HIGHER-ORDER BELIEFS AND THE PERSISTENCE OF SOCIAL NORMS

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., collective participation in social sanctions occurs in any equilibrium if (certain Folk-theorem-type conditions hold and) there are higher order beliefs, at some finite order n and above. 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 and above.

JEL Codes: D01, D02, D83, Z10

Date: October 2015

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

1. Introduction

Why people follow social norms and how social norms evolve are questions that have received a great deal of attention within the social sciences (see, for example, Elster 1989, Fehr and Fischbacher 2004 and Bicchierri 2011 for reviews).

The dominant view within sociology is that people are 'hard-wired' to follow social norms and 'hard-wired' to inflict a punishment on those who deviate from them. The internalisation of norms plays an important in, for example, Talcott Parsons' theory of socialisation (Parsons 1951). While this premise provides a robust answer to the question why social norms persist over long periods of time, it provides no insight about behaviour when norms have not been internalised.

A contrasting view that has emerged within economics is that people weigh the costs and benefits of following a norm. If an individual deviates from the norm he or she expects to face social sanctions from others; and sanctioning behaviour itself is sustained by the threat of social sanctions from others. These theories can help to characterise the set of conditions under which a particular social norm can be sustained. Well-known theories which make use of these mechanisms include George Akerlof's explanation of the endurance of the caste system in India (Akerlof 1976) and Avner Greif's explanation of contract enforcement in medieval trade (Greif 1993). But they typically generate multiple equilibria, and no ready answer as to why one equilibrium as opposed to another is obtained in practice.

In this paper, we present an alternative to these theories that offers insights about the nature of social norms and their persistence over time. Our underlying assumption is that people derive utility from ostracising those who they believe to have 'bad character'. But people differ in their beliefs about how the character of a person may be inferred. We consider a maxim about bad character which may be expressed as follows: 'Anyone who engages in an act X has bad character. Anyone who associates with someone with bad character also has bad character.'

If belief in this maxim is shared within a population then, whenever a person i engages in act X, people will update their beliefs about i to 'bad character'; and ostracise i thereafter whenever they have the opportunity to do so. It is evident that if experiencing ostracism is sufficiently unpleasant, people will refrain from act X. In this paper, we show the

surprising result that we obtain the same outcome – i.e. people refrain from act X – even if people do not hold these beliefs, and do not believe that others hold these beliefs, and do not believe that others believe that others hold these beliefs, etc. but the beliefs hold true above some finite nth order. This equilibrium is unique and renegotiaton-proof (Farrell and Maskin, 1989).

The intuition behind this result is as follows: a person j, who does not believe in the maxim himself but believes that others do, will behave as if the maxim is indeed true (i.e. avoid engaging in act X, and ostracising those who do) to avoid being ostracised himself (assuming that being ostracised is sufficiently costly). Then a person k who neither believes in the maxim nor believes that others do, but believes that others have the same beliefs as j will also behave as if the maxim is indeed true to avoid being ostracised by people like j. We can continue with this reasoning *ad infinitum*. Thus people refrain from act X even when no one believes in the maxim, and no one believes that others do, and no one believes that others believe that others do, etc. so long as the maxim is held to be true above some finite nth order. A necessary condition for people to engage in act X is that the absence of belief in the maxim is common knowledge.

The implication of this result is that a social norm may persist long after people have stopped believing in the moral code or maxim that triggered social sanctions against anyone who engaged in act X. Furthermore, an event which causes the lack of belief to become common knowledge would cause the social norm to unravel suddenly. The literature documents a variety of cases where a social norm has endured over long periods with little change followed by abrupt decline (see, for example, Bicchierri 2011). The theory presented here provides a potential explanation for such phenomena.

The absence of common knowledge – about whether certain social norms have been internalised in a population – is plausible when the norms in question deal with sensitive social issues that are not often discussed in public. Kuran (1995) provides a range of historical examples – from the Indian caste system to racial affirmative action in the United States – where individuals are engaged in 'preference falsification'; i.e. they refrained from actions that express their true beliefs or preferences for fear of the repercussions that such a revelation would bring, thus leading to situations where true beliefs regarding the social norm may not have been common knowledge.

The main technical result in this work is akin to that 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.

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 (2013) 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.

This paper is also broadly related to a rich theoretical literature within economics on 'herding' or 'conformist' behaviour within a population. Potential drivers of such behaviour that have been considered in the literature include positive payoff externalities that generate strategic complementarities (Cooper and John, 1988; Chamley 1999), status or reputational concerns that directly affect an agent's utility (Bernheim 1994; Kuran 1987), and the imitation of the behaviour of potentially better-informed agents (Banerjee 1992). Brock and Durlauf (2001) analyses a model of social interactions that also generates conformist behaviour and nests a number of these mechanisms. However, to the best of our knowlege, the potential role of higher-order beliefs in sustaining norms of social behaviour has not been explored in this literature.

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 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 types of behaviour although they have not internalised the norms that forbid such behaviour) 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. The formal model is presented in Section 2. In sections 2.1-2.5, we develop a framework for analysing the role of higher-order beliefs in maintaining social norms. Section 2.6 demonstrates the main result of the paper, and properties of the related equilibria are discussed in Section 2.7. The dynamic implications of the model are explored in Section 3.

2. 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 $i, j \in \mathcal{I}$, $i \neq j$. 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 can, potentially, play a critical role in sustaining a social taboo, and a credible threat of social ostracism.

