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

Date: December 2012

z.wahhaj@kent.ac.uk. School of Economics, Keynes College, University of Kent, Canterbury CT2 7NP, United Kingdom.

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 repercussions 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 noone 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 anything. Immediately, everyone gives up the pretence. But if everyone already knew that

[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

the emperor was naked, should the child's declaration make any difference in people's behaviour? One possible explanation is that noone, 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 theoretical 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 i is in a position to 'engage in 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 'moral character' which may be 'good' or 'bad'. A community member will receive some psychological reward from ostracising a person who has 'bad moral character', and, therefore, would willingly engage in such an act of ostracism in the absence of any other incentives or disincentives.

What 'bad moral character' 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 have 'good' or 'bad moral character'. 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 c_i which describes his or her 'moral character': $c_i = 1$ if person i has 'good moral character' and $c_i = 0$ if he or she has 'bad moral character'. We assume that c_i is known to person i but unobservable to any other community member. Prior beliefs are given by $\Pr(c_i = 1) = 1 - \varepsilon$ where ε is positive but negligibly small. The payoffs in the stage-game are given by

$$(1)\quad u^i(a_i, a_{-i}, e) = -\sum_{j \neq i} \left[\mathbf{I}\left(e_o^{ji}\right) \alpha_o^{ji} P + \mathbf{I}\left(e_o^{ij}\right) \alpha_o^{ij} \left\{ Q - (1 - Ec_j) R \right\} \right] + \sum_{j \in \mathcal{I}} \mathbf{I}\left(e_w^j\right) \alpha_w^j W$$

where $a_i = (\alpha_o^i, \alpha_w^i)$, $\alpha_o^i = (\alpha_o^{ij})_{j \neq i}$, $e = (e_o^i, e_w^i)_{i \in \mathcal{I}}$, $e_o^i = (e_o^{ij})_{j \neq i}$ and $\mathbf{I}(e)$ is an indicator function which takes a value of 0 or 1 depending on whether or not event e has occurred. Q represents the cost of engaging in an act of social ostracism, and R is a reward from ostracizing a person with 'bad moral character'; 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 Q 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 c_i, $i \in \mathcal{I}$. This can be interpreted as meaning that they do not have any intrinsic views about the 'moral character' of the public act. Nevertheless, a 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. $(\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2..)$ *is a sequence of subsets of* $\mathcal{I}$ *defined as follows:*

$\mathcal{L}_0 = \emptyset$

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

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

The set $\mathcal{L}_t$ 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 $j \in \mathcal{L}_{t-1}$, choose $\alpha^{ij}_{o,t} = 1$; if $e^{ij}_{o,t} = 1$ and $j \notin \mathcal{L}_{t-1}$, 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 $i \notin \mathcal{L}_t$ and the alternate strategy $\bar{s}^i_2$ if $i \in \mathcal{L}_t$. This constitutes a subgame perfect equilibrium of the repeated game $\mathcal{G}(\infty)$ if

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

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

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

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

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

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

The 'blacklist' $\tilde{\mathcal{L}}_t$ is the opposite of $\mathcal{L}_t$. 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 $j \in \tilde{\mathcal{L}}_{t-1}$, choose $\alpha^{ij}_{o,t} = 1$; if $e^{ij}_{o,t} = 1$ and $j \notin \tilde{\mathcal{L}}_{t-1}$, 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 $i \notin \mathcal{L}_t$ 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. 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.

4. A Dynamic Framework for Modelling Interactive Knowledge and Beliefs

We denote by Ω_t the set of all possible states of the world in period t. A state will include information on the history of all past actions in the game, the 'type' of each player i, and other time-invariant, payoff-relevant, characteristics about the world.

Therefore, the set of states can be represented as follows:

$$\Omega_t \subseteq \mathcal{H}_t \times \prod_{i \in \mathcal{I}} \Theta_i \times \Sigma \tag{5}$$

where $\mathcal{H}_t$ is the set of all possible histories in period t; and Θ_i is the type-space for person i; and Σ the set of possible values for other time-invariant payoff-relevant characteristics of the world. For reasons we discuss later, not every element of the set represented on the right-hand side of (5) may be a feasible state; therefore we allow for the possibility that Ω_t is a subset of this set.

