Moral Codes and the Dynamics of Social Norms

Zaki Wahhaj
University of Kent*

February 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 to these theories, that offers insights about the persistence of social norms and the manner in which they may evolve. Our underlying assumption is that people derive utility from ostracising those who they believe to have 'bad character' but there is uncertainty and learning about how the character of a person may be inferred. Using this framework, we can account for both social norms that persist for long periods and unravel suddenly and those that evolve gradually over time. We apply the model to explain the phenomenon of persistent low labour market participation of women in Bangladesh.

JEL Codes: D01, D02, D83, Z10

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

1 Introduction

The role of social norms in human behaviour, why people follow them and how they evolve are topics which have received a great deal of attention within the social sciences over the years (see, for example, Elster 1989, Fehr and Fischbacher 2004 and Bicchierri 2011 for reviews of this literature).

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). Relatedly, Bowles and Gintis (2011) argue that there is a high prevalence of 'social preferences' – which leads to ethical or moral behaviour – across human societies and over time. While the premise that people are 'hard-wired' to follow social norms provides a robust answer to the question why they persist over long periods of time, it provides limited insight about the conditions that would cause social norms to evolve.

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). They typically generate multiple equilibria, and provide no simple insight about how we move from one equilibrium to another.

In this paper, we present an alternative to these theories, that offers insights about the persistence of social norms and the manner in which they may evolve. The key elements of the model of social behaviour explored in this paper are the following. (i) People carry an attribute called 'moral character'. A person's 'moral character' may be either 'good' or 'bad'. (ii) They derive utility from associating with someone whom they believe to have 'good character' and disutility from associating with someone whom they believe to have 'bad character'. (iii) The character of individuals are inferred on the basis of 'maxims' which take the following form: "A person who engages in act X has good/bad character". People may have different beliefs regards the truth of a maxim and they update their beliefs both about a person's character and the truth of a maxim based on his or her choice of actions.

This approach creates dynamics in social norms, even in the absence of exogenous shocks, and a unique equilibrium path. Using this approach, we can account both for social norms that (i) persist for long periods and unravel suddenly; and (ii) those that evolve gradually over time.

While it has been argued that social norms are typically slow-moving (Roland 2004),

the literature also 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). Therefore, it is important for a theory of changing social norms to explain both gradual and sudden changes, which the mechanism modelled in this paper is able to do.

The theory developed here is related to a number of papers in the existing literature on social behaviour. Bernheim (1994) proposes a model where people's consumption choices are motivated not only by intrinsic utility but also status concerns, and a person's status is determined by public perceptions of his/her 'predispositions'. If status concerns are important enough, this can lead to conformist behaviour and, if intrinsic preferences are assumed to evolve over time, it can account for both persistent and transitory norms. In relation to Bernheim (1994), a key contribution of this work is that it can account for social norms evolving under its own dynamics, in the absence of any exogenous shocks or changes in the environment.

Benabou and Tirole (2011) develop a theory of moral behaviour based on the idea that individuals make assessments regarding their own moral qualities based on choices that they and their peers have made in the past. Taboos and ostracising behaviour arises to block information that could damage self-image. Similar to Benabou and Tirole (2011), the agents in our model infer the moral qualities of others based on their past behaviour which can result in ostracism and taboos, but the mapping from actions to moral qualities itself evolves as a function of the history of past behaviour.

Benabou and Tirole (2006) use a setting similar to that of Bernheim (1994) – where individuals are motivated by altruism, greed as well as reputational concerns – to investigate the optimal level of (extrinsic) incentives for generating pro-social behaviour. Ellison and Fudenberg (1993) investigate how 'rules of thumb' for social learning – which are akin to the 'maxims' used in this paper – affect the pattern of technology adoption in a society. Wahhaj (2012) uses a similar framework to that used in this paper to investigate the role of higher-order beliefs in sustaining social norms.

In the second part of the paper, we provide an application of the theory to the phenomenon of low female labour force participation in certain traditional societies using the case of Bangladesh. Although women in Bangladesh have experienced dramatic social changes in the last three decades – including lower fertility, increased schooling, and increased employment opportunities in the manufacturing sector – only a small fraction are engaged in work outside of their homes. The sociological literature indicates that their choices are constrained by strong social norms restricting their presence in public spaces (the practice of 'purdah') as well as the traditional division of gender roles, which assigns women to domestic work within the home. Making use of the theoretical model and ethnographic and quantitative evidence, we offer an explanation as to why these social norms have evolved little until now and why they may do so in the future.