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 unobservable to any community member, even for person i. Prior beliefs are given by $\Pr(c_i = 1) = 1 - \varepsilon$ where $\varepsilon > 0$. 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 $a_i = (\alpha_o^i, \alpha_w^i)$, $\alpha_o^i = \left(\alpha_o^{ij}\right)_{j \neq i}$, $e = \left(e_o^i, e_w^i\right)_{i \in \mathcal{I}}$, $e_o^i = \left(e_o^{ij}\right)_{j \neq i}$ 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 $P, Q, R > 0$.

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. In this infinitely repeated game, the moves by nature described above are independent across periods. The infinite repetition ensures that there is, in particular, always a future period in which one may be subject to social ostracism by others.

Consider, first, the case where past behaviour regarding act 'X' 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 'morality' of act 'X'. Even so, we know from the Folk Theorem that, if β is sufficiently close to 1, a variety of behaviour can be sustained in a subgame-perfect equilibrium. For example, we may have an equilibrium in which all individuals engage in act 'X' whenever they have the opportunity to do so, and no one faces social sanctions; and we may also have an equilibrium in which all individuals refrain from act 'X', and anyone who engages in act 'X' is sanctioned. Wahhaj (2012) illustrates and provides the formal conditions for these equilibria.

Introducing a 'moral code' or maxim in this framework narrows down the set of possible equilibria and provides a basis for predicting what norms can be sustained within the group for different types of beliefs. We represent the maxim in terms of feasible states of the world. We also allow for individuals to believe in a maxim which may, in fact, be false. Therefore, we need to distinguish between an individual's knowledge, which is always accurate, and his or her beliefs, which may be inaccurate. These concepts are formally defined the following section.

2.1. **A Framework for Modelling Interactive Knowledge and Beliefs .** We shall propose an alternative type of equilibrium for the game described above where players update their beliefs about the moral character of others based on their past actions. Specifically, we shall introduce a 'maxim' about moral character, a rule for mapping past actions of other players to beliefs about their moral character. Members of the community may have different priors regarding the truth of the maxim; and they may differ in terms of their higher order beliefs regarding the maxim. To investigate how such a maxim may affect behaviour within the community, we need an epistemic framework where it is possible for individuals to hold false beliefs. We introduce such a framework below.

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:

$$\Omega_t \subseteq \mathcal{H}_t \times \prod_{i \in \mathcal{I}} \Theta_i \times \Sigma \tag{2}$$

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.[1] For reasons we discuss later, not every element of the set represented on the right-hand side of (2) 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 $\prod_{j\neq i}\Theta_j \times \Sigma$:

$$\Gamma_i : \Theta_i \to \Delta\left(\prod\nolimits_{j\neq i}\Theta_j \times \Sigma\right)^2 \tag{3}$$

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.

2.2. **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*'). 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* as defined by Battigalli and Bonanno (1999). 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) $\mathcal{B}_t^i(\omega) \subseteq \mathcal{K}_t^i(\omega)$

[1]For the analysis, we construct Θ_i as a particular countable type-space according to a procedure described in Section 2.4.

(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.

2.3. **The Evolution of Beliefs.** Next, we specify how the belief and knowledge correspondences, defined in Section 2.2, 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

$$\mathcal{B}_0^i = \left\{ \omega \in \prod_{i \in \mathcal{I}} \Theta_i \times \Sigma : p_{\theta_i,0}^i(\omega) > 0 \right\} \tag{4}$$

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 $\hat{\sigma}_t(a_t|\omega_{t-1}, \sigma)$ that the action a_t will take place after state ω_{t-1} has been realised, given a strategy profile σ (to be defined in more detail in the next section). 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:

$$p_{\theta_i,t}^i(\omega_t|h_t) = \frac{p_{\theta_i,t-1}^i(\omega_{t-1}|h_{t-1})\,\hat{\sigma}_t(a_t|\omega_{t-1},\sigma)}{\sum_{\omega'_{t-1} \in \Omega_{t-1}} p_{\theta_i,t-1}^i\left(\omega'_{t-1}|h_{t-1}\right)\hat{\sigma}_t\left(a_t|\omega'_{t-1},\sigma\right)} \tag{5}$$

Thus, equation (5) 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 $p^i_{\theta_i,t-1}(.|h_{t-1})$ 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

$$\mathcal{B}^i_t(\omega_t) = \left\{\omega \in \Omega_t : p^i_{\theta_i,t}(\omega|h_t) > 0\right\} \tag{6}$$

Equation (5) 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 (5) is positive.

2.4. **Representing the History of Actions and a Maxim.** 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 $\mathcal{E} = \{e^{ij}_o : i, j \in \mathcal{I}, i \neq j\} \cup \{e^i_w : i \in \mathcal{I}\}$, and represent the relevant actions in a period as a tuple $(e, a) \in \mathcal{E} \times \{0, 1\}$. Thus, the tuple $(e^i_w, 0)$, for example, indicates that person i had an opportunity to engage in act 'X' 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 Θ_i, which we define in more detail below. Besides the player types, the time-invariant characteristics of the game will include the moral character of each player as defined above: $c_i \in \{0, 1\}$, $i \in \mathcal{I}$. 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 $\Sigma = \{0, 1\}^{n+1}$.