We define the function Γ_i as a mapping from player i's type to a subjective prior, defined on $\prod_{j \neq i} \Theta_j \times \Sigma$:

$$\Gamma_i : \Theta_i \to \Delta\left(\prod\nolimits_{j \neq i} \Theta_j \times \Sigma\right)^3 \tag{6}$$

Thus, a player's type describes what he or she believes about the types of the other players, and other time-invariant characteristics of the world at the beginning of the game. One's own beliefs about the types of other players include, by construction, one's beliefs about *their* beliefs regarding Σ, their beliefs about the types of others, etc. Thus, the mapping implicitly describes higher order beliefs.

4.1. **Belief and Knowledge Correspondences.**

We shall distinguish between beliefs and knowledge in the model. Informally, if one has 'knowledge' of an event, the event is necessarily true. By contrast, one may hold 'beliefs' that are false. To model beliefs and knowledge, we adopt the knowledge and belief framework presented by Battigalli and Bonanno (1999) (which the authors call a '*KB-frame*').

Battigalli and Bonanno (1999) define a 'belief frame' as a tuple $\mathcal{F} = (\Omega, P)$, where Ω is a set of possible states, and $P : \Omega \to 2^{\Omega}$ is a correspondence. They define the following properties that a belief frame may satisfy:

- *Seriality :* $\forall \omega \in \Omega,\ P(\omega) \neq \emptyset$
- *Reflexivity :* $\omega \in P(\omega)$
- *Transitivity :* $\forall \alpha, \beta \in \Omega$, if $\beta \in P(\alpha)$ then $P(\beta) \subseteq P(\alpha)$
- *Euclideanness :* $\forall \alpha, \beta \in \Omega$, if $\beta \in P(\alpha)$ then $P(\alpha) \subseteq P(\beta)$

Given the state space defined in the preceding section, we define a *knowledge correspondence* and a *belief correspondence* for each player i: $\mathcal{K}_t^i : \Omega_t \to 2^{\Omega_t}$, $\mathcal{B}_t^i : \Omega_t \to 2^{\Omega_t}$ for $t = 1, 2, 3, ...$ Each correspondence $\mathcal{K}_t^i$ will satisfy *reflexivity, transitivity and euclideanness*, while $\mathcal{B}_t^i$ will satisfy *seriality, transitivity and euclideanness.* It is easy to verify that the knowledge correspondence $\mathcal{K}_t^i$, with the afore-mentioned properties, is equivalent to person i's information set. While 'reflexivity' implies that one never rules out the true state of the world in terms of one's knowledge, no such restriction is imposed regarding one's beliefs.

Beliefs and knowledge will be linked together through the following conditions, adopted from Battigalli and Bonanno (1999):

(R1) $\mathcal{B}_t^i(\omega) \subseteq \mathcal{K}_t^i(\omega)$

(R2) if $\omega' \in \mathcal{K}_t^i(\omega)$ then $\mathcal{B}_t^i(\omega') = \mathcal{B}_t^i(\omega)$

Condition (R1) implies that if a state can be ruled out on the basis of one's knowledge, then it does not belong in one's belief set. Condition (R2) implies that if two states are indistinguishable in terms of one's knowledge, then one should also hold the same beliefs in those two states.

4.2. **A Syntactic Language.**

Our aim in this paper is to articulate a theory about social norms in which a certain maxim about moral character is central to how individuals form beliefs about others. The framework introduced in the preceding sections is sufficient for describing both the first-order and higher-beliefs of each individual, and specifying how these beliefs evolve according to the history of actions within the game. However, it will be convenient and intuitive to formulate the maxim about moral character using a syntactic language. The syntactic language will be based on the semantic framework used in the preceding sections, and any reasoning using the former can also be done using the latter.

The following definitons and notation are adopted from Battigalli and Bonnano (1999) and Aumann (1999). The syntactic language consists of letters of an 'alphabet' $\mathfrak{X}$ (representing atomic propositions) and the symbols $\neg$, $\vee$, (,), and b_i and k_i for each $i \in \mathcal{I}$.

Given $\mathfrak{X}$, we can construct a *formula* according to the following rules:

(1) each $x \in \mathfrak{X}$ is a formula
(2) if x, y are formulae, then $(x) \vee (y)$ is a formula
(3) if x is a formula, then so are $\neg (x)$, $k_i (x)$ and $b_i (x)$, for each $i \in \mathcal{I}$

We denote by Φ the set of all formulae derived from $\mathfrak{X}$ using these rules. We use the symbols $\implies$ and $\wedge$, used as in $x \implies y$ and $(x) \wedge (y)$, as abbreviations for $(\neg (x) \vee y)$ and $\neg (\neg (x) \vee \neg (y))$ respectively. Parantheses may be omitted if doing so does not result in any ambiguity. This syntactic language should be interpreted as follows. The symbol k_i stands for 'person i knows that ...', while b_i stands for 'person i believes that...'. The symbol $\neg$ stands for 'not', and $\vee$ stands for 'or'. From the definition of $\implies$, one can verify that the symbol retains its standard meaning in mathematics, and stands for 'implies that ...'; while $\wedge$ stands for 'and'.