Social norms – and, in particular, social sanctions against behaviour which contradicts a set of informal rules – is often an important element in the functioning of informal institutions. It appears, for example, in theoretical explanations of informal risk-sharing in village societies (Kimball 1988; Fafchamps 1992; Coate and Ravallion 1994), the effectiveness of joint liability credit contracts in eliciting high repayment rates (Besley and Coate 1994) and the allocation of resources within households (Kazianga and Wahhaj 2013). Therefore, a theory about endogenous change in social norms can provide insights about how changes in informal institutions can occur in societies where they continue to play an important role today.

The remainder of this paper is organised as follows. In the next section, we provide an example via a simple model to highlight the main theoretical insights in the paper. A more general model is developed in Section 3 and the theoretical results are presented in Section 4. The application of the theory to the subject of female labour market participation is discussed in Section 5.

2 An Example on a Changing Moral Code

We begin by presenting a theoretical example which highlights the mechanism for changing social norms investigated in this paper. 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 $E\left[r\left(2c-1\right)|a\right]$ 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 has good 'moral character' and 0 otherwise. There is a maxim which states that "A person who has engaged in act 'X' has bad character". Let μ be a binary variable which takes the value of 1 if the maxim is true and 0 otherwise. Thus, there are four possible states of the world, which can be described by the pair (c, μ).

Person j's prior beliefs about the state of the world are as follows:

$$\begin{aligned} \Pr\left(c=0\right) &= \varepsilon \\ \Pr\left(\mu=0\right) &= \delta \end{aligned}$$

and the realisations of c and μ are independent.

We are now in a situation to analyse the behaviour of i and j in this game. Suppose person i has engaged in act 'X' in period 1. Then, there are three possible states of the

world which are feasible in period 2: $(c=1,\mu=0),(c=0,\mu=1)$ and $(c=0,\mu=0)$. The other state of the world, described by $(c=1,\mu=1)$, is ruled out because if the maxim were true, then i cannot have good 'moral character' (given that he has engaged in act 'X').

Therefore, person j's posterior beliefs about person i's character is given by

$$\begin{aligned}\Pr(c=1|a=1) &= \frac{\Pr(c=1,\mu=0)}{\Pr(c=0)+\Pr(c=1,\mu=0)} \\ &= \frac{(1-\varepsilon)\delta}{\varepsilon+(1-\varepsilon)\delta}\end{aligned}$$

Thus, if person i has engaged in act 'X', person j's utility from associating with him is equal to $r\left(2\frac{\delta(1-\varepsilon)}{\delta(1-\varepsilon)+\varepsilon}-1\right)$. Therefore, person j will ostracise person i if and only if

$$r\left(2\frac{\delta(1-\varepsilon)}{\delta(1-\varepsilon)+\varepsilon}-1\right)<0$$

$$\begin{aligned}&\Longrightarrow \quad 2\delta(1-\varepsilon)<\delta(1-\varepsilon)+\varepsilon \\ &\Longrightarrow \quad \delta<\frac{\varepsilon}{(1-\varepsilon)} \qquad (1)\end{aligned}$$

Likewise (since we have assumed that $r>w$), person i will refrain from act 'X' if and only if the condition in (1) holds. It is evident that the condition in (1) depends on the relative 'strengths' of j's beliefs in the moral code and in the good 'moral character' of person i.

If (1) is not satisfied, then person i engages in act 'X' and this action also leads j to update her beliefs in the maxim as follows:

$$\begin{aligned}\Pr(\mu=1|a=1) &= \frac{\Pr(c=0,\mu=1)}{\Pr(\mu=0)+\Pr(c=0,\mu=1)} \\ &= \frac{\varepsilon(1-\delta)}{\delta+\varepsilon(1-\delta)} \qquad (2)\end{aligned}$$

From (2), it is evident that if $\delta>0$ and $\varepsilon<1$, then $\Pr(\mu=1|a=1)<1-\delta$. In words, if person j initially had some doubt about the truth of the maxim and believed with positive probability that i has good 'moral character', then she becomes even less confident in the maxim once she has observed i engage in act 'X'. This change in belief can be a driving mechanism for changes in social behaviour. It is this mechanism that we will explore in the remainder of the paper.

