Moral Codes, Higher-Order Beliefs and The Persistence of Social Norms[*]

Zaki Wahhaj
University of Kent[†]

July 2017

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 present an alternative theory that draws elements from both of these approaches to offer insights about the nature of social norms and their persistence over time. 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., collective participation in social sanctions must occur in equilibrium if (certain Folk-theorem-type conditions hold and) there are higher order beliefs, at some finite order n and above.

JEL Codes: D01, D02, D83, Z10

Keywords: Informal Institutions; Common Belief; Social Sanctions; Social Taboos

[*] I am grateful for support from the University of Oxford and the University of Kent, where the research for this paper was carried out.

[†] 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 insights about the behaviour of individuals who are not 'hard-wired'.

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 theory that draws elements from both these approaches to offer insights about the nature of social norms and their persistence over time. Our underlying assumption is that people derive utility from associating with those who they believe to be 'moral' and ostracising those who they believe to be 'amoral'. We define a 'moral code' as a classification of possible actions into 'moral' choices and 'immoral' choices, coupled with the statement that 'it is immoral to associate with a person of amoral character'. When individuals choose an action, they are subject to an exogenous tremble and, if the moral code is 'true', 'moral' people are more likely to tremble when they make an 'immoral' choice and less likely to tremble when they make a 'moral' choice, compared to 'amoral' people. Others observe the action taken, which may or may not include trembling, but not the intended choice.

Whenever a person i is observed to make an 'immoral' choice, those who believe the moral code is true will update beliefs to assign i a higher probability of 'amoral character'; and, if beliefs are sufficiently strong, ostracise i thereafter whenever they have the opportunity to do so. It is evident that if individuals incur sufficiently high disutility from experiencing ostracism and there is common belief in the moral code, then they will all refrain from choosing 'immoral' actions. In this paper we show the surprising result that we obtain the same result even in a population where the moral code is believed to be false, and people believe that others believe the moral code is false, and believe that others believe that others believe that the moral code is false, etc. but the moral code is held to be true above some finite nth order.

The intuition behind this result is as follows: a person j, who does not believe in the moral code himself but believes that others do, will behave as if the moral code is indeed true (i.e. avoid choosing immoral actions, 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 moral code nor believes that others do, but believes that others have the same beliefs as j will also behave as if the moral code is indeed true to avoid being ostracised by people like j. We can continue with this reasoning *ad infinitum*. Thus people refrain from the immoral choice even when no one believes in the moral code, and no one believes that others do, and no one believes that others believe that others do, etc. so long as the moral code is held to be true above some finite nth order.

The implication of this result is that a social norm can be sustained by a moral code – as defined above – even when people do not believe in the moral code but the absence of belief is not common knowledge. We show how this can translate into a persistence of social norms in a dynamic population where newcomers are fully informed about the beliefs of existing players but not vice versa. 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.

The theoretical framework we analyse has some parallels with the literature on global games: games of incomplete information where players receive private signals about the underlying state (see Morris and Shin 2003 and Morris 2007 for reviews). Similar to global games, we investigate a setting with incomplete information regarding some payoff-relevant characteristics – in our case, the 'moral character' of other players. The private signals in global games leads to heterogeneity in first-order and higher-order beliefs about the underlying state. Simiarly, we assume heterogeneous prior beliefs (including higher-order beliefs) regarding the underlying state, more precisely, about whether a 'moral code' is a true representation of the world.

Our result regarding higher order beliefs in a moral code also has a parallel in the global games literature: where it has been shown that a small amount of incomplete information can provide a unique prediction about rationalisable behaviour even when the equivalent complete information game has multiple equilibria (Carlsson and van Damme 1993).[1] However, these results do not lend themselves easily to the type of game we analyse. For example, in our setting, small perturbations in beliefs about payoffs do not lead to a uniquely rationalisable outcome. Rather, our proposed mechanism relies on the idea that players' actions can affect their subsequent 'reputation' – which in turn can affect future expected payoffs – and people have heterogeneous first-order and higher-order beliefs regarding the appropriate rule for updating beliefs. Belief-updating based on past actions is key to our insights regarding the persistence of social norms.

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.

[1] These results have subsequently been generalised in the form of 'structure theorems on rationalisability' for normal-form games (Weinstein and Yildiz 2007), finite extensive-form games (Chen 2012) and infinite-horizon dynamic games (Weinstein and Yildiz 2013).

The remainder of this paper is organised as follows. In the next section, we present a simple example to illustrate how we represent a 'moral code' and how it can shape behaviour in a game. The formal model is presented in Section 3. In sections 3.1-3.4, we develop a framework for analysing the role of higher-order beliefs in maintaining social norms. Section 3.5 demonstrates the main result of the paper, and additional properties of the equilibrium are discussed in Section 3.6. We discuss the implications of the model – in particular the persistence of social norms despite changing beliefs – in Section 4.

2 Behaviour under a Moral Code

We begin by presenting a theoretical example to illustrate what we mean by a 'moral code' and how it can shape behaviour in a game. Consider two individuals i and j who interact over two periods. In period 1, person i has the opportunity to engage in an act 'X' or refrain from it. Person j observes i's action, and, in period 2, j has the opportunity to 'associate' with person i or 'ostracise' person i.

Person i receives a utility of w if he engages in act 'X' and receives a utility of 0 otherwise. He receives an additional utility of r if person j chooses to 'associate' with him in period 2. We assume that $r > w$. Person j receives a utility of $E\left[r\left(2c-1\right)|a\right]$ in period 2 if she chooses to 'associate' with person i and a utility of 0 otherwise. In the expression above, a takes a value of 1 if i has engaged in the act and 0 otherwise. The binary variable c indicates person i's 'moral character'; c takes a value of 1 if i is 'moral' and 0 if i is 'amoral'.

When choosing an action, player i 'trembles' with a small probability. More precisely, if i chooses action 'not X', the observed action (i.e. the action that j observes) is 'X' with probability $\left(1-mc\right)\varepsilon$ and, when i chooses action 'X', the observed action is 'not X' with probability $\left(1-mc\right)\varepsilon + mc$ where ε is positive and close to zero and $m \in \{0,1\}$. Thus, if $m = 0$, then player i trembles with probability ε regardless of his moral character. If $m = 1$, a moral player does not tremble when choosing action 'not X' and always trembles when choosing action 'X'. An amoral player still trembles with probability ε for either choice of action. Our interpretation of this formulation is that when $m = 1$, there is a 'moral choice' and an 'immoral choice' and a moral person is less likely to tremble when making a moral choice, and more likely to tremble when making an 'immoral choice'.[2] If $m = 0$, then all actions are identical in terms of their moral value and, therefore, there is no difference in the behaviour of moral and amoral individuals. Thus, $m = 1$ stands for a 'moral code' which says (roughly) that it is immoral to engage in act 'X'. Player j does not observe m and c

[2] Here, we make stark assumptions about trembling by a moral player to simplify the analysis but the same logic would hold as long as a moral player is more (less) likely to tremble than an amoral player when taking a moral (an immoral) action.

but has prior beliefs about the state of the world as follows:

$$\begin{aligned}\Pr(c=1) &= \gamma \\ \Pr(m=1) &= \mu\end{aligned}$$

and the realisations of c and m are independent.