Next, we describe how a maxim or 'moral code' can be represented within this framework. Consider the following maxim: "A person who has engaged in act X has bad moral character." In each period t, only a subset of states in Ω_t will be consistent with this

maxim, which we can represent as follows:

$$\mathcal{M}_t^0 = \left\{(h_t, \boldsymbol{\theta}, \mathbf{c}, 1) \in \Omega_t : \text{for each } i \in \mathcal{I},\ (c_i = 0)\ \text{ or }\ \left(\left(e_w^i, 1\right) \subsetneq h_t\right)\right\} \tag{7}$$

where $\boldsymbol{\theta} = (\theta_1, .., \theta_n)$, $\mathbf{c} = (c_1, .., c_n)$. In words, $\mathcal{M}_t^0$ includes only those states of the world in period t where $\mu = 1$ – indicating that the maxim holds true – and every person i has either never engaged in act X or else he/she has bad moral character. We can construct more complex moral codes which relate not only to people's actions but also their beliefs, as follows:

$$\mathcal{M}_t^k = \left\{(h_t, \boldsymbol{\theta}, \mathbf{c}, 1) \in \Omega_t : \text{for each } i \in \mathcal{I},\ (c_i = 0)\ \text{ or }\ \left(\mathcal{B}_t^i(\omega_t) \subseteq \bigcap_{j=0}^{k-1} \mathcal{M}_t^j\right)\right\}\ \text{ for } k \in \mathbb{N}^+$$

As per the definition above, $\mathcal{M}_t^1$ captures those states of the world where $\mu = 1$ and every person i believes that the true state is in $\mathcal{M}_t^0$ or else he/she has bad moral character. Similarly, $\mathcal{M}_t^2$ captures those states of the world where every person i believes that the true state is in $\mathcal{M}_t^0 \cap \mathcal{M}_t^1$ or else he/she has bad moral character and so on. We wish to analyse the implications of a maxim which says that the statements corresponding to $\mathcal{M}_t^0$, $\mathcal{M}_t^1$, etc. all hold true. Therefore, we limit the state space in each period t to the following subset:

$$\Phi_t = \left(\bigcap_{j=0}^{\infty} \mathcal{M}_t^j\right) \cup \{(h_t, \boldsymbol{\theta}, \mathbf{c}, 0) \in \Omega_t\}$$

Thus Φ_t includes all states of the word in Ω_t in which either the maxim is false ($\mu = 0$) or the actions, beliefs and characters of individuals are consistent with the statements corresponding to $\mathcal{M}_t^0$, $\mathcal{M}_t^1$, etc. In words, we can describe the maxim collectively represented by the statements $\mathcal{M}_t^0$, $\mathcal{M}_t^1$, etc. as follows: "Anyone who commits the public act has bad moral character, and anyone who lacks belief in this statement also has bad moral character."

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 $\varepsilon > 0$, and realisations of moral character are believed to be independent across players. We define θ_b as a type of player that believes the maxim is true and believes that all other players are of type θ_b; and similarly, we define θ_n as a type of player that believes that the maxim is false and believes that all other players are of type θ_n.

Using θ_b and θ_n, we construct other types as follows: a player of type $\theta_b(X)$ believes that the maxim is true and assigns positive probability to each type in the set X, while a player of type $\theta_n(X)$ believes that the maxim is false and assigns positive probability to all types in the set X. Thus, type $\theta_n(\{\theta_b\})$ does not believe in the maxim and belives that all other players are of type θ_b; type $\theta_b(\{\theta_b, \theta_n\})$.believes in the maxim and believes that all other players are either of type θ_b or θ_n; type $\theta_n(\{\theta_b, \theta_n(\theta_b)\})$ does not believe in the maxim and believes all other players are either of type θ_b or type $\theta_n(\theta_b)$, etc. We denote by Θ the set of all possible types that can be constructed in this manner.

2.5. **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 chooses a specific action in period t, contingent 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 $\theta_i \in \Theta$, 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 (5). 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).

2.6. **Characterisation of Equilibria.** Next, we provide a characterisation of equilibria of the game. The formal reasoning is provided in the proof of Proposition 2.1. But the structure of the argument will be evident from the following informal description.

We begin by considering the possible strategies for an individual of type θ_b. 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 θ_b individual would ostracise one who has committed the public act, regardless of the strategies pursued by others.[3]

Furthermore, since a θ_b individual believes that everyone else in the community is of type θ_b, who, by definition, believe that the maxim is true, she would expect to be ostracised by everyone were she to engage in the public act. Therefore, if the benefit of engaging in the act is lower than the disutility from being subject to ostracism, as implied by the condition in (9), she would refrain from doing so.

$$W < \frac{\beta\,(n-1)\,\delta_o}{1-\beta} P \tag{9}$$

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 θ_b person, who believes that the maxim is true, will conclude that such a person also has bad moral character. Therefore, if R is sufficiently high, a θ_b person will ostracise an individual who has failed to ostracise someone who has committed the public act.[4] Reasoning in the same manner, we can show that a θ_b 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.

[3]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

$$R - Q > \frac{\beta\,(n-1)}{1-\beta}\left(\delta_o P + \delta_w |W|\right) \tag{8}$$

a θ_b individual should ostracise someone who has committed the public act regardless of the repercussions.

[4]The condition is the same as the one derived in the preceding footnote.

There remains only the question of whether, in an equilibrium, a θ_b 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 Q and no significant direct reward. Therefore, she would only do so if she is given additional incentives for it.