The atomic propositions contained in $\mathfrak{X}$ can be used to give content to the states of the world described earlier. To do this, we introduce the functions $f_t : \mathfrak{X} \rightarrow 2^{\Omega_t}$, $t = 1, 2, ...$ The function f_t provides a mapping from each atomic proposition to the set of states in period t where they hold true. For each $\phi, \psi \in \Phi$, we define the 'truth set' in period t, denoted by $\|\phi\|_t$ or $\|\psi\|_t$, as a subset of Ω_t constructed according to the following recursive rules:

(i) If $\phi = x$ where $x \in \mathfrak{X}$, then $\|\phi\|_t = f_t(x)$
(ii) $\|\neg\phi\|_t = \Omega_t \setminus \|\phi\|_t$
(iii) $\|\phi \vee \psi\|_t = \|\phi\|_t \cup \|\phi\|_t$
(iv) $\|b_i\phi\|_t = \{\omega \in \Omega_t : \mathcal{B}_t^i(\omega) \subseteq \|\phi\|_t\}$
(v) $\|k_i\phi\|_t = \{\omega \in \Omega_t : \mathcal{K}_t^i(\omega) \subseteq \|\phi\|_t\}$

The recursive rules simply establish in which states of the world a particular formula holds true by giving the standard interpretation to the symbols $\neg$ and $\vee$. They also give precise meaning to the symbols b_i and k_i: according to rule (iv), the formula $b_i\phi$ holds true in some state ω, if ϕ holds true in all the states that person i believes are possible when state ω is realised; the formula $k_i\phi$ holds true in some state ω, if ϕ holds true in all the states that person i cannot rule out in terms of his or her knowledge when state ω is realised.

Using the definitions of the 'truth sets', we can establish the two following lemmas which are key to any reasoning done using the syntactic language.

Lemma 4.1. *For each $\phi, \psi \in \Phi$, if $\omega \in \|\phi\|_t$ and $\omega \in \|\phi \implies \psi\|_t$ in some period t, then $\omega \in \|\psi\|_t$.*

Lemma 4.2. *For each $\phi, \psi \in \Phi$, if $\omega \in \|b_i\phi\|_t$ and $\omega \in \|b_i(\phi \implies \psi)\|_t$ in some period t for some $i \in \mathcal{I}$, then $\omega \in \|b_i\psi\|_t$. For each $\phi, \psi \in \Phi$, if $\omega \in \|k_i\phi\|_t$ and $\omega \in \|k_i(\phi \implies \psi)\|_t$ in some period t for some $i \in \mathcal{I}$, then $\omega \in \|k_i\psi\|_t$.*

According to Lemma 4.1, if the formulae ϕ and $\phi \implies \psi$ hold true in some state ω, then the formula ψ must also hold true in state ω. According to Lemma 4.2, if the formulae $b_i\phi$ and $b_i(\phi \implies \psi)$ hold true in some state ω (i.e. person i believes ϕ and person i believes that ϕ implies ψ), then $b_i\psi$ also holds true in state ω. And the same result holds for the knowledge symbol k_i.

4.3. **The Evolution of Beliefs.** Next, we specify how the belief and knowledge correspondences, defined in Section 4.1, relate to the subjective priors, and how beliefs evolve in the game. We assume that, in each period, each type of each player has knowledge of the history of the game, and nothing else:

Assumption 1. *If h_t is the history corresponding to state ω in period t, and $E(h_t) \subset 2^{\Omega_t}$ is the event that the history h_t has been realised, then $\mathcal{K}_t^i(\omega) = E(h_t)$.*