We are now in a situation to analyse the behaviour of i and j in this game. Suppose player j observes i engage in act 'X' in period 1. If the equilibium strategy is for player i to play 'not X' then according to Bayes' rule, player j updates beliefs as follows:

$$\Pr(c=1|\text{'not X'}) = \mu\left\{\frac{\gamma}{\gamma+(1-\gamma)(1-\varepsilon)}\right\}+(1-\mu)\gamma > \gamma \tag{1}$$

$$\Pr(c=1|\text{'X'}) = \mu\left\{\frac{0}{\gamma.0+(1-\gamma)\varepsilon}\right\}+(1-\mu)\gamma < \gamma \tag{2}$$

On the other hand, if the equilibrum strategy is for player i to play 'X', then j updates beliefs as follows:

$$\Pr(c=1|\text{'not X'}) = \mu\left\{\frac{\gamma}{\gamma+(1-\gamma)\varepsilon}\right\}+(1-\mu)\gamma > \gamma \tag{3}$$

$$\Pr(c=1|\text{'X'}) = \mu\left\{\frac{0}{(1-\gamma)(1-\varepsilon)}\right\}+(1-\mu)\gamma < \gamma \tag{4}$$

Therefore, whatever the equilibrium strategy may be, the probability that i is moral increases when i plays 'not X' and decreases when i plays 'X'. Recall that player j receives a utility of $E\left[r(2c-1)|a\right]$ if she chooses to 'associate' with person i and a utility of 0 otherwise. Consider the following conditions:

$$E(c=1|\text{'not X'}) > \frac{1}{2}\ \&\ E(c=1|\text{'X'}) < \frac{1}{2} \tag{5}$$

We can verify, using (1)-(4) that the conditions in (5) are satisfied for both possible pure equilibrium strategies for i if, for example, $\gamma = \frac{1}{2}$ and $\mu > 0$. Then player j will associate with i if and only if i is observed to take action 'not X'. Therefore, since we have assumed that $r > w$, person i will choose 'not X'. Thus, there is a unique pure strategy equilibrium in which player i chooses 'not X' and player j associates with i if and only if i is observed to take action 'not X'.

The game illustrates what we mean by a 'moral code' and how belief in a 'moral code' can shape equilibrium behaviour in a game. In the following sections, we develop a more general version of this game to explore situations where there may be no belief in the moral code (i.e. $\mu = 0$) but higher order beliefs that the moral code is an accurate description of the world.

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 in the population. We define a stage game $\mathcal{G}$ in which two types of random events may occur:

(i) Let e_w^i be the event that person i is in a position to 'engage in act X'. If event e_w^i occurs, then person i has a choice of action a_w^i which can take a value of 0 or 1; $a_w^i = 1$ represents the action 'X', and $a_w^i = 0$ represents the action 'not X'.

(ii) Let e_o^{ij} be the event that person i has an opportunity to 'associate with' person j. If event e_o^{ij} occurs, then person i has a choice of action a_o^{ij} which can take a value of 0 or 1, where $a_o^{ij} = 1$ represents the action that person i 'associate with j', and $a_o^{ij} = 0$ represents the action 'ostracise j'.

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 $n\delta_w + n(n-1)\delta_o \leq 1$.

An individual who engages in act X receives a private payoff of W in that period. If individual i chooses to associate with j, then i receives a utility of $E[R(2c_j - 1)]$ and j receives a utility equal to R, where c_i and c_j are binary variables representing their moral character (As in the previous section, c_i takes a value of 1 if i is 'moral' and 0 if i is 'amoral').[3] We assume that c_i is unobservable for any individual $j \in \mathcal{I}$, including person i.[4] Each c_i, $i \in \mathcal{I}$, is independant with $\Pr(c_i = 1) = \gamma$ and we assume $\gamma \in \left(\frac{1}{2}, 1\right)$. Thus, the payoffs in the stage-game can be written as

$$u^i(a_i, a_{-i}, e) = \mathbf{I}\left(e_w^i\right) a_w^i W + \sum_{j \neq i} \left\{\mathbf{I}\left(e_o^{ij}\right) a_o^{ij} E(2c_j - 1) + \mathbf{I}\left(e_o^{ji}\right) \alpha_o^{ji}\right\} R \tag{6}$$

where $a_i = (a_o^i, a_w^i)$, $a_o^i = (a_o^i)_{j \neq i}$, $e = (e_o^i, e_w^i)_{i \in \mathcal{I}}$, $e_o^i = (e_o^{ij})_{j \neq i}$ 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.

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.

[3] The asymmetry between the payoffs received by the person who makes the decision whether to associate or ostracise and the person who is subject to this decision simplifies the Folk-theorem conditions but is not essential to the analysis. It can be justified by the argument that an individual would wish not to associate with another who he believed to be 'amoral' but would not wish to ostracised by the same person if the other had the opportunity to do so.

[4] The assumption that an individual does not know her own character has some parallels in the literature. See, for example, Benabou and Tirole (2011). The assumption raises the question whether an individual may have an incentive to learn about her own character through experimentation. We address this point in Section 3.6.

Consider, first, the case where individuals 'tremble' – in the sense defined in the preceding section – with probability ε whenever they choose an action, and the probability of trembling is unaffected by their moral character. Then there is no possible learning about moral chracter within the game. We can show that if β is sufficiently close to 1, a variety of behaviour can be sustained in a perfect public Equilibrium. For example, consider the following strategy for each player i: When event e_w^i occurs, choose $a_w^i = 0$; when event e_o^{ij} occurs, choose $a_o^{ij} = 0$ if and only if j has never previously been observed to play 'X', or associate with someone who has been observed to play 'X', or associate with someone who has associated with someone who has been observed to play 'X', etc. Such a strategy profile constitutes a perfect public equilibrium if and only if the following conditions hold true:

$$W < \frac{\beta_\varepsilon (n-1) \delta_o}{1-\beta_\varepsilon} \{R(1-2\varepsilon)\} \tag{7}$$

$$R(2\gamma - 1) < \frac{\beta_\varepsilon (n-1) \delta_o}{1-\beta_\varepsilon} \{R(1-2\varepsilon)\} \tag{8}$$

where $\beta_\varepsilon = \beta [1 - \{\delta_w + (n-1)\delta_o\} \varepsilon]$. The left-hand sides of (7) and (8) represent, respectively, for some person i, the utility gain from 'X' and the utility gain from associating with another individual. The right-hand side of (7) and (8) represent the expected present discounted value of the utility cost of social ostracism in subsequent periods. The term $(n-1)\delta_o$ is the probability that i experiences an event where another can choose to associate with or ostracise her. The term $R(2\gamma - 1)$ represents the expected utility to i of associating with a person who has a probability γ of being 'moral'. The term $(1 - 2\varepsilon)$ represents the increase in the probability of ostracism when others are choosing to ostracise i rather than associate with i, taking into account that, in both instances, there is an ε probability of a tremble. We discount by β_ε rather than β because in each period i has a probability $\{\delta_w + (n-1)\delta_o\}\varepsilon$ of trembling in which case she will be subject to social ostracism in subsequent periods in any case.

The conditions in (7) and (8) ensure that the disutility of social ostracism is higher than the utility gain from action 'X'; and also higher than the utility gain from associating with another individual. Then we obtain an equilibrium where there is a taboo against act 'X', sustained by the threat of social ostracism, and ostracism against those who violate the taboo is itself sustained by the threat of social ostracism.

But under the same conditions (7) and (8), we can also have an equilibrium where all players opt for the following strategy: When event e_w^i occurs, choose $a_w^i = 1$; when event e_o^{ij} occurs, choose $a_o^{ij} = 1$ if and only if j has never previously been observed to play 'not X', or associate with someone who has been observed to play 'not X', or associate with someone who has associated with someone who has been observed to play 'not X', etc.