3 Formal Model

In this section, we develop a more general model for analysing changes in social norms. 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 α_w^i which can take a value of 0 or 1; $\alpha_w^i = 1$ represents the action that 'person i engages in act X', and $\alpha_w^i = 0$ represents the action that he does not.

(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 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 associate' with j, and $\alpha_o^{ij} = 0$ represents the action that he does not.

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 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 $E[R(2c_i - 1)]$, where c_i and c_j are binary variables representing their moral character. We assume that c_i is unobservable for any individual $j \in \mathcal{I}$, including person i. Each c_i, $i \in \mathcal{I}$, is independant with $\Pr(c_i = 0) = \varepsilon$ where $\varepsilon > 0$.

If either of them choose not to associate, both individuals receive 0. Thus, the payoffs in the stage-game can be written as

$$u^i(a_i, a_{-i}, e) = \mathbf{I}\left(e_w^i\right)\alpha_w^i W + \sum_{j \neq i}\left\{\mathbf{I}\left(e_o^{ij}\right)\alpha_o^{ij} + \mathbf{I}\left(e_o^{ji}\right)\alpha_o^{ji}\right\} E[R(2c_j - 1)] \tag{3}$$

where $a_i = (\alpha_o^i, \alpha_w^i)$, $\alpha_o^i = (\alpha_o^{ij})_{j \neq i}$, $e = (e_o^i, e_w^i)_{i \in \mathcal{I}}$, $e_o^i = (e_o^{ij})_{j \neq i}$ and $\mathbf{I}(e)$ is an indicator function which takes a value of 0 or 1 depending on whether or not event e has occurred.

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.

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' into this framework – as in the example provided in the preceding section – narrows down the set of possible equilibria and provides a basis for analysing how norms may evolve, as well as the conditions under which they would evolve.

We represent a 'moral code' in terms of feasible states of the world. We also allow for individuals to believe in a moral code 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 in the following section.

3.1 A Framework for Modelling Interactive Beliefs

We shall propose a 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:

$$\Omega_t = \mathcal{H}_t \times \prod_{i \in \mathcal{I}} \Theta_i \times \Sigma \tag{4}$$

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.[1]

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

[1] For the analysis, we construct Θ_i as a particular countable type-space according to a procedure described in Section 3.2.

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, in each state of the world, a particular maxim about moral character, will be either true or false. We introduce a variable μ which takes a value of 0 if the maxim is false and 1 if the maxim is true (μ will be used to specify the players' prior beliefs in the next section). 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 maxim.

3.2 Representing Moral Codes and Beliefs

Given the state space defined above, we can represent a maxim or 'moral code' by a subset of states which are consistent with a particular statement linking people's actions and moral character. Consider, as an example, the following statement: "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 moral code, as described below:

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

where $\boldsymbol{\theta} = (\theta_1, .., \theta_n)$ and $\mathbf{c} = (c_1, .., c_n)$. In words, $\mathcal{M}_t$ includes only those states of the world in which, for each person i, either i has never engaged in act X or i has bad character. Note that, for any $t > 0$, if $\omega \in \Omega_t - \mathcal{M}_t$, the state is, by construction, inconsistent with the statements and, therefore, μ should only take a value of 0 in such a state. As such, we need to exclude from the subsequent analysis all states in Ω_t in which $\mu = 1$ but the history is inconsistent with the maxim. We do so by introducing a subset $\Phi_t \subset \Omega_t$ as follows:

$$\Phi_t = \mathcal{M}_t \cup \{(h_t, \boldsymbol{\theta}, \mathbf{c}, \mu) \in \Omega_t : \mu = 0\} \text{ for } t = 1, 2, .. \tag{5}$$

Thus, Φ_t includes all states in Ω_t in which either (i) the maxim is labelled as false ($\mu = 0$) or (ii) the actions, beliefs and character of individuals are consistent with the statements corresponding to $\mathcal{M}_t$. Thus, Φ_t is the set of feasible states in each period $t > 0$ and the subsequent analysis is restricted to these sets.