A player's beliefs at the start of the game, before any actions have taken place should, intuitively, correspond to the support of the subjective priors; i.e. the set of states to which a person assigns positive probability at the start of the game. Therefore, for player i of type θ_i, we let

$$\mathcal{B}_0^i = \left\{ \omega \in \prod_{i \in \mathcal{I}} \Theta_i \times \Sigma : p_{\theta_i,0}^i(\omega) > 0 \right\} \tag{7}$$

where $p_{\theta_i,0}^i(.)$ is the function generated by the mapping Γ^i. By Assumption (1), player i has knowledge of the updated history of the game in each of the subsequent periods. We assume that he revises his subjective probabilities on the basis of this new knowledge using Bayes' rule. To be precise, let $h_t = (h_{t-1}, a_t)$ be the history realised in period t and let a_t be the period t action profile corresponding to this history. Let ω_t be a possible period t state of the world, corresponding to the action profile a_t subsequent to state ω_{t-1} in period $t-1$ (note that the complete history h_t need not hold true in state ω_t). For a given strategy profile (which will be defined in more detail in the subsequent sections), we can compute the conditional *objective* probability $\sigma_t(a_t|\omega_{t-1})$ that the action a_t will take place after state ω_{t-1} has been realised. Then, the players' subjective probability that the true state of the world is ω_t, conditional on history h_t, can be computed as follows:

$$p_{\theta_i,t}^i(\omega_t|h_t) = \frac{p_{\theta_i,t-1}^i(\omega_{t-1}|h_{t-1})\,\sigma_t(a_t|\omega_{t-1})}{\sum\limits_{\omega'_{t-1} \in \Omega_{t-1}} p_{\theta_i,t-1}^i\left(\omega'_{t-1}|h_{t-1}\right)\sigma_t\left(a_t|\omega'_{t-1}\right)} \tag{8}$$

Thus, equation (8) gives player i's subjective probability that state ω_t has been realised in period t, when he observes history h_t, using his subjective probability function $p_{\theta_i,t-1}^i(.|h_{t-1})$ from the previous period. The belief sets from period 1 onwards should correspond to these revised probabilities. To be precise, if ω_t is the true state in period t and h_t is the corresponding history, then player i's belief set can be written as

$$\mathcal{B}_t^i(\omega_t) = \left\{\omega \in \Omega_t : p_{\theta_i,t}^i(\omega|h_t) > 0\right\} \tag{9}$$

Equation (8) provides a valid procedure for updating player i's subjective probabilities after observing the actions a_t if and only if, in the preceding period, he had assigned positive probabilities to at least some states in which action a_t is chosen with positive probability, i.e. the denominator of (8) is positive.

5. Applying the Knowledge-Belief Framework to the Social Norms Game

In this section, we describe how the knowledge-belief framework presented above can be used to represent and analyse the role of beliefs in the game introduced at the beginning of Section 3.

The history that is relevant to the game is the move by nature (which determines which random event will occur) and the choice of action by the player who is required to take an action when that event occurs. Therefore, we denote nature's set of possible actions in any period t by $\mathcal{E} = \{e_o^{ij} : i, j \in \mathcal{I}, i \neq j\} \cup \{e_w^i : i \in \mathcal{I}\}$, and represent the relevant actions in a period as a tuple $(e, a) \in \mathcal{E} \times \{0, 1\}$. Thus, the tuple $(e_w^i, 0)$, for example, indicates that person i had an opportunity to commit the public act but chose not to commit the act. The relevant history from the beginning of the game up to period t can be written as $h_t = (e_1, a_1, e_2, a_2, ..., e_t, a_t)$ where e_τ denotes the move by nature, and a_τ the choice of action by the relevant player, in period τ. So, the set of possible histories in period t is given by

$$\mathcal{H}_t = \{\mathcal{E} \times \{0, 1\}\}^t$$

For each player i, we represent the set of possible types by $\Theta_i = \{0, 1, 2, 3, ...\}$. Besides the player types, the time-invariant characteristics of the game will include the moral character of each player. For each person i, we denote moral character by the variable c_i which takes a value of 0 or 1; we use $c_i = 1$ to mean that person i has 'good moral character' and $c_i = 0$ to mean that person i has 'bad moral character'. Furthermore, in each state of the world, a particular maxim about moral character, to be defined below, will be either true or false. We represent these possibilities by a variable μ which takes a value of 0 if the maxim is false and 1 if the maxim is true. So we can represent the set of time-invariant payoff-relevant characteristics by $\Sigma = \{0, 1\}^{n+1}$.