Introducing a 'moral code' into this framework – as in the example provided in the preceding section – narrows down the set of possible equilibria and allows more precise predictions about behaviour. We represent a 'moral code' in terms of mappings from information sets to the set of feasible actions, indicating which actions are deemed 'moral' and 'immoral' (according to a particular code) at each information set. We also allow for individuals to have inaccurate first-order and higher-order beliefs about a moral code. 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 in the following section.

3.1 A Framework for Modelling Interactive 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:

$$\Omega_t = \mathcal{H}_t \times \prod_{i \in \mathcal{I}} \Theta_i \times \Sigma \tag{9}$$

where $\mathcal{H}_t$ is the set of all possible histories in period t; 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.

The history that is relevant to the game is the move by Nature (which determines which random event will occur) and the observed outcome when that event occurs (Recall that the observed outcome may differ from the player's action and the latter is not observed publicly). 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 outcome in a period as a tuple $(e, \tilde{a}) \in \mathcal{E} \times \{0, 1\}$. Thus, the tuple $(e_w^i, 0)$, for example, indicates that person i had an opportunity to engage in act 'X' but is observed not to commit the act. The relevant history from the beginning of the game up to period t can be written as $h_t = (e_1, \tilde{a}_1, e_2, \tilde{a}_2, ..., e_t, \tilde{a}_t)$ where e_τ denotes the move by Nature, and $\tilde{a}_\tau$ the observed outcome in period τ. So, the set of possible histories in period t is given by

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

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, we introduce a variable m which takes a value of 0 if the moral code is false and 1 if the moral code is true (As discussed in the next section, m will affect the probability of 'trembles' in each state of the world). So we can represent the set of time-invariant payoff-relevant characteristics by $\Sigma = \{0, 1\}^{n+1}$. We discuss in the next section how we use the state space to specify the details of the moral code.

3.2 Representing Moral Codes and Beliefs

We can represent a 'moral code' by a mapping, in each period t, from the set of information sets $\mathcal{H}_{t-1} \times \mathcal{E}$ to the set of feasible actions that are deemed 'moral' or 'immoral'. As the choice of actions at any information set is binary, the moral code can simply be represented by the set of information sets at which engaging in an act is moral – and its opposite is immoral – or vice versa. Consider the following statement: "When one has the choice of engaging in act X, the moral choice is always to avoid doing so". First, we identity, in each period t, the set of information sets where this statement is pertinent:

$$\begin{aligned} \mathcal{W}_1 &= \left\{ e \in \mathcal{E} : e = e_w^i \text{ for some } i \in \mathcal{I} \right\} \\ \mathcal{W}_t &= \left\{ (h_{t-1}, e) \in \mathcal{H}_{t-1} \times \mathcal{E} : e = e_w^i \text{ for some } i \in \mathcal{I} \right\}, \text{ for } t = 2, 3, ... \end{aligned}$$

To construct our moral code, we combine the statement above with the following: "When one has the choice of associating with another person, the moral choice is to do so if and only if that person has never previously made an immoral choice". This statement is pertinent at the following information sets: $\mathcal{O}_1 = \emptyset$ and

$$\mathcal{O}_t = \left\{ \begin{array}{c} (h_{t-1}, e) \in \mathcal{H}_{t-1} \times \mathcal{E} : e = e_o^{ij},\ (h_{s-1}, e') \subset h_{t-1}, \\ (h_{s-1}, e') \in \mathcal{W}_s \cup \mathcal{O}_s,\ e' \in \left\{ e_w^j, e_o^{jk} \right\},\ \tilde{a}_s = 1,\ i, j, k \in \mathcal{I} \end{array} \right\}, \text{ for } t = 2, 3, ...$$

In words, $\mathcal{O}_t$ is the period t set of information sets at which an individual i has the opportunity to associate with or ostracise j, and j has previously engaged in act 'X', or associated with some individual k at a similar information set in a previous period. Let $\mathcal{M}_t = \mathcal{W}_t \cup \mathcal{O}_t$ and let $\mathcal{M}_t^C$ be the complement of $\mathcal{M}_t$, for $t = 1, 2, 3, ...$ According to the moral code, at each information set in $\mathcal{M}_t$, the moral action is for the player to refrain from the act in question; and at each information set in $\mathcal{M}_t^C$ the moral action is for the player to engage in the act in question. The moral code affects trembling within the game as follows. Given action a_t at information set (h_{t-1}, e), the observed outcome $\tilde{a}_t$ is given by

$$\begin{aligned} \Pr(\tilde{a}_t \neq a_t) &= (1 - mc)\varepsilon + mc \text{ if } (h_{t-1}, e) \in \mathcal{M}_t \ \&\ a_t = 1 \qquad (10) \\ \Pr(\tilde{a}_t \neq a_t) &= (1 - mc)\varepsilon \text{ if } (h_{t-1}, e) \in \mathcal{M}_t \ \&\ a_t = 0 \qquad (11) \\ \Pr(\tilde{a}_t \neq a_t) &= (1 - mc)\varepsilon + mc \text{ if } (h_{t-1}, e) \in \mathcal{M}_t^C \ \&\ a_t = 0 \qquad (12) \\ \Pr(\tilde{a}_t \neq a_t) &= (1 - mc)\varepsilon \text{ if } (h_{t-1}, e) \in \mathcal{M}_t^C \ \&\ a_t = 1 \qquad (13) \end{aligned}$$

The probabilities above are defined such that after choosing an action, players always 'tremble' with probability ε – except that, if $m = 1$, a 'moral' player never trembles when engaging in a moral action and always trembles when engaging in an immoral action.[5] If

[5] As in the example in Section 2, we make stark assumptions about trembling by a moral player to simplify the analysis but the same logic would hold as long as a moral player is more (less) likely to tremble than an amoral player when taking a moral (an immoral) action.

$m = 0$ this effectively means the moral code is false. If the moral code is false, then all players are equally likely to tremble when undertaking any action.

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

$$\Gamma_i : \Theta_i \rightarrow \Delta\left(\prod_{j\neq i}\Theta_j \times \Sigma\right) \tag{14}$$

where $\Delta(S)$ is the set of all probability functions defined on the set S. 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. As Σ includes m, these probability functions also enable us to specify their prior beliefs regarding the moral code, i.e. before any actions have occurred in the game.

We define the player types as follows. First, θ_b is a type of player with the following prior beliefs: (i) the moral code is true; (ii) any player is moral with probability $\gamma > 0$, with realisations of moral character being independent across players; (iii) all other players are of type θ_b. A player of type θ_n (i) believes that the moral code is false, (ii) has prior beliefs about people's moral character identical to those of θ_b, (iii) and believes that all other players are of type θ_n.

Formally, let $\Gamma_i(\theta_i) = p^i_{\theta_i,0}(.)$. Suppose the vector $\mathbf{c}$ has x elements of 1 and $(n-x)$ elements of 0. Then we have the following prior probabilities when player i is, respectively, of type θ_b and θ_n.

$$p^i_{\theta_b,0}(\boldsymbol{\theta},\mathbf{c},m) = \begin{cases} (1-\gamma)^x\,\gamma^{(n-x)} \text{ for } \boldsymbol{\theta} = (\theta_b,..,\theta_b)\,,\ m=1 \\ 0 \text{ if } m=0 \text{ or } \boldsymbol{\theta} \neq (\theta_b,..,\theta_b) \end{cases} \tag{15}$$

$$p^i_{\theta_n,0}(\boldsymbol{\theta},\mathbf{c},m) = \begin{cases} (1-\gamma)^x\,\gamma^{(n-x)} \text{ for } \boldsymbol{\theta} = (\theta_n,..,\theta_n)\,,\ m=0 \\ 0 \text{ if } m=0 \text{ or } \boldsymbol{\theta} \neq (\theta_n,..,\theta_n) \end{cases} \tag{16}$$

Using θ_b and θ_n, we construct other types as follows: a player of type $\theta_b(Y)$ believes that the moral code is true and assigns positive probability to each type in the set Y, while a player of type $\theta_n(Y)$ believes that the moral code is false and assigns positive probability to all types in the set Y. Thus, type $\theta_n(\{\theta_b\})$ does not believe in the moral code and belives that all other players are of type θ_b; type $\theta_b(\{\theta_b,\theta_n\})$.believes in the moral code 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 moral code 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.