The only possible type of equilibrium where individuals with high probability of good moral character are ostracised is if such ostracism occurs with some frequency less than 1 and the behaviour is supported by the threat of increased frequency of ostracism.[5] Therefore, if R is sufficiently large, the strategies pursued by θ_b individuals in a equilibrium must take the following form.

$$\sigma_t^i\left(h_{t-1}, e_w^i, \theta_b\right) = 0 \tag{10}$$

$$\sigma_t^i\left(h_{t-1}, e_o^{ij}, \theta_b\right) = 1 \text{ if } j \in \mathcal{I}_b\left(h_{t-1}\right) \tag{11}$$

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

where $\mathcal{I}_b\left(h_t\right) = \left\{j \in \mathcal{I} : \left(e_w^j, 1\right) \in h_t \text{ or } \left(\left(h_\tau, e_o^{jl}, 0\right) \subseteq h_t \text{ and } l \in \mathcal{I}_b\left(h_\tau\right)\right)\right\}$ and $\pi \in [0, 1)$. In words, any strategy pursued by a θ_b individual in an equilibrium must include the following instructions: '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.' The strategy may further indicate that those who do not belong to any of these categories are to be ostracised with some frequency less than 1. Given this strategy, the following is the necessary and sufficient condition for θ_b individuals to not engage in the public act:

$$W < \frac{\beta\left(n-1\right)\delta_o}{1-\beta}\left(1-\pi\right)P \tag{13}$$

By definition, individuals of type $\theta_n\left(\{\theta_b\}\right)$ believe that all other community members are of type θ_b. For ease of notation, we use henceforth the abbreviation θ_{nb} for the type $\theta_n\left(\{\theta_b\}\right)$. In an equilibrium where θ_b individuals are playing a strategy which satisfies (10)-(12), a θ_{nb} individual expects to be ostracised by everyone else if she engages in the public act. Therefore, she would not do so if (13) holds. She also reasons that if she fails

[5]Note that the act of ostracism against individuals not believed to have bad moral character cannot be probabilistic; as it is not possible to enforce such behaviour using the threat of punishment. But the act of ostracism may be linked to other elements of the game such as the current time period.

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 θ_b 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 the following condition holds:

$$Q - \varepsilon R < \frac{\beta (n-1) \delta_o}{1-\beta} (1-\pi) P \tag{14}$$

for which the following is a sufficient condition for any $\pi \in [0,1)$:

$$Q - \varepsilon R < \frac{\beta (n-1) \delta_o}{1-\beta} P \tag{15}$$

Then, there remains only the question of whether θ_{nb} individuals will ostracise those who do not belong to any of these categories; i.e. those that θ_b individuals do not believe to have bad moral character. As we reasoned above, if they do so, it must be with some frequency less than 1. Therefore, if the θ_b individuals are playing the strategy specified in (10)-(12), then under conditions (13) and (14), the strategy pursued by θ_{nb} individuals must take the following form.

$$\sigma_t^i \left(h_{t-1}, e_w^i, \theta_{nb}\right) = 0 \tag{16}$$

$$\sigma_t^i \left(h_{t-1}, e_o^{ij}, \theta_{nb}\right) = 1 \text{ if } j \in \mathcal{I}_b \left(h_{t-1}\right) \tag{17}$$

$$\sigma_t^i \left(h_{t-1}, e_o^{ij}, \theta_{nb}\right) = \pi_1 \text{ if } j \notin \mathcal{I}_b \left(h_{t-1}\right) \tag{18}$$

where $\pi_1 \in [0,1)$. By definition, individuals of type $\theta_n \left(\{\theta_b, \theta_{nb}\}\right)$ believe that all other community members are of type θ_b or θ_{nb}. For ease of notation, we use the abbreviation θ_{nnb} to denote individuals of type $\theta_n \left(\{\theta_b, \theta_{nb}\}\right)$. Given the strategy for θ_{nb} individuals specified in (16)-(18) and the strategy for θ_b individuals specified in (10)-(12), we can show, using the same reasoning as above, that under conditions equivalent to (13) and (14) (i.e. with π_1 instead of π on the right-hand side of the inequalities), the strategy pursued by θ_{nnb} individuals in an equilibrium take the same form.

By reasoning iteratively, we can show that the strategies pursued by individuals of type $\theta_n \left(\{\theta_b, \theta_{nb}, \theta_{nnb}\}\right)$, $\theta_n \left(\{\theta_b, \theta_{nb}, \theta_{nnb}, \theta_{nnnb}\}\right)$, $\theta_n \left(\{\theta_b, \theta_{nb}, \theta_{nnb}, \theta_{nnnb}, \theta_{nnnnb}\}\right)$, etc. – (where $\theta_{nnnb} = \theta_n \left(\{\theta_b, \theta_{nb}, \theta_{nnb}\}\right)$, $\theta_{nnnnb} = \theta_n \left(\{\theta_b, \theta_{nb}, \theta_{nnb}, \theta_{nnnb}\}\right)$, and so on) – must take the same form too. In other words, the reasoning applies to any population in which higher order beliefs regarding the type of other individuals, above a finite nth order, is θ_b. Let $\Theta_b = \{\theta_b, \theta_{nb}, \theta_{nnb}, \theta_{nnnb}, ...\}$. We have now established the following.

Proposition 2.1. *If the conditions in (9), (15) and (8) hold, and all individuals in the population belong to a type in $\Theta_b \subset \Theta$, the equilibrium strategy pursued by each type includes refraining from the public act, and ostracising 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.*

Proposition 2.1 implies that, if beliefs regarding the types of other individuals in the population above any finite nth order includes only type θ_b, then the maxim about moral character can be used to pin down equilibrium behaviour regarding the public act. By contrast, it should be evident that the reasoning behind Proposition 2.1 does not apply to individuals of type θ_n. Consequently, it also does not apply if higher order beliefs regarding the type of other individuals includes θ_n. In other words, the maxim about moral character cannot be used to pin down the equilibrium behaviour of individuals whose first-order or higher-order beliefs regarding the type of other individuals include θ_n.

The reasoning behind Proposition 2.1 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.