The following will constitute the 'alphabet' of the syntactic language. Let θ_{ix} be the occurrence that person i has type-x. Let $w_{i,\tau}$ 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'. With some abuse of notation, let c_i be the occurrence that person i 'has good moral character'. Thus, the 'alphabet' is given by[4]

$$\begin{aligned}\mathfrak{X} &= \{\theta_{ix} : i \in \mathcal{I}, x \in \Theta_i\} \cup \{o_{ij,t} : i, j \in \mathcal{I}, i \neq j, t \in \mathbb{N}\} \\ &\quad \cup \{w_{i,t} : i \in \mathcal{I}, t \in \mathbb{N}\} \cup \{c_i : i \in \mathcal{I}\}\end{aligned}$$

For ease of notation, we may drop the time subscript when using the alphabet if the exact time period of the occurrence is not relevant. Thus, w_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.' Using this alphabet, we can describe a 'maxim' regarding moral character as follows:

(i) $m_0 = (w_i \implies \neg (c_i)$ for each $i \in \mathcal{I})$
(ii) $m_n = (\neg (b_i m_{n-1}) \implies \neg (c_i)$ for each $i \in \mathcal{I})$ for $n = 1, 2, ...$
(iii) $m = m_0 \wedge m_1 \wedge m_2 ...$

In words, the formula m_0 says that 'anyone who commits the public act has bad moral character'; m_n says that 'anyone who does not believe in m_{n-1} has bad moral character' where n is a positive integer. Finally, m can be interpreted as saying that 'anyone who commits the public act has bad moral character, and anyone who contradicts this maxim in any way also has bad moral character.'

Given m, we can construct truth-sets $\|m\|_t$ and $\|\neg m\|_t \subseteq \Omega_t$ using the recursive rules defined in the previous section to identify the states in which the maxim holds true. Recall that we also associated a variable μ with each state of the world to indicate whether the maxim holds in a particular state. The reason for constructing these two alternative ways for indicating whether the maxim holds true or not in each state will become clear in the following discussion. To ensure that the two approaches are consistent, we limit the set of possible states to the subset $\Phi_t \subset \Omega_t$, defined as follows.

$$\Phi_t = \{\omega \in \Omega_t : (\mu = 1 \ \& \ \omega \in \|m\|_t) \text{ or } (\mu = 0 \ \& \ \omega \in \|\neg m\|_t)\}$$

The subjective priors, represented by the mapping $\Gamma_i : \Theta_i \to \Delta\left(\prod_{j \neq i} \Theta_j \times \Sigma\right)$, defines for each type of a person whether he believes the maxim is true and his beliefs about the type and moral character of each of the other players. We specify subjective prior beliefs as follows. Any player has bad moral character with probability $\varepsilon > 0$; players of type 0 believe that the maxim is true, while higher types believe that the maxim is false. A

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

player of type-x, where $x \in \{0, 1, 2, 3, ...\}$ believe all other players have a type between 0 and $(x - 1)$.

5.1. **Strategies and Equilibrium.** We represent player' i's strategy using a sequence of functions of the form $\sigma_t^i : \mathcal{H}_{t-1} \times \mathcal{E} \times \Theta_i \longrightarrow [0, 1]$ where $t \in \mathbb{N}$. The function σ_t^i specifies the probability with which person i's chooses a specific action in period t, contigent on the past history, nature's move in the current period and person i's type. Specifically, $\sigma_t^i(h_{t-1}, e_w^i, \theta)$ denotes the probability that player i of type θ chooses $a_w^i = 1$ (i.e. chooses to engage in the public act) when the event e_w^i occurs following history h_{t-1}, and $\sigma_t^i(h_{t-1}, e_o^{ij}, \theta)$ denotes the probability that player i of type θ chooses the action $a_o^{ij} = 1$ (i.e. chooses to ostracise person j) when event e_o^{ij} occurs following history h_{t-1}.

We represent person i's full strategy by $\sigma^i = (\sigma_t^i)_{t \in \mathbb{N}}$ and a strategy profile of the game by $\sigma = (\sigma_i)_{i \in \mathcal{I}}$. Using σ, and the prior beliefs $p_{\theta_i,0}^i(.)$ for each player $i \in \mathcal{I}$ and each player type $\theta_i \in \Theta_i$, we can compute the posterior beliefs of each player at each information set, $E(h_t)$.

We define an indirect utility function $V^i(.)$ as follows:

$$V^i(\sigma_i, \sigma_{-i}) = \sum_{t=1}^{\infty} \beta^{t-1} \sum_{h_t \in \mathcal{H}_t} \Pr(h_t|\sigma)\, u^i(a_i, a_{-i}, e)$$

where $h_t = (h_{t-1}, a, e)$, $a = (a_i, a_{-i})$ and $u^i(.)$ is as defined in (1). We define an equilibrium as a strategy profile σ, prior beliefs $p_{\theta_i,0}^i(.)$ and posterior beliefs $p_{\theta_i,t}^i(.)$ such that

$$\sigma_i \in \arg\max_{\sigma_i} EV^i(\sigma_i, \sigma_{-i})$$

and at each information set $E(h_t)$ that person i believes will be reached with positive probability given $p_{\theta_i,t-1}^i(.)$, beliefs will be updated using Bayes' rule as described in (8). At each information set $E(h_t)$ that person i believes will be reached with zero probability, beliefs will satisfy the *consistency* criterion proposed by Kreps and Wilson (1982).