3.3 The Evolution of Beliefs

Next, we specify how the beliefs held by players evolve in the game. Given the state space defined above, we define a *belief correspondence* for each player i: $\mathcal{B}^i_t : \Omega_t \rightarrow 2^{\Omega_t}$ for $t = 1, 2, 3, ...$

The belief correspondence describes, for each player, in each period and in each possible state of the world, the states to which he or she assigns positive probability. 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. Therefore, for player i of type θ_i, we let

$$\mathcal{B}_0^i = \left\{ \omega \in \prod_{i \in \mathcal{I}} \Theta_i \times \Sigma : p^i_{\theta_i,0}(\omega) > 0 \right\} \tag{17}$$

Recall that players observe actions in the game (after trembling) in each period but not the moral character of players or the moral code. We assume that he revises his subjective probabilities on the basis of observed actions using Bayes' rule wherever possible. To be precise, let $h_t = (h_{t-1}, e_t, \tilde{a}_t)$ be the history realised in period t. Let ω_t be a possible period t state of the world and ω_{t-1} the period $t-1$ state implied by ω_t and h_t. For a given strategy profile σ (which will be defined in more detail in the next section), we can compute the conditional *objective* probability $\hat{\sigma}_t(\tilde{a}_t|\omega_{t-1}, e_t, \sigma)$ that the outcome $\tilde{a}_t$ will be observed after state ω_{t-1} and the move e_t by Nature have 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

$$p^i_{\theta_i,t}(\omega_t|h_t) = \frac{p^i_{\theta_i,t-1}(\omega_{t-1}|h_{t-1})\,\hat{\sigma}_t(\tilde{a}_t|\omega_{t-1}, e_t, \sigma)}{\sum\limits_{\omega'_{t-1} \in \Omega_{t-1}} p^i_{\theta_i,t-1}\left(\omega'_{t-1}|h_{t-1}\right)\hat{\sigma}_t\left(\tilde{a}_t|\omega'_{t-1}, e_t, \sigma\right)} \text{ if } \omega_t \subset \mathbf{E}(h_t) \tag{18}$$

$$p^i_{\theta_i,t}(\omega_t|h_t) = 0 \text{ if } \omega_t \subsetneq \mathbf{E}(h_t) \tag{19}$$

where $\mathbf{E}(h_t)$ denotes the event that history h_t has been realised. Thus, equations (18) and (19) give 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.

Note that equation (18) 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 (18) is positive. If action a_t was a zero probability event given i's prior beliefs, and the strategy profile σ, we ensure that i's beliefs following action a_t satisfy the *consistency* criterion, proposed by Kreps and Wilson (1982) (discussed further in the next section).

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 period t belief set can be written as

$$\mathcal{B}_t^i(\omega_t) = \left\{ \omega \in \Omega_t : p^i_{\theta_i,t}(\omega|h_t) > 0 \right\} \tag{20}$$

3.4 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, $\sigma_t^i \left(h_{t-1}, e_w^i, \theta_i\right)$ denotes the probability that player i of type θ_i chooses $a_w^i = 1$ (i.e. chooses to engage in act 'X') when the event e_w^i occurs following history h_{t-1}, and $\sigma_t^i \left(h_{t-1}, e_o^{ij}, \theta_i\right)$ denotes the probability that player i of type θ_i chooses the action $a_o^{ij} = 1$ (i.e. chooses to associate with person j) when event e_o^{ij} occurs following history h_{t-1}.

We represent person i's full strategy by $\sigma_i = \left(\sigma_t^i\right)_{t \in \mathbb{N}^+}$ and a strategy profile of the game by $\sigma = \left(\sigma_i\right)_{i \in \mathcal{I}}$. As per the notation introduced in Sections 3.2 and 3.3, we represent players' posterior beliefs by probability functions $p_{\theta_i,t}^i \left(.|h_t\right)$ for each player $i \in \mathcal{I}$, player type $\theta_i \in \Theta_i$, period $t \in \mathbb{N}^+$ and history $h_t \in \mathcal{H}_t$. We define an indirect utility function $V^i\left(.\right)$ as follows:

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

where $h_t = \left(h_{t-1}, e, a\right)$, $a = \left(a_i, a_{-i}\right)$, and $u^i\left(.\right)$ is as defined in (6). Similarly, we denote by $V_t^i\left(\sigma, \theta_i|h_t\right)$ the expected payoff to person i following history h_t. We define an equilibrium as a strategy profile σ, prior beliefs $p_{\theta_i,0}^i\left(.\right)$ and posterior beliefs $p_{\theta_i,t}^i\left(.\right)$ such that

1. for each $i \in \mathcal{I}$, $t \in \mathbb{N}^+$, $h_t \in \mathcal{H}_t$,

   $$V_t^i\left(\sigma_i, \sigma_{-i}, \theta_i|h_t\right) \geq V_t^i\left(\sigma_i', \sigma_{-i}, \theta_i|h_t\right)$$

   for all other player i strategies σ_i';

2. at each information set $\mathbf{E}\left(h_t\right)$ that person i believes will be reached with positive probaility, $p_{\theta_i,t}^i\left(.\right)$ are obtained from (18)-(19);

3. at each information set $\mathbf{E}\left(h_t\right)$ that person i believes will be reached with zero probaility, $p_{\theta_i,t}^i\left(.\right)$ will satisfy the *consistency criterion* (Kreps and Wilson, 1982) based on the rules for updating beliefs described by (18)-(19).

3.5 Characterisation of Equilibria

There are three conditions which are key for the following analysis. The first two are given by (7) and (8) which ensure, respectively, that an individual would refrain from choosing 'X' and ostracise others when required to do so when the alternative is perpetual social ostracism (and there is a probability ε of trembling by each player).