2.7. **Properties of Equilibria in which the Social Taboo is sustained .** In this section, we discuss some important qualities of the type of equilibrium described in Proposition 2.1. The simplest type of equilibrium obtains if every member of the community is of type θ_b. Then they all believe in the association between the public act and the notion of 'bad moral character' embodied in the maxim 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 θ_{nb} 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 maxim to hide their true beliefs, because they fear being accused of bad moral character otherwise.

In a community consisting entirely of θ_{nnb} individuals, everyone believes, accurately, that their neighbours do not believe in the maxim. This can be seen from the fact that if individuals i and j are of type θ_{nnb}, then we have, by construction, $\mathcal{B}_0^i \subset E\left(\mathcal{B}_0^j \subset \mathcal{M}_0\right) \cup E\left(\mathcal{B}_0^j \subsetneq \mathcal{M}_0\right)$ (since i believes j to be of either type θ_b or type θ_{nb}; a θ_b individual believes in the maxim but a θ_{nb} individual does not) and $\mathcal{B}_0^j \subsetneq \mathcal{M}_0$ (since a θ_{nnb} individual does not believe in the maxim).[6] However, they have inaccurate beliefs about what their neighbours believe about whether others believe in the maxim (since, by construction, $\mathcal{B}_0^i \subset E\left(\mathcal{B}_0^j \subset E\left(\mathcal{B}_0^i \subset \mathcal{M}_0\right)\right)$ but $\mathcal{B}_0^j \subset E\left(\mathcal{B}_0^i \subsetneq \mathcal{M}_0\right)$). In other words, the second-order beliefs are inaccurate. And this causes everyone to behave in accordance with the maxim to hide their true beliefs, because they fear being accused of bad moral character otherwise.

In a community consisting entirely of type $\theta_{n...nb}$ individuals (where n may be repeated any finite number of times in the subscript) , everyone has accurate beliefs up to any finite 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 2.1 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 should affect a person's behaviour. Also, unless the reward R is common knowledge, the reasoning used in Proposition 2.1 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

[6]Here, $\mathcal{M}_0$ denotes, as per (7), the subset of states in which the maxim holds true at $t = 0$. Since no actions have yet taken place, we have $\mathcal{M}_0 = E\left(\mu = 1\right)$.

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 2.1 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 2.6, a θ_b player is playing his or her dominant strategy following each possible history. In other words, a type θ_b player would do worse with any other continuation strategy profile. 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.

3. The Dynamics of Social Taboos

In Proposition 2.1, we described a type of 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.

We explore in this section under what conditions beliefs and equilibrium behaviour in the game would evolve over time. If a player assigns a probability of 1 to the event that the maxim is true at the start of the game – as type θ_b players do in our construction in Section 2.4 – then Bayesian updating cannot lead to a change in beliefs in future periods. Therefore, we redefine ex-ante beliefs regarding the maxim as follows. Instead of assuming that θ_b and $\theta_b(X)$ players believe in the maxim, we assume that they believe that the maxim is false with some probability $\delta > 0$ (and true with probability $1-\delta$). We retain all other assumptions about prior beliefs described in Section 2.4.

Then, if an individual engages in the public act, or fails to ostracise someone who has engaged in the public act, etc. a θ_b individual will update her subjective probability that the individual has bad moral character from ε to $\frac{\varepsilon}{(1-\varepsilon)\delta+\varepsilon}$.[7] 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 2.6 will still go through.

Rather than having a static population as in Section 2, we introduce, for each individual i, a small exogenous probability ζ that he or she will exit the game (migrate or die), and another individual will take his or her place.

Let $\mathcal{I}(0) = \{1, 2, .., n\}$ be the set of individuals living in the community at the start of the game. Let $\mathcal{J}_n$ be the set of n-element subsets of $\mathbb{N}^+$. We denote by $\mathcal{I}(t) \in \mathcal{J}_n$ the set of community members in period t. The set of possible states in period t can be represented as

$$\Omega_t \subseteq \mathcal{H}_t \times \mathcal{J}_n \times (\Theta_b)^n \times \Sigma$$

[7] 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 θ_b, that the former has bad moral character is given by

$$p^j_{\theta_b,t+1}\left(E(c_i = 0) \mid \left(h_{t-1}, e^i_w, 1\right)\right)$$

$$\begin{aligned}
&= \frac{\left[p^j_{\theta_b,t}\left(E(c_i=0)\cap \mathcal{M}_{t+1}\right) + p^j_{\theta_b,t}\left(E(c_i=0)\cap (\mathcal{M}_{t+1})^C\right)\right]}{\left[p^j_{\theta_b,t}\left(E(c_i=0)\cap \mathcal{M}_{t+1}\right) + p^j_{\theta_b,t}\left(E(c_i=1)\cap (\mathcal{M}_{t+1})^C\right) + p^j_{\theta_b,t}\left(E(c_i=0)\cap (\mathcal{M}_{t+1})^C\right)\right]} \\
&= \frac{\varepsilon(1-\delta)+\varepsilon\delta}{\varepsilon(1-\delta)+(1-\varepsilon)\delta+\varepsilon\delta} \\
&= \frac{\varepsilon}{(1-\varepsilon)\delta+\varepsilon}
\end{aligned}$$

At the start of each period t, each player $i \in \mathcal{I}(t-1)$ may 'die' with exogenous probability ζ, and replaced by a new player with index $\max\{i : i \in \mathcal{I}(t-1)\} + 1$. Each new player (as well as those initially present in the community) will be assigned a type from Θ_b according to a probability distribution $F_\Theta \in \Delta(\Theta_b)$, and assigned 'bad moral character' with probability ε. The realisation of type and moral character will be independent across players.

