, b_iB \in B^i(\omega)$, then

$$A \succ_{i,\omega} B \text{ or } B \succ_{i,\omega} A \text{ (but not both)}$$

(2) *Transitivity:* i.e. if $b_iA, b_iB, b_iC \in B^i(\omega)$, then

$$A \succ_{i,\omega} B,\ B \succ_{i,\omega} C \implies A \succ_{i,\omega} C$$

Thus, the operator $\succ_{i,\omega}$has the same properties as the *strict preference* relation (Mas-Colell, Whinston, and Green 1995; page 6). Next, we define the following algorithm for updating beliefs:

Assumption 4. *Let ω_t and ω_{t+1} be the states of the world realised in periods t and $t+1$ respectively. Then, if $B \in B^i(\omega_t)$ and there exists no A in $B^i(\omega_t)$ such that $A \succ_{i,\omega_t} B$ and $(b_iA \wedge h_{t+1}) \implies \neg B$, then $b_iB \in \omega_{t+1}$. Otherwise, $b_iB \notin \omega_{t+1}$.*

Furthermore, we impose the condition that the precedence of beliefs does not change over time.

Assumption 5. *Let ω_t and ω_{t+1} be the states of the world realised in periods t and $t+1$ respectively. If $A, B \in B^i(\omega_t)$ and $A, B \in B^i(\omega_{t+1})$, then if $A \succ_{i,\omega_t} B$, then we must have $A \succ_{i,\omega_{t+1}} B$.*

We can now provide a formal definition of what is meant for one belief to *supercede* another:

Definition 5.1. *If $A, B \in B^i(\omega_t)$, and $A \succ_{i,\omega_t} B$, then person i believes in both statements A and B in state ω_t but his belief in statement A supersedes his belief in statement B.*

Definition 5.2. *Let ω_t and ω_{t+1} be the states of the world realised in periods t and $t+1$ respectively. If $A \succ_{i,\omega_t} B$ and $(b_i A \wedge h_{t+1}) \implies \neg B$, then person i abandons his belief in statement B in the sense that, by Assumption 4, $b_i B \notin \omega_{t+1}$.*

With these definitions, we can formally construct the argument made in Section 4.

Recall that the formula $(\neg m_i)$ corresponds to the statement 'person i is not immoral'. Suppose that, whenever $b_j(\neg m_i) \in \omega_t$ then $(\neg m_i) \succ_{j,\omega_t} A$ for each $A \in B^j(\omega_t)$; i.e. whenever person j believes in the statement $(\neg m_i)$, this belief supersedes all others.

It follows that, if $b_j(\neg m_i), b_j T \in \omega_t$ and $a^i_{w,t} = 1$, then, by Assumption 4, $b_j T \notin \omega_{t+1}$ where ω_{t+1} denotes the state realised in period $t+1$. In words, if person j believes in both statements $(\neg m_i)$ and T in some state ω_t, and person i commits the public act in period t, then person j must abandon belief in the statement T in the following period.

It follows that, if $b_l b_j(\neg m_i), b_l b_j T \in \omega_t$, and $a^i_{w,t} = 1$, then, by Assumption 4, $b_l b_j T \notin \omega_{t+1}$. In words, if person l believes that person j believes in both statements $(\neg m_i)$ and T, and person i commits the public act in period t, then person l must conclude that person j has abandoned belief in statement T in the following period.

Reasoning iteratively, it should be evident that, in a community of θ-types, if $(\neg m_i)$ is common belief within the community, and person i commits the public act in some period t, then $\neg T$ is common belief in the subsequent period. In the absence of any belief in the statement T, the social taboo against the public act cannot be sustained.

6. Conclusion

In this paper, we proposed a mechanism for sustaining a credible threat of sanctions in a population against some behaviour distinct from both the dominant economic and sociological approaches to the issue. The norm is underpinned by a simple moral code: 'a person who commits X is immoral'. Individuals in the population can vary in terms of whether or not they believe the statement is true, what they believe about what others believe, about what others believe they believe, etc. Nevertheless, we show that if it is regarded as true at some higher order level in the population – e.g. everyone believes that others believe that others believes ... that others believe the statement is true – there is an equilibrium in which everyone behaves as if the moral code were true.