5.2. **Characterisation of Equilibria of the Epistemic Game.** Next, 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 has bad moral character. Therefore, if the disutility from associating with a person of bad moral character (represented by the variable R) is sufficiently high, a type-0 individual would ostracise one who has committed the public act, regardless of the strategies pursued by others.[5]

If an individual fails to ostracise someone who has committed the public act, it implies that he does not believe that the maxim is true. The maxim implies that such a person has bad moral character. Therefore, a type-0 person, who believes that the maxim is true, will conclude that such a person also has bad moral character and choose to ostracise him. Reasoning in the same manner, we can show that a type-0 individual will also ostracise anyone who has failed to ostracise someone who has committed the public act, anyone who has failed to ostracise someone who has failed to ostracise someone who has committed the public act, etc.

Furthermore, since a type-0 individual believes that everyone else in the community is of type-0, who, by definition, believe that the maxim is true, she would expect to be ostracized by everyone were she to engage in the public act. Therefore, if the condition in (2) holds, she would refrain from doing so.

There remains only the question of whether, in an equilibrium, a type-0 individual would ostracise someone in situations other than those described above. Given prior beliefs, in these situations she assigns the person a high probability (close to 1) of *good* moral character. Ostracising such a person carries a cost of P and no significant reward. Therefore, she would do so only if she is given additional incentives for it. Since we have argued that type-0 individuals will always refrain from engaging in the public act, this cannot be used to provide the appropriate incentives These incentives cannot take the form of a threat of ostracism from other community members if one expects to be ostracised by others every period in any event.

The only possible type of equilibrium where individuals with high probability of good moral character are ostracised is if such ostracism occurs with some probability less than

[5] 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 $\frac{\beta(n-1)}{1-\beta}\left(\delta_o P + \delta_w \|W\|\right)$. Therefore, if

$$R - Q > \frac{\beta\,(n-1)}{1-\beta}\left(\delta_o P + \delta_w \|W\|\right) \tag{10}$$

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

one and the behaviour is supported by the threat of increased frequency of ostracism against those who fail to carry it out. We can rule out such strategies using the Farrell-Maskin criteria of renegotiation-proofness (Farrell and Maskin, 1989). This is because the players can switch to a pareto-superior equilibrium.by choosing not to ostracise someone who has failed to ostracise someone with good moral character. Therefore, in any renegotiation-proof equilibrium, type-0 individuals will not ostracise others when they have a high probability of good character.

Therefore, if R is sufficiently large, the following is the only possible strategy for type-0 individuals in a renegotiation-proof equilibrium.

$$\sigma_t^i\left(h_{t-1}, e_w^i, 0\right) = 1 \tag{11}$$

$$\sigma_t^i\left(h_{t-1}, e_o^{ij}, 0\right) = 1 \text{ if } j \in \mathcal{I}_b\left(h_{t-1}\right) \tag{12}$$

$$\sigma_t^i\left(h_{t-1}, e_o^{ij}, 0\right) = 0 \text{ if } j \notin \mathcal{I}_b\left(h_{t-1}\right) \tag{13}$$

where $\mathcal{I}_b\left(h_t\right) = \left\{j \in \mathcal{I} : (e_w^j, 1) \in h_t \text{ or } (\left(e_o^{jl}, 0\right) \in h_t \text{ and } l \in \mathcal{I}_b\left(h_{t-1}\right))\right\}$. In words, the specified strategy is as follows: 'Do not engage in the public act. Ostracise anyone who previously engaged in the public act, or failed or ostracise someone who engaged in the public act, or failed to ostracise someone who failed to ostracise someone who engaged in the public act, etc. Do not ostracise others.'