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

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

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 μ, these probability functions also enable us to specify their prior beliefs regarding the maxim, i.e. before any actions have occurred in the game.

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: $\mathcal{B}_t^i : \Phi_t \to 2^{\Phi_t}$ 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

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

Player i receives 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 $h_t = (h_{t-1}, e_t, a_t)$ 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 $\hat{\sigma}_t(a_t|\omega_{t-1}, e_t, \sigma)$ that the action a_t will take place after state ω_{t-1} and the move by nature e_t 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:

$$p_{\theta_i,t}^i(\omega_t|h_t) = \frac{p_{\theta_i,t-1}^i(\omega_{t-1}|h_{t-1})\,\hat{\sigma}_t(a_t|\omega_{t-1}, e_t, \sigma)}{\sum_{\omega'_{t-1} \in \Phi_{t-1}} p_{\theta_i,t-1}^i(\omega'_{t-1}|h_{t-1})\,\hat{\sigma}_t(a_t|\omega'_{t-1}, e_t, \sigma)} \text{ if } \omega_t \subset \mathbf{E}(h_t) \tag{8}$$

$$p_{\theta_i,t}^i(\omega_t|h_t) = 0 \text{ if } \omega_t \subsetneq \mathbf{E}(h_t) \tag{9}$$

where $\mathbf{E}(h_t)$ denotes the event that history h_t has been realised. Thus, equations (8) and (9) 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 $p_{\theta_i,t-1}^i(.|h_{t-1})$ from the previous period.

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

$$\mathcal{B}_t^i(\omega_t) = \left\{\omega \in \Phi_t : p_{\theta_i,t}^i(\omega|h_t) > 0\right\} \tag{10}$$

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, $\sigma_t^i(h_{t-1}, e_w^i, \theta)$ denotes the probability that player i of type θ chooses $a_w^i = 1$ (i.e. chooses to engage in act 'X') 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 (3). We define an equilibrium as a strategy profile σ, prior beliefs $p_{\theta_i,0}^i(.)$ and posterior beliefs $p_{\theta_i,t}^i(.)$ such that

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

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

4 Analysis

The concepts and notation introduced in subsections 3.1-3.3 provides us with the means to represent a 'moral code' using the standard knowledge-belief framework, as well as allow

people to hold conflicting beliefs about the 'moral code'. The aim of this section is to show under what conditions such a moral code generates dynamics in social behaviour as well as the nature of these dynamics.

We assume that all individuals share the same prior beliefs regarding the moral code and the character of other members of society: $\Pr(\mu = 0) = \delta$; i.e. they assign a probability of δ to the event that the moral code is false.

There are three conditions which are key for the following analysis. The first is the standard Folk Theorem condition:

$$W \leq \frac{\beta(n-1)\delta_o}{1-\beta}\left[R(1-\varepsilon)\right] \tag{11}$$

On the left-hand side of the condition in (11), we have the immediate gain in utility to an individual from engaging in act 'X' in any given period. On the right-hand side of the condition, we have the expected utility loss to the individual from being ostracised by the rest of society in subsequent periods, assuming that the probability of encountering a person of bad character in this society is equal to ε. If the condition in (11) does *not* hold, then a restriction or taboo against act 'X' cannot be sustained by the threat of social exclusion.

Next, we assume that, when the beliefs about a person's character corresponds to the prior beliefs, the cost of ostracising this person is too high for a threat of ostracism to be credible. The utility forgone from engaging in ostracism against such a person is $R(1-\varepsilon)$. On the other hand, the minmax punishment that can be inflicted on an individual is for the rest of the society to inflict perpetual ostracism, which would result in an expected utility loss of $\frac{\beta(n-1)\delta_o}{1-\beta}\left[R(1-\varepsilon)\right]$. Therefore, it is impossible to sustain social ostracism against someone whose probability of good character corresponds to prior beliefs if and only if

$$\begin{aligned} R(1-\varepsilon) &> \frac{\beta(n-1)\delta_o}{1-\beta}\left[R(1-\varepsilon)\right] \\ \Longrightarrow 1-\beta &> \beta(n-1)\delta_o \end{aligned} \tag{12}$$

The condition in (12) rules out the possibility that a person of 'good' reputation (i.e. probability of bad character equal to $1-\varepsilon$) is ever ostracised in equilibrium.