In societies around the world, we find a variety of moral injunctions against behaviour of one sort or another: incest, blashphemy, adultery and so on. Whether, and to what extent people have internalised the moral code that underlie these injunctions (i.e. the incestuous, the blashphemous or the adulterous is immoral) is difficult to assess. But our result implies that, even if belief in the moral code is extremely 'weak' – in the sense that people may have only higher order beliefs regarding its veracity – there is an equilibrium in which they continue to respect the moral injunction.

Nevertheless, the moral code is critical in sustaining a credible threat of sanctions against the proscribed behaviour. In the model, it is common knowledge that one derives utility from ostracizing an 'immoral' person (although individuals can disagree on who is or isn't immoral) and this allows people to infer the private beliefs of others from their public actions. This is an example and reflexion of the assertion by Herbert Gintis that 'Humans have a social epistemology ... we have reasoning processes that afford us forms of knowledge and understanding, especially the understanding and sharing of the content of other minds, that are unavailable to merely "rational" creatures' (Gintis, 2009; page xv).

The theoretical mechanism suggests a particular strategy for bringing an end to inefficient or oppressive social norms. It requires that the moral code be contradicted by one whose own moral standing in the society is impeccable. If the norm were initially sustained purely through higher-order beliefs, then the fact that there is no belief in the moral code in the population becomes common knowledge after the statement of contradiction is made. Therefore, the social norm unravels. By contrast, if adherence to the norm is driven, not

by first-order and higher-order beliefs regarding a moral code but by expectations about other people's behaviour, there is no specific reason why such a statement would change people's behaviour regarding the norm.

7. Appendix A

In this section, we briefly consider alternative equilibria to the epistemic game analysed in sections 3.1-3.3. We assume throughout that the condition in (6) holds, and therefore, a type-0 individual would always ostracise a person whom they believed to be immoral.

First, we note that if $C > \frac{\beta(n-1)}{1-\beta}\delta_w \|W\|$, then it is not possible to have an equilibrium where type-0 individuals *always* ostracise *every* individual. The reason is that an act of ostracism against a person who is not immoral involves a cost of at least C, and the maximum reward that one can receive for such an act (when everyone is being 'subjected to ostracism' whatever their past actions) is $\frac{\beta(n-1)}{1-\beta}\delta_w \|W\|$.

However, it is possible to have equilibria where type-0 individuals ostracize those whom they do *not* believe to be immoral with some probability $\pi \in (0,1)$ if the following condition holds:

$$C < \frac{\beta\,(n-1)\,\delta_o}{1-\beta}\,(1-\pi)\,P \tag{7}$$

The condition in (7) ensures that the threat of being ostracized at all times, as opposed to sometimes, is sufficient to induce community members to engage in an act of ostracism even when it is costly.

Similarly, the threat of constant ostracism (as opposed to occasional ostracism) would be sufficient to dissuade community members from engaging in the public act if the following condition holds:

$$W < \frac{\beta\,(n-1)\,\delta_o}{1-\beta}\,(1-\pi)\,P \tag{8}$$

Therefore, we can use the reasoning in Section 3.3 to argue that if conditions (1), (7) and (8) hold, then the following constitutes an equilibrium:

Type-0 individuals play the strategy $s_0^i(\pi)$: If $e_{w,t}^i = 1$, then choose $\alpha_{w,t}^i = 0$. If $e_{o,t}^{ij} = 1$ then, if $b^i m_j$, choose $\alpha_{o,t}^{ij} = 1$; otherwise, choose $\alpha_{o,t}^{ij} = 1$ with probability π and $\alpha_{o,t}^{ij} = 0$ with probability $(1-\pi)$;

All higher types play the strategy $s_1^i(\pi)$: If $e_{w,t}^i = 1$, then choose $\alpha_{w,t}^i = 0$. If $e_{o,t}^{ij} = 1$ then, if $(h_t \wedge T) \implies m_j$, choose $\alpha_{o,t}^{ij} = 1$; otherwise, choose $\alpha_{o,t}^{ij} = 1$ with probability π and $\alpha_{o,t}^{ij} = 0$ with probability $(1-\pi)$.