The third condition is based on an assumption that when one person believes another to be amoral, the cost of associating with this person is sufficiently high that the former will ostracise the latter regardless of the strategies pursued by others. The cost of associating with an amoral person is R. On the other hand, the minmax punishment that can be inflicted on an individual is for the rest of the society to engage in perpetual ostracism, which would result in an expected utility loss of at most $\frac{\beta(n-1)\delta_o}{1-\beta}R$. This expression is identical to the right-hand side of (7) and (8) except that we replace the term ε by 1 and β_ε by β. The reason is that if a player i believes the moral code to be true and assigns another player j a probability 1 of being moral following some history, then i will also assign j a zero probability of trembling when making a moral choice. Therefore, the following is a sufficient condition for one to choose to ostracise an amoral person regardless of the strategies pursued by others in the population:

$$R > \frac{\beta(n-1)\delta_o}{1-\beta}R$$

$$\Longrightarrow \frac{1-\beta}{\beta} > (n-1)\delta_o \tag{21}$$

Population of θ_b individuals: First, we consider equilibrium behaviour in a population consisting entirely of individuals of type θ_b. The formal reasoning is provided in the proof of Proposition 1. But the structure of the argument will be evident from the following informal description. As per (15), each individual i has prior beliefs that the moral code is true and that all other individuals are also of type θ_b. The procedure for updating beliefs, described in Section 3.3, implies that i will take the moral code to be true in every period, regardless of the previous history of the game. This is formally shown in Lemma 2 in the Appendix. If some person j is seen to engage in act 'X' in some period t, then i updates her beliefs and assigns a probability of 1 to the event that j is amoral. Then, if the condition in (21) is satisfied, i will ostracise j regardless of the strategies pursued by other players. If the condition in (7) is satisfied, the threat of ostracism will mean that j does not choose act 'X'.[6]

Similarly, if j engages in ostracism against someone who has not violated the moral code (e.g. who has not engaged in act 'X'), then i updates her beliefs and assigns a probability of 1 to the event that j is amoral. Then, if the condition in (21) holds, j will necessarily be subject to ostracism in subsequent periods. The threat of ostracism will mean that j will not choose to ostracise anyone who has not violated the moral code.

Following this reasoning, we obtain a full characterisation of equilibrium strategies. We denote by $\gamma_i(h_t)$ the belief about person i's moral character held by an individual of type θ_b

[6] Note that j may assign i a lower probability of trembling than that implied by (7), at least following some histories, because j believes that the moral code is true. Nevertheless, the right-hand side of (7) provides the minimum cost to j of being subject to perpetual social ostracism following any history. Therefore, (7) provides a sufficient condition.

following history h_t.

Proposition 1 *If the population consists entirely of individuals of type θ_b and the conditions in (7) and (21) hold, the equilibrium strategies must include the following: After history (h_{t-1}, e_w^i), individual i chooses action 'not X' if and only if $\gamma_i(h_{t-1}) \geq \frac{1}{2}$. After history (h_{t-1}, e_o^{ji}), individual j ostracises i if and only if the latter has previously engaged in act 'X', or failed to ostracise someone who has engaged in act 'X', or failed to ostracise someone who has failed to ostracise someone who has engaged in act 'X', etc.*

According to Proposition 1, if the conditions in (7) and (21) hold and $\gamma \geq \frac{1}{2}$, then individuals will choose 'not X' whenever event e_w^i occurs. Recall that when someone is observed to engage in act 'X' or violates the moral code in any other way, individuals of type θ_b update their beliefs and conclude that the person in question is amoral, i.e. $\gamma_i(h_t) = 0$. Proposition 1 implies that such a person would face ostracism. It also implies that a person is not subject to ostracism outside of these circumstances. While the equilibrium described in Proposion 1 is unique, the observed behaviour is probabilistic given that players can tremble.

Population of θ_b and $\theta_n(\{\theta_b\})$ individuals: Next, we consider equilibrium behaviour in a population consisting of individuals of type θ_b and $\theta_n(\{\theta_b\})$. For ease of notation, we let $\theta_{nb} = \theta_n(\{\theta_b\})$. Individuals of type θ_b believe in the moral code and believe that everyone else in the population are also of type θ_b. Therefore, their behaviour corresponds to that described in Proposition 1. Individuals of type θ_{nb} do not believe in the moral code. But they do believe that others are of type θ_b and, therefore, that anyone who engages in act 'X' will be ostracised. Therefore, under condition (7) they will not engage in act 'X'. Furthermore, they believe that they will be ostracised if they do not participate in ostracism when the moral code dictates it. The expected cost of being ostracised exceeds the utility gain from associating with any other individual in the population under condition (8). Therefore, individuals of type $\theta_n(\{\theta_b\})$ will also ostracise those who engage in act 'X', and those who fail to ostracise those who engage in act 'X', etc. Following the same reasoning as in the case of type θ_b, they will not ostracise those whose actions have been in accordance with the moral code.

Population of individuals of type θ_b, $\theta_n(\{\theta_b\})$, $\theta_n(\{\theta_b, \theta_{nb}\})$, $\theta_b(\{\theta_b, \theta_{nb}\})$, etc.: For ease of notation, we let $\theta_{nnb} = \theta_n(\{\theta_b, \theta_{nb}\})$ and $\theta_{bnb} = \theta_b(\{\theta_b, \theta_{nb}\})$. Individuals of type θ_{nnb} do not believe in the moral code and believe that others are of type θ_b or θ_{nb}. Based on the reasoning above, individuals of type θ_{nnb} believe that they will be ostracised by everyone if they engage in act 'X'. Therefore, they will refrain from act 'X'. They also believe that they will be ostracised by everyone if they fail to ostracise anyone who has engaged in act 'X', or fail to ostracise anyone who has failed to ostracise anyone who has engaged in act 'X', etc. and so they will choose to ostracise anyone who has violated the moral code. Based on the same reasoning as above, they will not ostracise anyone who has not violated the moral code. Thus, individuals of type θ_{nnb} play the same strategy in equilibrium as those of type θ_b and θ_{nb}. Note that this reasoning did not depend on the individual's *own* belief regarding the moral code. Therefore, the argument also works for individuals of type θ_{bnb}.

By reasoning iteratively, we can show that the strategies pursued by individuals of type $\theta_n(\{\theta_b, \theta_{nb}, \theta_{nnb}, \theta_{bnb}\})$, $\theta_b(\{\theta_b, \theta_{nb}, \theta_{nnb}, \theta_{bnb}\})$, etc. – will also opt for the same strategy. In particular, 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 us denote this set of types by $\Theta_b = \{\theta_b, \theta_{nb}, \theta_{nnb}, \theta_{bnb}, ...\}$. Then, we can obtain the following result shown in the Appendix.

Proposition 2 *If all individuals in the population belong to a type in $\Theta_b \subset \Theta$ and the conditions in (7), (8), and (21) hold, the equilibrium strategy must include the following: After history (h_{t-1}, e_w^i), individual i chooses action 'not X' if and only if $\gamma_i(h_{t-1}) \geq \frac{1}{2}$. After history (h_{t-1}, e_o^{ji}) individual j ostracises person i if and only if $\gamma_j(h_{t-1}) \geq \frac{1}{2}$ or $\theta_j = \theta_b$, and person i has previously engaged in act 'X', or failed to ostracise someone who has engaged in act 'X', or failed to ostracise someone who has failed to ostracise someone who has engaged in act 'X', etc.*

Proof. See the Appendix. ■

Proposition 2 describes behaviour which is identical to that in Proposition 1 with two key differences. First, we require an additional condition (8), which implies that individuals who do not believe in the moral code can be induced to ostracise even those who they believe to be moral when they themselves are faced with the threat of ostracism. Second, those who do not believe in the moral code and are being subject to ostracism due to a poor reputation (i.e. $\gamma_i(h_t) < \frac{1}{2}$) cannot be induced to ostracise others who have violated the moral code.

Proposition 2 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 moral code can be used to pin down equilibrium behaviour. By contrast, it should be evident that the reasoning behind Proposition 2 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.

3.6 Additional Properties of Equilibria

In this section, we discuss some important qualities of the type of equilibrium described in Propositions 1 and 2. The simplest type of equilibrium obtains if every individual in the population is of type θ_b. Then they all believe in the association between act 'X' and the notion of 'moral character' embodied in the moral code and behave accordingly. Thus we obtain a population of *homo sociologicus* who avoid the forbidden act and ostracise 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: In a population 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 act 'X' and the notion of 'moral character' but they all believe that everyone else does. They follow the behaviour implicitly prescribed by the moral code, because if they do not (they believe) others will conclude they have amoral character and ostracise them.