Individuals form their beliefs regarding the type of any new member of the community according to their own types; i.e. a θ_b individual believes any new member is also of type θ_b, a θ_{nnb} individual assigns, for any new individual in the community, positive probabilities to the types θ_b and θ_{nb}, etc. We assume that each new player who enters the community will have knowledge of the previous history of the game (as stated in Assumption 1).

We retain the definitions of strategy and equilibrium provided in Section 2.5 to analyse behaviour in this dynamic community. Given the probability of exit, individuals will discount future payoffs by a factor $\beta\zeta$. As individuals believe new entrants to the community to have a type from the same set as those whom they replace, the reasoning provided in Section 2.6 regarding equilibrium strategies would still apply (even if, as argued above, θ_b individuals are assumed to have a small amount of 'doubt' about the truth of the maxim). Therefore, any equilibrium strategy profile will include the behaviour described in the statement of Proposition 2.1 if the conditions in (9), (15) and (8) hold for the discount factor $\beta\zeta$.

We have thus established that even if members of the community are replaced by new individuals with some probability in each period, and the set of possible types correspond to that defined in Section 2.4, the social taboo against the public act will be maintained in each period in any equilibrium.

This raises the question whether, given our set of possible types Θ_b, there exist some parameter values for which equilibrium behaviour regarding the public act changes after a finite number of periods. This can happen only if θ_b individuals change their beliefs regarding the truth of maxim; which, in turn, can happen only if a member of the community engages in the public act. Of course, if the parameter values are such that the conditions in (9), (15) or (8) are not satisfied, then such a result is readily obtained. However, a more interesting case of unravelling occurs when some member of the community has an exceptionally low probability of bad moral character.

Let us assume that, on rare occasions, a new entrant to the community is believed to have 'stronger reputation' of good moral character compared to others. Formally, we can imagine a device that, whenever a new entrant appears in the community, produces a public signal with probability ρ_g for individuals of good moral character, and with probability ρ_b for individuals of bad moral character, where $\rho_g > \rho_b$. For a new entrant for whom such a signal has been observed, the posterior probability of bad moral character will equal (using Bayes' rule) $\varepsilon_\rho = \frac{\rho_b \varepsilon}{\rho_g(1-\varepsilon)+\rho_b \varepsilon} < \varepsilon$.[8]

Therefore, if the signal is observed for a new entrant to the community l, all players assign a probability ε_ρ to the event that he or she has bad moral character. If individual l subsequently has the opportunity to engage in the public act (i.e. event e_w^l is realised in some period t) and chooses to do so ($\alpha_w^l = 1$), then this probability will be revised up to $\frac{\varepsilon_\rho}{(1-\varepsilon_\rho)\delta+\varepsilon_\rho}$; but note that if ε_ρ is small relative to δ – which would mean that θ_b individuals have more confidence in the good moral character of l than the truth of the maxim – then this expression, close to $\frac{\varepsilon_\rho}{\delta}$, is smaller than 1. Then the posterior probability that l has bad moral character may not be sufficiently high for θ_b individuals to ostracise him in the absence of other incentives.

Moreover, if individual l commits the public act, θ_b individuals will revise downward their subjective probability that the maxim is true to $\frac{(1-\delta)\varepsilon_\rho}{(1-\delta)\varepsilon_\rho+\delta}$.[9] If ε_ρ is small relative to δ, then the posterior probability will be close to 0. In this case, θ_b individuals will not have sufficient incentive to ostracise individuals who commit the public act in subsequent

[8]To compute this expression, we use the formula

$$\Pr(\text{bad}|\text{signal}) = \frac{\Pr(\text{signal}|\text{bad})\Pr(\text{bad})}{\Pr(\text{signal})}$$

[9]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 θ_b, that the maxim is true is given by

$$\begin{aligned}
& p_{0,t+1}^j\left(\mathcal{M}_{t+1}|\left(h_{t-1}, e_w^l, 1\right)\right) \\
= \;& p_{0,t+1}^j\left(E\left(c_i=0\right)\cap\mathcal{M}_{t+1}|\left(h_{t-1}, e_w^l, 1\right)\right) \\
= \;& \frac{\left[p_{0,t}^j\left(E\left(c_i=0\right)\cap\mathcal{M}_{t+1}\right)\right]}{\left[p_{0,t}^j\left(E\left(c_i=0\right)\cap\mathcal{M}_{t+1}\right)+p_{0,t}^j\left(E\left(c_i=1\right)\cap\left(\mathcal{M}_{t+1}\right)^C\right)+p_{0,t}^j\left(E\left(c_i=0\right)\cap\left(\mathcal{M}_{t+1}\right)^C\right)\right]} \\
= \;& \frac{\varepsilon_r\left(1-\delta\right)}{\varepsilon_r\left(1-\delta\right)+\left(1-\varepsilon_r\right)\delta+\varepsilon_r\delta} \\
= \;& \frac{\varepsilon_r}{\left(1-\varepsilon_r\right)\delta+\varepsilon_r}
\end{aligned}$$

periods; or to ostracise individuals who fail to ostracise anyone who has committed the public act, etc. As a result, higher types will also not have sufficient incentive to adopt the behaviour described in the statement of Proposition 2.1.

Therefore, we can construct an equilibrium where, in the continuation game following the public act by individual l, all individuals engage in the public act whenever they have the opportunity to do so, and no one is ostracised for engaging in the public act, or failing to ostracise someone who has engaged in the public act, etc.; while behaviour in the community prior to the public act by individual l corresponds to the strategies described in the statement of Proposition 2.1. Thus, the social taboo 'unravels' within the game.