Clearly, the conditions in (7) and (8) are more easily satisfied for smaller π. Moreover, among the class of strategy profiles taking the form $(s_0^i(\pi), s_1^i(\pi))$, lowering π leads to a Pareto improvement.

References

[1] Akerlof, George. "The Economics of Caste and of the Rat Race and Other Woeful Tales", *The Quarterly Journal of Economics*, Volume 90, 1976.

[2] Aumann, Robert. "Interactive Epistemology I: Knowledge", *International Journal of Game Theory*, 1999, Volume 28.

[3] Battigalli, P. and G. Bonanno. "Recent Results on Belief, Knowledge and the Epistemic Foundations of Game Theory", *Research in Economics*, Volume 53.

[4] Bicchieri, Cristina. "Social Norms", *The Standard Encyclopedia of Philosophy*, 2011.

[5] Chen, Yi-Chun. "A Structure Theorem for Rationalizability in the Normal Form of Dynamic Games", mimeo, National University of Singapore, 2011.

[6] Coate, S. and M. Ravallion. "Reciprocity without Commitment: Characterization and Performance of Informal Insurance Arrangements", *Journal of Development Economics*, Volume 40, 1993.

[7] Elster, Jon. "Social Norms and Economic Theory", *The Journal of Economic Perspectives*, Volume 3, 1989.

[8] Fafchamps, Marcel. "Solidarity Networks in Preindustrial Societies: Rational Peasants with a Moral Economy", *Economic Development and Cultural Change*, Vol. 41(1), October 1992.

[9] Farrell, J., and E. Maskin. "Renegotiation in repeated games", *Games and Economic Behavior*, 1989, Volume 1.

[10] Fudenberg, Drew and Jean Tirole. *Game Theory*, MIT Press, Cambridge, Massachusetts, 1991.

[11] Greif, Avner. "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition", *The American Economic Review*, Volume 83, 1993.

[12] Gintis, Herbert, *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*, Princeton University Press, 2009.

[13] Harsanyi, John. "Games with Incomplete Information Played by 'Bayesian' Players", *Management Science*, 1967, 1968.

[14] Hersholt, Jean. *The Complete Andersen*, The Limited Editions Club, New York, Volumes I-VI, 1949.

[15] Kimball, Miles. "Farmers' Cooperatives as Behavior towards Risk", *American Economic Review*, Volume 78, 1988.

[16] Kuran, Timur, *Private Truths, Public Lies: The Social Consequences of Preference Falsification*, Harvard University Press, 1995.

[17] Mackie, Gerry. "Ending Footbinding and Infibulation: A Convention Account", *American Sociological Review*, Volume 61, 1996.

[18] Mackie, Gerry. "Female Genital Cutting: The Beginning of the End", in Bettina Shell-Duncan and Ylva Hernlund, eds, Female Circumcision: Multidisciplinary Perspectives (Boulder, CO: Lynne Reinner Publishers)
[19] Mas-Colell, A., M. Whinston and J. Green. *Microeconomic Theory*. Oxford University Press, 1995.
[20] Parsons, Talcott. *The Social System*. Routeledge, New York, 1951.
[21] Roland, Gerard. "Understanding Institutional Change: Fast-Moving and Slow-Moving Institutions", *Studies in Comparative International Development*, Winter 2004, Volume 38.
[22] Rubinstein, Ariel. "The Electronic Mail Game: Strategic Behavior Under 'Almost Common Knowledge'", *The American Economic Review*, Volume 79, 1989.
[23] Weinstein, Jonathan and Muhamet Yildiz. "A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements", *Econometrica*, Volume 75(2), March 2007.
[24] Weinstein, Jonathan and Muhamet Yildiz. "A Structure Theorem for Rationalizability in Infinite-Horizon Games", mimeo, Massachusetts Institute of Technology, 2010.

University of Kent