In a population consisting entirely of θ_{nnb} individuals, everyone believes, accurately, that others do not believe in the moral code. This can be seen from the fact that if individuals i and j are of type θ_{nnb}, then we have, by construction, $\mathcal{B}_0^i \subset \mathbf{E}\left(\mathcal{B}_0^j \subset \mathbf{E}(m=1)\right) \cup \mathbf{E}\left(\mathcal{B}_0^j \subsetneq \mathbf{E}(m=1)\right)$ (since i believes j to be of either type θ_b or type θ_{nb}; a θ_b individual believes in the moral code but a θ_{nb} individual does not) and $\mathcal{B}_0^j \subsetneq \mathbf{E}(m=1)$ (since a θ_{nnb} individual does not believe in the moral code). However, they have inaccurate beliefs about what others believe about whether others believe in the moral code (since, by construction, $\mathcal{B}_0^i \subset \mathbf{E}\left(\mathcal{B}_0^j \subset \mathbf{E}\left(\mathcal{B}_0^i \subset \mathbf{E}(m=1)\right)\right)$ but $\mathcal{B}_0^j \subset \mathbf{E}\left(\mathcal{B}_0^i \subsetneq \mathbf{E}(m=1)\right))$. In other words, the second-order beliefs are inaccurate. And this causes everyone to behave in accordance with the moral code, because they believe that if they do not they will be ostracised by others.

In a population 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 behave in accordance with the moral code, because if they do not, they will be ostracised by others.

Renegotiation: An important critic of the 'economic' explanation for sustained social norms – whereby potential deviators from the social norm face the threat of social ostracism, and those who deviate from ostracising behaviour face a further threat of social ostracism, etc. – is that these mechanisms are not 'renegotiation-proof' (Farrell and Maskin, 1989; Bernheim and Ray 1989). In the context of the model presented here, the criterion means that it cannot be that individuals 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.

In the type of equilibrium described in Section 3.5, 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. There is no other possible equilibrium where players of all types are at least as well-off and, therefore, renegotiation cannot be welfare improving. Thus, the criticism based on renegotiation-proofness does not apply to mechanism for sustaining social norms described in Section 3.5.

Strategic Experimentation: The fact that the probability of trembling depends on the underlying state of the world (as represented by Σ) raises the question whether individuals have an incentive to 'experiment' to learn about the state. If an individual is certain that the moral code is true ($m = 1$) or that the moral code is false ($m = 0$), then they will not update their beliefs about m following any history (this is shown formally, for players of type θ_b, in Lemma 1 in the Appendix). On the other hand, they will update their beliefs about their own moral character if they tremble. In particular, if a person of type θ_b chooses an action that corresponds to the 'moral choice' and trembles, she would conclude that she is amoral (because the moral code implies that a moral person does not tremble). Likewise, if she chooses an action that corresponds to the 'immoral choice' and does not tremble, she would again conclude that she is amoral. In both instances, according to the equilibrium described in Proposition 1 and Proposition 2, she would be subjected to perpetual ostracism by others. And there is no strategic gain from her new perceived knowledge that would override the cost of perpetual ostracism. Therefore, she has no incentive to experiment in this manner.

4 The Persistence of Social Norms

In the preceding section, we showed that there is a unique equilibrium in which a social taboo against act 'X' is sustained when all individuals are either (i) of type θ_b, or (ii) have nth order and higher order beliefs that all individuals are of type θ_b. There are two potential concerns with these results. First, the equilibrium is unique only for a specific moral code. Therefore, these results do not allow us to predict what type of equilibrium would emerge without prior knowledge of the relevant moral code for a population. Second, as the results depend on players belonging to a specific type set, they raise the question how restrictive are the assumptions imposed on the belief structure. We address these points below.

Social scientists have exerted a good deal of effort in documenting, for a range of different societies, the moral codes that have traditionally played an important role in social life.[7] These moral codes may be rooted in religion, customary law, or tradition and, in the case of the most important moral codes, it is reasonable to assume that, at some point in the past, belief in them was widespread and common knowledge. We can represent such a state of affairs within the framework of our model by assuming that the population was composed entirely of individuals of type θ_b. Then, according to Proposition 1, everyone would abide by the moral code. (See Posner 1998 and Ellickson 2001 for related discussions on the emergence of social norms).

The key insight of the theory follows from Proposition 2, which implies that individuals continue to abide by the moral code – i.e. refrain from actions that are forbidden by it, ostracise those who violate the code, etc. – even when they cease to believe in the moral code, as long as they retain higher order beliefs in it above some finite nth order. Under what conditions would individuals retain higher holder beliefs in a moral code when they

[7] George Murdock's *Ethnographic Atlas* (Murdock 1967), for example, provides a compilation of customary practices and norms across a large number of ethnic groups or societies drawing on ethnographic studies.

do not believe in it themselves? To address this question, we present here an example of a dynamic population where such a belief structure emerges over time.

Rather than having a static population as in Section 3, we introduce, for each individual i, in each period, 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 in the population 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 individuals in period t. The set of possible states in period t can be represented as

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

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 in $\mathcal{I}(0)$ – will be assigned a 'moral character' ('moral' with probability γ and 'amoral' with probability $1-\gamma$). The realisation of moral character will be independent across players. (Recall that players do not observe their own moral character or the moral character of others). Each new player will also be assigned a belief about the moral code: either a belief that the moral code is true (probability μ_0) or that it is false (probability $1-\mu_0$). Each new player observes the types of all existing players when she joins the population. However existing players (erroneously) believe that each new player is of the same type as the one she has replaced.[8] This last assumption will have significant implications for the belief structure in the game.

Suppose that those initially present in the population are of type θ_b - i.e. they believe that the moral code is true and that all others in the population are also of type θ_b. By construction, these beliefs are indeed accurate. When one such player is exogenously replaced by another – who may or may not believe in the moral code – this replacement is ignored by others in the population. The newcomer observes the type of the existing individuals in the population, and thus forms the belief (accurately) that all others are of type θ_b. Therefore, the newcomer is of type θ_b or type θ_{nb}, depending on her own beliefs. When another newcomer joins the population in a subsequent period, the process is repeated. Beliefs of the existing players do not change. But the newcomer is of type θ_b or θ_{nb} or θ_{nnb} or θ_{bnb}. Reasoning thus, we conclude that after any finite number of periods, all players will be of some type $\theta \in \Theta_b$. Formally, we have the following result.

Lemma 1 *Consider a population where all individuals are of type θ_b in period $t = 0$. At the start of every period, each player faces an exogenous probability ζ of death and replacement*

[8] Note that multiple players may die and be replaced in the same period. In this case, we assume that the new players who replace them do not observe each others' type but, rather, believe that each of the others is of the same type as the individual he or she has replaced.

by a new player. We assume that (i) Each new player observes the beliefs of existing players and, with some probability $\mu_0 \in (0,1)$, is assigned the belief that the moral code is true (and otherwise that the moral code is false); (ii) Existing players believe that each new player has the same belief as the player he or she has replaced. Then, in every period t, each player has a type $\theta \in \Theta_b$.

Proof. See the Appendix. ■

The following corollary follows immediately from Lemma 1 and Proposition 2.