Following the reasoning provided in Section 2, if an individual i observes j engage in act 'X' then i's posterior beliefs about person j's character is given by the expression $\frac{(1-\varepsilon)\delta}{\varepsilon+(1-\varepsilon)\delta}$. If event e_{ij} occurs subsequently, the utility gain to i from associating with j is equal to $R\left(2\frac{\delta(1-\varepsilon)}{\delta(1-\varepsilon)+\varepsilon}-1\right)$. And it follows that ostracism will take place if and only if

$$\delta < \frac{\varepsilon}{(1-\varepsilon)} \tag{13}$$

If the condition in (13) holds, then person i derives utility from ostracising any person j who has engaged in act 'X'. Therefore, the threat of perpetual ostracism against anyone who has engaged in act 'X' is, in fact, credible.

Under these three conditions, we obtain the equilibrium described in the following proposition.

Proposition 1 *If the conditions in (11)-(13) hold, then there exists a unique equilibrium as follows: Individuals do not engage in act 'X', and individuals engage in ostracism if and only if they are paired with someone who has previously engaged in act 'X'.*

4.1 Heterogeneity in Player Types

As per Proposition 1, a social taboo against act 'X' is perpetually sustained under conditions (11)-(13) and we observe no dynamics in behaviour. In this section, we discuss two possible types of heterogeneity in the characteristics of individuals which cause behaviour regarding act 'X' to evolve over time. Let us denote by ε_i and W_i the prior belief regarding the character of person i and the utility gain to person i from engaging in act 'X' in some period.

Heterogeneity in Utility Gain from 'X': We start with the case where $\varepsilon_i = \varepsilon$ for all i, but W_i is heterogeneous and its distribution is described by the c.d.f. $F(.)$. Suppose also that $F(\underline{W}) \in (0,1)$ where $\underline{W}$ corresponds to the value for which the condition in (11) is satisfied with equality. Furthermore, suppose the conditions in (12) and (13) are satisfied.

By construction, if $W_i > \underline{W}$, then i would engage in act 'X' whenever he or she has the opportunity to do so. This is because the utility derived from engaging in act 'X' exceeds the maximum punishment that can be inflicted on i. Whenever i engages in act 'X', this will result in the updating of beliefs, and the posterior probability about the good character of i will be lower than the prior probability. As (13) is assumed to hold, person i will be subject to ostracism. However, the posterior probability of the truth of the moral code will also decline on each occasion that a new individual engages in act 'X'. We can show that after a finite number of such occurences, the condition in (13) will fail to hold for the updated beliefs regarding the truth of the moral code. In subsequent periods, all individuals in the society will engage in the moral code and they will not face ostracism. Formally, we have the following result.

Proposition 2 *Suppose that $\varepsilon_i = \varepsilon$ for all i, and the distribution of W_i is described by the c.d.f. $F(.)$ with $F(\underline{W}) \in (0,1)$, where $\underline{W}$ is the value of W for which the condition in (11) holds with equality. Suppose also that the conditions in (12) and (13) hold. Then individuals with $W_i > \underline{W}$ will engage in act 'X' in each period that event e_w^i occurs and be subject to ostracism in subsequent periods. The probability of anyone engaging in act 'X' will remain constant till the number of distinct individuals who engage in the act reach a finite number $l > 0$. Thereafter, individuals with $W_i \in (0, \underline{W}]$ will engage in act 'X' in each period that event e_w^i occurs and will not be subject to ostracism in subsequent periods. Individuals who have not engaged in act 'X' in preceding periods will not be ostracised.*

Proposition 2 describes a type of behaviour in which social norm is initially stable – in the sense that it is obeyed and violated by fixed proportions of individuals. This stability occurs although belief in the underlying moral code is weakening over time. When the moral code becomes sufficiently weak, the social norm unravels suddenly. And, in subsequent periods, everyone engages in act 'X' whenever they have the opportunity to do so.

Heterogeneity in Initial Reputation: Next, let us consider the case where W is homogeneous, while ε is distributed in the population according to the c.d.f. $G(.)$. We assume that (11) and (12) are satisfied, and (13) is satisfied with equality for some $\underline{\varepsilon}$ such that $G(\underline{\varepsilon}) \in (0,1)$.