Recall that type-1 individuals believe that all other community members are of type-0. Therefore, in an equilibrium where type-0 individuals are playing strategy specified in (11)-(13), a type-1 individual 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 fails to ostracise someone who has committed the public act, or fails to ostracise someone who has failed to ostracise someone who has committed the public act, etc., then type-0 individuals would conclude that she has bad moral character, and ostracise her thereafter. Therefore, it is optimal for her to ostracise anyone who has engaged in the public act, ostracise anyone who has failed to do the same, and so on if (3) holds. Moreover, she has no incentives to ostracise anyone in other situations. Therefore, if the type-0 individuals are playing the strategy specified in (11)-(13), then under conditions (2) and (3), the following is the unique optimal strategy for a type-1 individual.

$$\sigma_t^i\left(h_{t-1}, e_w^i, 1\right) = 1 \tag{14}$$

$$\sigma_t^i\left(h_{t-1}, e_o^{ij}, 1\right) = 1 \text{ if } j \in \mathcal{I}_b\left(h_{t-1}\right) \tag{15}$$

$$\sigma_t^i\left(h_{t-1}, e_o^{ij}, 1\right) = 0 \text{ if } j \notin \mathcal{I}_b\left(h_{t-1}\right) \tag{16}$$

Recall that type-2 individuals believe that all other community members are of type-1. Given the strategy for type-1 individuals specified in (14)-(16), we can show, using the same reasoning as above, that under conditions (2) and (3), type-2 individuals have the same optimal strategy. By reasoning iteratively, we can show that the same strategy is optimal for all higher types.

We have now established the following.

Proposition 5.1. *If the conditions in (2), (3) and (10) hold, and type-0 individuals are restricted to strategies that are renegotiation-proof, the unique optimal strategy for all types is to refrain from the public act, and ostracise no-one except those who have previously engaged in the public act, or failed to ostracise someone who has engaged in the public act, or failed to ostracise someone who has failed to ostracise someone who has engaged in the public act, etc.*

The reasoning behind Proposition 5.1 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.

5.3. 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 5.1. 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 'bad moral character' embodied in the maxim m 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 'bad moral character' but they all believe that everyone else does. They follow the behaviour implicitly prescribed by the maixim m to hide their true beliefs, because they fear being accused of bad moral character otherwise.

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

Necessity of Common Knowledge of the Notion of 'Moral Character': An important element of the equilibrium described in Proposition 5.1 is the psychological reward R that one obtains from ostracising a person of 'bad moral character'. Without this reward, there is no reason why belief in the maxim m should affect a person's behaviour. Also, unless the reward R is common knowledge, the reasoning used in Proposition 5.1 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 a person of 'bad moral character', whatever 'moral character' 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 5.1 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 5.2, 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.

6. The Dynamics of Social Taboos

In Proposition 5.1, we described an equilibrium in which no-one engages in the public act and no-one chooses to ostracise another person in any period. Consequently, beliefs about moral character and the truth of the maxim do not change over time; individuals retain their prior beliefs throughout the game.

However, an equilibrium of this kind can unravel suddenly; a long-standing social taboo may be given up collectively and permanently within a single period. We can use the theoretical framework developed in Section 5 to see how this would happen.

So that beliefs regarding the maxim may evolve throughout the game, we need to introduce some 'doubt' about the truth of the maxim for type-0 players. Instead of assuming that type-0 players believe in the maxim, we assume that they believe that the maxim is false with probability $\delta > 0$ (and true with probability $1 - \delta$), where δ may be infinitesimally small. We retain all other assumptions about prior beliefs described in Section 5.

Then, if an individual engages in the public act, or fails to ostracise someone who has engaged in the public act, etc. a type-0 individual will update her subjective probability that the individual has bad moral character from ε to $\frac{\varepsilon}{(1-\varepsilon)\delta+\varepsilon}$.[6] It is evident that if δ is close to zero and small relative to ε, the latter expression is close to 1. Therefore, all the reasoning in Section 5.2 will still go through.

Suppose there is an individual l in the community who has a much 'stronger reputation' of good moral character than other community members. Specifically, suppose that each player's prior subjective probability that l has bad moral character is $\varepsilon_l < \varepsilon$. If l engages in the public act, this probability will be revised up to $\frac{\varepsilon_l}{(1-\varepsilon_l)\delta+\varepsilon_l}$; but note that if ε_l is small relative to δ – which would mean that type-0 individuals have more confidence in the good moral character of l than the truth of the maxim – then this expression, close to $\frac{\varepsilon_l}{\delta}$, is smaller than 1. Then the posterior probability that l has bad moral character may not be sufficiently high for type-0 individuals to ostracise him in the absence of other incentives.