Corollary 1 *Consider a population where individuals face death and replacement as per the description in Lemma 1. Suppose the per-period discount factor is equal to $\beta_\zeta = \frac{\beta}{\zeta}$. Then, under conditions (7), (8) and (21), the characterisation of the equilibrium strategies in the statement of Proposition 2 must hold true.*

Proof. See the Appendix. ■

Corollary 1 highlights how a social taboo against act 'X' may persist even when beliefs regarding the moral code evolve – in the sense that newcomers may have different beliefs from those whom they replace. To obtain this result, we assumed that individuals are fully informed about the beliefs of existing players when they join the population but fail to update these beliefs when individuals 'die' and are replaced by others. By contrast, if the lack of belief in the moral code becomes common knowledge, the Proposition 2 does not apply. In this case, it may still be possible sustain a social taboo against act 'X' but, as discussed at the beginning of Section 3, multiple equilibria are feasible.

5 Conclusion

In this paper, we proposed a mechanism for sustaining a social norm in a population against some behaviour distinct from both the dominant economic and sociological approaches to the issue. The norm is underpinned by a 'moral code' which classifies certain possible actions as being either 'moral' or 'immoral'. In a world where the 'moral code' is 'true', those who are 'moral' are less likely to tremble when undertaking a 'moral' action and more likely to tremble when undertaking an 'immoral' action, compared to those who are 'amoral'. If the moral code is false, all individuals are equally likely to tremble when undertaking any action.

If those believed to be 'amoral' are subject to ostracism and everyone believes in the moral code, then we obtain the expected result that individuals always choose actions that conform to the moral code. Our main focus, however, is on how higher order beliefs regarding the moral code affect people's choices. We show that even when everyone believes that the moral code is false, and believes that everyone else believes that the moral code is false, and believes that everyone else believes that everyone else believes that the moral code is false, etc. their actions will conform to the moral code if it is believed to hold true above some finite nth order.

These results imply that if beliefs in the moral code evolve in a way that is not common knowledge – for example, the beliefs of newcomers are not observed by others in the population, then the social norm would persist even when no one believes that the moral code is true. This suggests that events or mechanisms that cause beliefs to become common knowledge can precipitate a sudden change in social norms.

6 Appendix

Lemma 2 *An individual of type θ_b believes the moral code is true in every period, regardless of the history of the game.*

Proof. of Lemma 2: Let (e_τ, a_τ) be the move by Nature and action in period τ, and h_τ be the full history at the end of this period. Let ω_τ be the true state of the world in period τ. Consider any $\hat{\omega}_\tau \in \{(h_\tau, \boldsymbol{\theta}, \mathbf{c}, m) \in \Omega_\tau : m = 0\}$ and denote by $\hat{\omega}_{\tau-1}$ the period $\tau - 1$ state implied by $\hat{\omega}_\tau$ and h_τ.

If there exists some $\omega'_{\tau-1} \in \Omega_{\tau-1}$ such that $p^i_{\theta_b,\tau-1}\left(\omega'_{\tau-1}\right) > 0$ and $\hat{\sigma}_t\left(a_\tau|\omega'_{\tau-1}, e_\tau, \sigma\right) > 0$, then we can use (15) to compute beliefs in period τ as follows:

$$p^i_{\theta_b,\tau}\left(\hat{\omega}_\tau|h_\tau\right) = \frac{p^i_{\theta_b,\tau-1}\left(\hat{\omega}_{\tau-1}\right)\hat{\sigma}_t\left(a_\tau|\hat{\omega}_{\tau-1}, e_\tau, \sigma\right)}{\displaystyle\sum_{\omega'_{\tau-1}\in\Phi_{\tau-1}} p^i_{\theta_b,\tau-1}\left(\omega'_{\tau-1}\right)\hat{\sigma}_t\left(a_\tau|\omega'_{\tau-1}, e_\tau, \sigma\right)} \tag{22}$$

for $\tau = 1, 2, ...$ Using the definition of θ_b in (15), we have $p^i_{\theta_b,0}\left(\hat{\omega}_0\right) = 0$. It follows from the equation above that $p^i_{\theta_b,1}\left(\hat{\omega}_1|h_1\right) = 0$ and, by iteration, we can show that $p^i_{\theta_b,\tau}\left(\hat{\omega}_\tau|h_\tau\right) = 0$ for each $\tau > 1$.

If, on the other hand, the denominator on the r.h.s. of (22) is zero, then we must construct beliefs such that they satisfy the *consistency* criterion. For this purpose, we let

$$p^{i,n}_{\theta_b,\tau}\left(\hat{\omega}_\tau|h_\tau\right) = \frac{p^{i,n}_{\theta_b,\tau-1}\left(\hat{\omega}_{\tau-1}\right)\hat{\sigma}_t\left(a_\tau|\hat{\omega}_{\tau-1}, e_\tau, \sigma^n\right)}{\displaystyle\sum_{\omega'_{\tau-1}\in\Phi_{\tau-1}} p^{i,n}_{\theta_b,\tau-1}\left(\omega'_{\tau-1}\right)\hat{\sigma}_t\left(a_\tau|\omega'_{\tau-1}, e_\tau, \sigma^n\right)}$$

where σ^n denotes a completely mixed strategy profile and $p^{i,n}_{\theta_b,0}(.) \equiv p^i_{\theta_b,0}(.)$. As before, we can show that $p^{i,n}_{\theta_b,\tau}\left(\hat{\omega}_\tau|h_\tau\right) = 0$ for each $\tau > 1$. Consider any sequence $\left\{\left(\sigma^n, p^{i,n}_{\theta_b,\tau}\right)\right\}_{n=1}^{\infty}$ and let

$$\left(\sigma, p^i_{\theta_b,\tau}\right) = \lim_{n\longrightarrow\infty}\left\{\left(\sigma^n, p^{i,n}_{\theta_b,\tau}\right)\right\}_{n=1}^{\infty}$$

Since $p^{i,n}_{\theta_b,\tau}\left(\hat{\omega}_\tau|h_\tau\right) = 0$ for each σ^n, it follows that $p^i_{\theta_b,\tau}\left(\hat{\omega}_\tau|h_\tau\right) = 0$.

Thus, a person of type θ_b assigns zero probability to any state in which the moral code is false. ■

Proof. of Proposition 1: First, we show that, under condition (21), it is optimal for a person of type θ_b to ostracise those who have been observed to engage in act 'X' regardless of the strategies pursued by other players. Suppose that $(e_w^j, 1) \in h_t$ and the move by Nature in period $t+1$ is e_o^{ij}; i.e. person i has the opportunity to ostracise person j in period $t+1$ and person j has previously engaged in act 'X'. Using Lemma 2, we have $p^i_{\theta_b,t}(\mathbf{E}(m=1)) = 1$. Then, using (10), (11) and (18), we have $p^i_{\theta_b,t}(\mathbf{E}(c_j = 1)) = 0$. The last step follows from the definition of $\mathcal{M}_t$ and the fact that $(e_w^j, 1) \in h_t$. Therefore, if e_o^{ij} occurs in period $t+1$ and person i chooses to ostracise j, he receives an additional utility of $E_i(1-c_j)R = R$. Under condition (21), 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 $t+1$. Therefore, $\sigma^i_{t+1}(h_t, e_o^{ij}, \theta_b) = 1$.