Note that it is possible to sustain the social taboo even after the public act by individual l using the strategies described in the first part of Section 2. However, there is no a priori reason why they would adopt such behaviour over any other.

4. 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 maxim: 'Anyone who engages in an act X has bad character. Anyone who associates with someone with bad character also has bad 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 then in any equilibrium of the game everyone behaves as if the maxim were true.

To show this result formally, we make use of an infinitely repeated game where players are randomly given the opportunity to engage in a publicly observable act, and punish one another. Standard folk theorem reasoning would show that, if the players are sufficiently patient, a wide range of behaviour can be sustained in this setting. Yet, we show that a maxim, such as the one described above, can 'pin down' behaviour – specifically whether players engage in the public act or not – even if no one believes in the maxim, everyone is aware of this, and is aware that everyone is aware, etc. up to a finite nth order.

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 assess. But our result implies that, even if belief in the maxim is extremely 'weak' – in the sense that people may have only higher order beliefs regarding its veracity – they will continue to respect the moral injunction.

The theoretical mechanism suggests a particular formula 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.

5. Appendix

Proof. of Proposition 2.1: (Individuals of type θ_b): First, we show that, under conditions (15) and (8), it is optimal for a person of type θ_b to ostracise those who have engaged in the public act, regardless of the strategies pursued by other players.

Let $h_t = (h_{t-1}, e_o^{ij})$ be the history of actions in the game at the beginning of period t and suppose that $(e_w^j, 1) \in h_{t-1}$; i.e. person i has the opportunity to ostracise person j in period t and person j has previously engaged in the public act. Denote by ω_t the state of the world in period t. For ease of notation, let

$$\mathcal{M}_t = \left(\bigcap_{j=0}^{\infty} \mathcal{M}_t^j \right)$$

If person i is of type θ_b, then we have

$$\begin{aligned} \mathcal{B}_t^i(\omega_t) &= \mathcal{M}_t \cap \mathcal{K}_t^i(\omega_t) \\ &= \mathcal{M}_t \cap E(h_t), \text{ by Assumption 1} \quad (19) \\ &\subset E(c_j = 0) \quad (20) \end{aligned}$$

This last relation follows from the definition of $\mathcal{M}_t^0$ and that $(e_w^j, 1) \in h_t$. Therefore, if person i chooses to ostracise person j in period t, he receives an additional utility of $E_i\left(1 - c_j\right) R = R$. Under condition (8), this additional utility exceeds the cost of any punishment that the other players can inflict on i in subsequent periods. Therefore, it is optimal for person i to ostracise person j in period t. Therefore, $\sigma_t^i\left(h_{t-1}, e_o^{ij}, \theta_b\right) = 1$.

Next, we show that it is optimal for a person of type θ_b to ostracise someone who has failed to ostracise someone who has engaged in the public act, regardless of the strategies pursued by other players.

Suppose, again, that $h_t = \left(h_{t-1}, e_o^{ij}\right)$. Suppose that $h_s = \left(h_{s-1}, e_o^{jl}, 0\right)$ and $h_r = \left(h_{r-1}, e_w^l, 1\right)$ and $r < s < t$, $h_r \subset h_s \subset h_t$; i.e. person l engaged in the public act in period r and person j failed to ostracise person l in a subsequent period s (when l had the opportunity to do so) and person i has the opportunity to ostracise person j in a subsequent period t.

Using the reasoning above, $\sigma_s^j\left(h_{s-1}, e_o^{jl}, \theta_b\right) = 1$; i.e. if person j were of type θ_b, then he would ostracise person l in period s. Therefore, if person i is of type θ_b, then we have

$$
\begin{aligned}
\mathcal{B}_t^i\left(\omega_\tau\right) &= \mathcal{M}_t \cap E\left(h_t\right) \\
&\subset \mathcal{M}_t \cap E\left(h_{s-1}, e_o^{jl}, 0\right) && (21) \\
&\subset \mathcal{M}_t \cap E\left(\theta_j \neq \theta_b\right) && (22) \\
&\subset \mathcal{M}_t \cap \left(\mathcal{B}_s^j\left(\omega_s\right) \subsetneq \mathcal{M}_s^0\right) && (23) \\
&\subset E\left(c_j = 0\right) && (24)
\end{aligned}
$$

where ω_s is the state of the world in period s after history h_s has been realised. The steps in (21)-(23) follow from the fact that if j chose not to ostracise l in period s, despite l having engaged in the public act in a preceding period r, it must be that l's beliefs do not include $\mathcal{M}_s^0$. The step in (24) follows from the definition of $\mathcal{M}_t^1$.Therefore, under condition (8), it is optimal for person i to ostracise person j in period t. Therefore, $\sigma_t^i\left(h_{t-1}, e_o^{ij}, \theta_b\right) = 1$.

Using the same type of reasoning, we can show that it is optimal for a person of type θ_b to ostracise someone who has failed to ostracise someone who has failed to ostracise someone who has failed to ostracise someone who has engaged in the public act, regardless of the strategies pursued by other players; and that the same applies to longer 'chains' of non-ostracising behaviour against someone who has engaged in the public act.

Next, we show that under condition (9), a person of type θ_b will refrain from engaging in the public act. Recall that, by definition, if person j is of type θ_b, then he assigns, ex-ante, a probability of 1 to all other members of the community being of type θ_b. Suppose $(e_w^j, 1) \in h_s$. We have already shown that $\sigma_t^i(h_{t-1}, e_o^{ij}, \theta_b) = 1$ if $(e_w^j, 1) \in h_{t-1}$. Therefore, if $h_s \subset h_{t-1}$, $s \leq t-1$, then person j assigns a probability of 1 to being ostracised if the event e_o^{ij} occurs in any period subsequent to period s. Under condition (9), the expected cost of this punishment exceeds any benefit from engaging in the public act in period s. Therefore, $\sigma_s^j(h_{s-1}, e_w^j, \theta_b) = 0$.