[6] To see this, note that if an individual i engages in the public act in some period t, then the posterior belief of an individual j, of type-0, that the former has bad moral character is given by

$$p^j_{\theta_j,t+1}\left(\|\neg(c_i)\|_{t+1} \mid \left(h_{t-1}, e^i_w, 1\right)\right)$$

$$\begin{aligned}
&= p^j_{\theta_j,t+1}\left(\|\neg(c_i) \wedge m\|_{t+1} \mid \left(h_{t-1}, e^i_w, 1\right)\right) \\
&= \frac{\left[p^j_{0,t}\left(\|\neg(c_i) \wedge m\|_t\right) + p^j_{0,t}\left(\|\neg(c_i) \wedge \neg(m)\|_t\right)\right]\sigma^i_t\left(h_{t-1}, e^i_w, 1\right)}{\left[p^j_{0,t}\left(\|\neg(c_i) \wedge m\|_t\right) + p^j_{0,t}\left(\|c_i \wedge \neg(m)\|_t\right) + p^j_{0,t}\left(\|\neg(c_i) \wedge \neg(m)\|_t\right)\right]\sigma^i_t\left(h_{t-1}, e^i_w, 1\right)} \\
&= \frac{\varepsilon(1-\delta) + \varepsilon\delta}{\varepsilon(1-\delta) + (1-\varepsilon)\delta + \varepsilon\delta} \\
&= \frac{\varepsilon}{(1-\varepsilon)\delta + \varepsilon}
\end{aligned}$$

Therefore, we can reason that a person with a 'strong reputation' of good moral character will, indeed, engage in the public act. Moreover, when he commits the public act, type-0 individuals will revise downward their subjective probability that the maxim is true, to $\frac{(1-\delta)\varepsilon_l}{(1-\delta)\varepsilon_l+\delta}$.[7] If ε_l is small relative to δ, then the posterior probability will be close to 0. In this case, type-0 individuals will cease to ostracise individuals who commit the public act in subsequent periods; and cease to ostracise individuals who fail to ostracise anyone who has committed the public act, etc. Then, it would not be optimal for higher types to follow the strategy prescribed by Proposition 5.1. Thus, the equilibrium will unravel.

7. 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 has bad moral character'. 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 have bad moral character) is difficult to

[7] To see this, note that if individual l engages in the public act in some period t, then the posterior belief of an individual j, of type-0, that the maxim is true is given by

$$p^j_{0,t+1}\left(\|m\|_{t+1} \mid \left(h_{t-1}, e^l_w, 1\right)\right)$$

$$\begin{aligned}
&= p^j_{0,t+1}\left(\|\neg\left(c_i\right) \wedge m\|_{t+1} \mid \left(h_{t-1}, e^l_w, 1\right)\right) \\
&= \frac{\left[p^j_{0,t}\left(\|\neg\left(c_i\right) \wedge m\|_t\right)\right]\sigma^l_t\left(h_{t-1}, e^l_w, 1\right)}{\left[p^j_{0,t}\left(\|\neg\left(c_i\right) \wedge m\|_t\right) + p^j_{0,t}\left(\|c_i \wedge \neg\left(m\right)\|_t\right) + p^j_{0,t}\left(\|\neg\left(c_i\right) \wedge \neg\left(m\right)\|_t\right)\right]\sigma^l_t\left(h_{t-1}, e^l_w, 1\right)} \\
&= \frac{\varepsilon_l\left(1-\delta\right)}{\varepsilon_l\left(1-\delta\right) + \left(1-\varepsilon_l\right)\delta + \varepsilon_l\delta} \\
&= \frac{\varepsilon_l}{\left(1-\varepsilon_l\right)\delta + \varepsilon_l}
\end{aligned}$$

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 a person of 'bad moral character' (although individuals can disagree on who has or hasn't 'bad moral character') and this allows people to infer the private beliefs of others from their public actions. This is an example and reflection 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.

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

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

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

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

[17] 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)

[18] Mas-Colell, A., M. Whinston and J. Green. *Microeconomic Theory.* Oxford University Press, 1995.

[19] Parsons, Talcott. *The Social System.* Routeledge, New York, 1951.

[20] Roland, Gerard. "Understanding Institutional Change: Fast-Moving and Slow-Moving Institutions", *Studies in Comparative International Development*, Winter 2004, Volume 38.

[21] Rubinstein, Ariel. "The Electronic Mail Game: Strategic Behavior Under 'Almost Common Knowledge'", *The American Economic Review*, Volume 79, 1989.

[22] Weinstein, Jonathan and Muhamet Yildiz. "A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements", *Econometrica*, Volume 75(2), March 2007.

[23] Weinstein, Jonathan and Muhamet Yildiz. "A Structure Theorem for Rationalizability in Infinite-Horizon Games", mimeo, Massachusetts Institute of Technology, 2010.

University of Kent