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 act 'X', regardless of the strategies pursued by other players. Suppose $h_r = \left(h_{r-1}, e_w^l, 1\right)$, $h_s = \left(h_{s-1}, e_o^{jl}, 0\right)$, and the move by Nature in some period $t+1$ is e_o^{ij}; where $r < s \leq t$ and $h_r \subset h_s \subseteq h_t$; i.e. person l engaged in act 'X' in period r, 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 $t+1$. Using the reasoning above, $p^i_{\theta_b,t}(\mathbf{E}(c_l = 1)) = 0$. Therefore, using (12), (13) and (18), we have $p^i_{\theta_b,t}(\mathbf{E}(c_j = 1)) = 0$. Therefore, if e_o^{ij} occurs in period $t+1$ and person i chooses to ostracise j, he receives an additional utility of $E_i(1-c_j)R = R$. Therefore, under condition (21), it is optimal for person i to ostracise person j in period $t+1$. Therefore, $\sigma^i_{t+1}(h_t, e_o^{ij}, \theta_b) = 1$.

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 engaged in act 'X', 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 act 'X'.

Suppose the move by Nature in some period $t+1$ is e_w^i and $\gamma_i(h_t) \geq \frac{1}{2}$. Based on the reasoning above, if i chooses action 'X' in period $t+1$, then i will be ostracised in subsequent periods. The expected disutility to i from such ostracism is at least equal the expression on the right-hand side of (7). Therefore, if the condition in (7) holds, then the expected disutility from action 'X' must exceed the benefits. Therefore, i will choose action 'not X'. By contrast, if $\gamma_i(h_{t-1}) < \frac{1}{2}$, then i will be subject to ostracism in subsequent periods even if he is not observed to engage in act 'X', while engaging in act 'X' yields a utility of W. Therefore i will choose action 'X'. ■

Proof. of Proposition 2: First note that, by definition, individuals of type θ_b believe that everyone else is of type θ_b. Therefore, under conditions (7) and (21), their equilibrium strategies will be as described in the statement of Proposition 1. Next, we can show that, given the strategy pursued by individuals of type θ_b, under condition (8) it is optimal for a person of type θ_{nb} to ostracise those who have previously been observed to engage in act 'X'. By definition, if person j is of type θ_{nb}, then he assigns, ex-ante, a probability of 1 to all other individuals being of type θ_b if $\gamma_j(h_{t-1}) \geq \frac{1}{2}$. We have already shown, in the proof of Proposition 1, that $\sigma_t^i(h_{t-1}, e_o^{ij}, \theta_b) = 1$ if $h_{t-1} \supseteq h_s \supset h_r$, $h_r \ni (e_w^l, 1)$, $h_s = (h_{s-1}, e_o^{jl}, 0)$, $r < s < t$. 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. The expected disutility to j from such ostracism is at least equal the expression on the right-hand side of (8). Therefore, under condition (8), the expected cost of this punishment exceeds the cost of ostracising person l in period s. Therefore, $\sigma_s^j(h_{s-1}, e_o^{jl}, \theta_{nb}) = 1$; i.e. it is optimal for person j to ostracise any person l who has previously engaged in act 'X'. If $\gamma_j(h_{t-1}) < \frac{1}{2}$, then j will expect to be subject to ostracism in subsequent periods regardless of his action following (h_{s-1}, e_o^{jl}). Therefore j is better-off associating with person l.

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 act 'X' if and only if $\gamma_j(h_{t-1}) \geq \frac{1}{2}$; 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 act 'X' if and only if $\gamma_j(h_{t-1}) \geq \frac{1}{2}$. Following the reasoning applied to individuals of type θ_b, we can also show that, under condition (7), when event e_w^i occurs, person i of type θ_{nb} will choose action 'not X' if and only if $\gamma_i(h_{t-1}) \geq \frac{1}{2}$.

Using the same type of reasoning as above, we can show that, under conditions (7) and (8), it is optimal for individuals of 'higher types' in Θ_b to choose 'not X', and ostracise those who have previously been observed to engage in act 'X', and ostracise those who have previously failed to ostracise someone has engaged in act 'X', etc. if and only if $\gamma_i(h_{t-1}) \geq \frac{1}{2}$. ■

Proof. of Lemma 1: Let us denote by θ_i the type of individual i. By assumption, $\theta_i = \theta_b$ for each $i \in \mathcal{I}(0)$. Let $\mathcal{I}^+(1)$ be the set of indices of players who join the population in period 1. By assumpion, each $i \in \mathcal{I}^+(1)$ believes that each $i' \in \mathcal{I}(1)$ is of type θ_b. Furthermore, if $i \in \mathcal{I}^+(1)$, then i either believes that the moral code is true (i.e. $m = 1$) or that the moral code is false (i.e. $m = 0$). Then, using the definitions in Section 3.2, we obtain $\theta_i \in \{\theta_b, \theta_{nb}\}$ for each $\theta_i \in \mathcal{I}^+(1)$. Let $\mathcal{I}^+(2)$ be the set of indices of players who join the population in period 2. Reasoning as above, we obtain $\theta_i \in \{\theta_b, \theta_{nb}, \theta_{nnb}, \theta_{bnb}\}$ for each $\theta_i \in \mathcal{I}^+(2)$. Similarly, we obtain $\theta_i \in \{\theta_b, \theta_{nb}, \theta_{nnb}, \theta_{bnb}, \theta_{nxnb}, \theta_{bxnb}\}$ for each $\theta_i \in \mathcal{I}^+(3)$ where $\theta_{bxnb} = \theta_b(\{\theta_b, \theta_{nb}, \theta_{bnb}, \theta_{nnb}\})$ and $\theta_{nxnb} = \theta_n(\{\theta_b, \theta_{nb}, \theta_{bnb}, \theta_{nnb}\})$. By iteration, we see that all types constructed in this manner will consist exclusively of θ_b above some finite nth order. Therefore, in each period t, for each $i \in \mathcal{I}(t)$, we have $\theta_i \in \Theta_b$. ■

Proof. of Corollary 1: Consider a player i who enters the game in some period $t > 0$. As per Lemma 1, $\theta_i \in \Theta_b$. Player i's strategy will be a best-response for the player types in period $t-1$ as i believes that all players who join the game in subsequent periods will be of the same type as those they replace. Because of the probability of death ζ in each period, player i discounts future payoffs by $\zeta\beta_\zeta = \beta$ per period. Therefore, under conditions (7), (8) and (21), player i's strategy will be as described in the statement of Proposition 2. ■

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] Banerjee, A (1992). "A Simple Model of Herd Behavior", *The Quarterly Journal of Economics*, Vol. 107(3), pp. 797-817.

[3] Bénabou, Roland, and Jean Tirole (2011). "Identity, Morals, and Taboos: Beliefs as Assets." *The Quarterly Journal of Economics*, Vol. 126(2), pp. 805-855.

[4] Bernheim, B. and D. Ray (1989). "Collective dynamic consistency in repeated games." *Games and Economic Behavior*, Vol. 1(4), pp.295-326.

[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] Chamley, C. (1999). "Coordinating Regime Switches", *Quarterly Journal of Economics*, Vol. 114(3), pp. 869-905.

[9] Carlsson, H. and E. van Damme (1993). "Global Games and Equilibrium Selection", *Econometrica*, Vol. 61, pp. 989-1018.

[10] Chen, Yi-Chun (2012). "A Structure Theorem for Rationalizability in the Normal Form of Dynamic Games", *Games and Economic Behaviour*, Vol. 75, pp. 587-597.

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

[15] Kuran, Timur (1987). "Preference Falsification, Policy Continuity and Collective Conservatism", *Economic Journal*, Vol. 47, pp. 642-665.

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

[17] Murdock, G.P. (1967). "Ethnographic Atlas: A Summary", *Ethnology*, Vol. 6(2), pp. 109-236.

[18] Parsons, Talcott (1951). *The Social System.* Routeledge, New York, 1951.

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

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

[21] 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.