By construction, if $\varepsilon_i < \underline{\varepsilon}$, then i would engage in act 'X' whenever he or she has the opportunity to do so. This is because i's ex-ante reputation of good character is sufficiently strong that he or she would not face ostracism even after engaging in act 'X'. After each occurance of this kind, the posterior probability that the moral code is true will decline. When the posterior probability declines, additional people will be able to engage in act 'X' without being ostracised. Therefore, the probability that act 'X' is violated will increase over time. But unlike the case of heteogeneous W, this increase will be gradual rather than abrupt. Formally, we have the following result.

Proposition 3 *Suppose that $W_i = W > 0$ for all i, and the distribution of ε_i is described by the c.d.f. $G(.)$ with $G(\underline{\varepsilon}) \in (0,1)$, where $\underline{\varepsilon}$ is the value of ε for which the condition in (13) holds with equality. Suppose also that the conditions in (11) and (12) hold. Then individuals with $\varepsilon_i < \underline{\varepsilon}$ will engage in act 'X' in each period that event e_w^i occurs and will not be subject to ostracism in subsequent periods. After each such occurence, the probability that an individual engages in act 'X' will (weakly) increase. For any individual i, the probability that he or she will engage in act 'X' – conditional on the occurence of event e_w^i – is (weakly) increasing over time and (weakly) increasing in the reputation of individual i at the start of the game. Ostracism will not occur in equilibrium.*

Proposition 2 describes a type of behaviour in which the social norm evolves gradually over time. It is first violated by those who initially have the strongest reputation in society. And because of their strong reputation, they are able to do so without 'losing their friends' (i.e. without being subject to ostracism). Their actions cause belief in the underlying moral code to weaken over time, allowing others to follow their example. Unlike the previous case, we do not observe a sudden unravelling of the social norm. But the evolution of the social norm is apparent both in terms of changes in beliefs and changes in behaviour.

4.2 Emergence of New Moral Codes

The analysis in the preceding section focused on dynamics in behaviour driven by declining beliefs in a moral code. It leaves open the question how a new moral code may arise, evolve

and shape behaviour.

There is no scope for an entirely new moral code to arise endogenously within the theoretical framework developed in Section 3. However, we can show that if there is an alternative moral code to which individuals initially assign some positive probability, then as beliefs in the dominant moral code declines (because of the updating of beliefs based on observed behaviour, as discussed in the preceding section), beliefs in the alternative code may strengthen. Furthermore, we can provide conditions under which the strengthening of beliefs in the second moral code can, in turn, influence subsequent behaviour.

To fix ideas, consider the alternative moral code "A person who has previously engaged in act X and has never declined to do so is of good moral character". Formally, we can represent this moral code by the subset of states in Ω_t that are consistent with it in each period t, as below:

$$\Phi'_t = \left\{(h_t, \boldsymbol{\theta}, \mathbf{c}) \in \Omega_t : \text{for each } i \in \mathcal{I},\ \left(\left(e^i_w, 1\right) \subsetneq h_t\right) \text{ or } \left(\left(e^i_w, 0\right) \subseteq h_t\right) \text{ or } (c_i = 1)\ \right\} \quad (14)$$

Note that if $(e^i_w, 1) \subseteq h_t$ and $(e^i_w, 0) \subsetneq h_t$ for some $i \in \mathcal{I}$, then the sets $\Phi_t \cap \mathtt{E}(h_t)$ and $\Phi'_t \cap \mathtt{E}(h_t)$ are mutually exclusive.[2] In other words, the two moral codes would contradict one another; the first moral code – represented by Φ_t, defined in (5) – implies i has bad moral character while the second moral code – represented by Φ'_t , defined in (14) – implies that i has good moral character.

Such a contradiction can arise if the two moral codes have different origins. For example, the first may stem from traditional beliefs while the second may be declared by a new religious, legal or moral authority (see Posner 1998 and Ellickson 2001 for related discussions on the emergence of new social norms).

As before, the binary variable μ indicates whether the moral code represented by (5) is true or false. We use μ' to indicate whether the second moral code represented by (14) is true ($\mu' = 1$) or false ($\mu' = 0$) and represent prior beliefs regarding this moral code by $\Pr(\mu = 0) = \delta'$. As the two moral codes cannot be simultaneously true, we must have $1 - \delta' \leq \delta$.