(Individuals of type $\theta_n(\{\theta_b\})$): Next, we show that, under condition (9), given the characterisation of the strategy pursued by individuals of type θ_b, it is optimal for a person of type $\theta_n(\{\theta_b\})$ to ostracise those who have previously engaged in the public act. By definition, if person j is of type $\theta_n(\{\theta_b\})$, then he assigns, ex-ante, a probability of 1 to all other members of the community being of type θ_b. Suppose $\left(e_w^l, 1\right) \in h_r$ and $\left(e_o^{jl}, 0\right) \in h_s$ where $r < s$ and $h_r \subset h_s$. We have already shown that, if $h_s \subset h_{t-1}$, then $\sigma_t^i(h_{t-1}, e_o^{ij}, \theta_b) = 1$. Therefore, person j assigns a probability of 1 to being ostracised if the event e_o^{ij} occurs in any period subsequent to period s. Under condition (15), the expected cost of this punishment exceeds the cost of ostracising person l in period s. Therefore, $\sigma_s^j\left(h_{s-1}, e_o^{jl}, \theta_n(\{\theta_b\})\right) = 1$; i.e. it is optimal for person j to ostracise any person l who has previously engaged in the public act. Using the same type of reasoning, we can show that it is optimal for person j to ostracise any person who has failed to ostracise someone who has engaged in the public act; and that it is also optimal for person j to engage in ostracism for longer 'chains' of non-ostracising behaviour against someone who has engaged in the public act. Following the reasoning applied to individuals of type θ_b, we can also show that, under condition (9), a person of type $\theta_n(\{\theta_b\})$ will refrain from engaging in the public act.

Using the same type of reasoning as above, we can show that, under conditions (9) and (15), it is optimal for individuals of 'higher types' in Θ_b not to engage in the public act, and ostracise those who have previously engaged in the public act, and ostracise those who have previously failed to ostracise someone has engaged in the public act, etc. □

References

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

[2] Aumann, Robert (1999). "Interactive Epistemology I: Knowledge", *International Journal of Game Theory*, 1999, Volume 28.
[3] Banerjee, A (1992). "A Simple Model of Herd Behavior", *The Quarterly Journal of Economics*, Vol. 107(3), pp. 797-817.
[4] Battigalli, P. and G. Bonanno (1999). "Recent Results on Belief, Knowledge and the Epistemic Foundations of Game Theory", *Research in Economics*, Volume 53.
[5] Bernheim, B. (1994) "A Theory of Confirmity", *Journal of Political Economy*, Vol. 102(4), pp. 841-877.
[6] Bicchieri, Cristina (2011). "Social Norms", *The Standard Encyclopedia of Philosophy*, 2011.
[7] Brock, W. and S. Durlauf (2001). "Discrete Choice with Social Interactions", *Review of Economic Studies*, Vol. 68(2), pp. 235-260.
[8] Camerer, C.F., and R.H. Thaler (1995). "Ultimatums, dictators and manners", *Journal of Economic Perspectives*, Vol. 9, pp. 209-219.
[9] Chamley, C. (1999). "Coordinating Regime Switches", *Quarterly Journal of Economics*, Vol. 114(3), pp. 869-905.
[10] Chen, Yi-Chun (2011). "A Structure Theorem for Rationalizability in the Normal Form of Dynamic Games", mimeo, National University of Singapore, 2011.
[11] Cooper, R. and A. John (1988). "Coordinating Coordination Failures in Keynesian Models", *Quarterly Journal of Economics*, Vol. 103(3), pp. 441-463.
[12] Elster, Jon (1989). "Social Norms and Economic Theory", *The Journal of Economic Perspectives*, Volume 3, 1989.
[13] Farrell, J., and E. Maskin (1989). "Renegotiation in repeated games", *Games and Economic Behavior*, 1989, Volume 1.
[14] Fudenberg, Drew and Jean Tirole (1991). *Game Theory*, MIT Press, Cambridge, Massachusetts, 1991.
[15] Greif, Avner (1993). "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition", *The American Economic Review*, Volume 83, 1993.
[16] Gintis, Herbert (2009). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*, Princeton University Press, 2009.
[17] Kuran, Timur (1987). "Preference Falsification, Policy Continuity and Collective Conservatism", *Economic Journal*, Vol. 47, pp. 642-665.
[18] Kuran, Timur (1995). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*, Harvard University Press, 1995.
[19] Parsons, Talcott (1951). *The Social System*. Routeledge, New York, 1951.
[20] Rubinstein, Ariel (1989). "The Electronic Mail Game: Strategic Behavior Under 'Almost Common Knowledge'", *The American Economic Review*, Volume 79, 1989.
[21] Wahhaj, Zaki (2012). "Social Norms, Higher Order Beliefs and the Emperor's New Clothes", *ThReD Working Paper 2012-023.*
[22] Weinstein, Jonathan and Muhamet Yildiz (2007). "A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements", *Econometrica*, Volume 75(2), March 2007.
[23] Weinstein, Jonathan and Muhamet Yildiz (2013). "Robust Predictions in Infinite-Horizon Games – An Unrefinable Folk Theorem", *Review of Economic Studies*, Volume 80(1), pp. 365-394.

University of Kent