The presence of a second moral code does not affect the Folk theorem condition (11) or the condition that ensures that individuals with no prior history are ostracised, (12). However, the posterior beliefs regarding the moral character of a person who engages in act 'X' becomes $\frac{(1-\varepsilon)\delta}{\varepsilon\delta' + (1-\varepsilon)\delta}$. Therefore, the condition under which individuals who have previously engaged in act 'X' are ostracised is given by

$$\delta < \frac{\varepsilon\delta'}{1-\varepsilon} \quad (15)$$

[2] $\mathtt{E}(h_t)$ represents the subset of states in which history h_t has occured. It is distinct from the expectations operator $E(.)$.

If belief in the second moral code is initially very weak – i.e. δ' is close to 1 – then the reasoning behind Proposition 1 would still apply. Specifically, in the case of homogeneous individuals, if conditions (11), (12) and (15) hold, we obtain an equilibrium in which no individual engages in act 'X' and no-one is ostracised.

Corollary 1 *to Proposition 1: Suppose that there are two moral codes, as described by (5) and (14), with initial beliefs given, respectively, by* $\delta, \delta' \in (0,1)$. *If individuals are homogeneous and the conditions in (11), (12) and (15) hold, then, for* δ' *sufficiently close to* 1, *there is a unique equilibrium in which individuals do not engage in act 'X', and individuals engage in ostracism if and only if they are paired with someone who has previously engaged in act 'X'.*

The key difference between Proposition 1 and its corollary is due to the difference between conditions 13 and 15. If $\delta' < 1$ – i.e. individuals assign some positive probability to the alternative moral code – then 15 is a stricter condition. In other words, the presence of the alternative moral code makes it more difficult to sustain a social taboo against act 'X'.

Next, we consider the case where there is heterogeneity in both the utility gain from engaging in 'X' and the initial reputation of individuals as described by the c.d.f.'s $F(.)$ and $G(.)$. Then the condition in (15) will not hold for individuals with high values of W_i and low values of ε_i. These individuals will choose to engage in 'X' and this will lead to an updating of beliefs regarding *both* moral codes. Specifically, if an individual i is observed to engage in 'X' in period t, and the full history is described by h_t, we obtain

$$\begin{aligned}
\Pr(\mu' = 1|h_t) &= 1 - \delta'_t = \frac{\left(1-\delta'_{t-1}\right)(1-\varepsilon_i)}{1-(1-\delta_{t-1})(1-\varepsilon_i)-\left(1-\delta'_{t-1}\right)\varepsilon_i} \\
\Pr(\mu = 1|h_t) &= 1 - \delta_t = \frac{(1-\delta_{t-1})(1-(1-\varepsilon_i))}{1-(1-\delta_{t-1})(1-\varepsilon_i)-\left(1-\delta'_{t-1}\right)\varepsilon_i}
\end{aligned}$$

We can show that, for ε_i sufficiently close to zero, we obtain $\delta_t > \delta_{t-1}$ and $\delta'_t < \delta_{t-1}$. In words, when an individual, with no prior history, is observed to engage in 'X', this leads to a decline in beliefs in the first moral code and a strengthening of beliefs in the alternative moral code, assuming prior belief in the good moral character of the individual is sufficiently strong.

Beliefs regarding the moral character of the individual who has engaged in act 'X' is updated as follows:

$$\Pr(c_i = 1|h_t) = \frac{\delta_{t-1}(1-\varepsilon_i)}{\delta_{t-1}(1-\varepsilon_i)+\delta'_{t-1}\varepsilon_i} \tag{16}$$

Using (16), we can show that if person i chooses to engage in act 'X', her reputation may worsen, as in section 4.1, or *improve* depending on the strength of beliefs in the two opposing moral codes Specfically, if $\delta'_{t-1} > \delta_{t-1}$ then act 'X' is damaging to her reputation while if

$\delta'_{t-1} < \delta_{t-1}$, this improves her reputation. And i's reputation is unaffected if $\delta'_{t-1} = \delta_{t-1}$. Therefore, if the second code replaces the first as the dominant moral code in some period t (i.e. $\delta'_t < \delta_t$), individuals with no prior history have an added incentive to engage in 'X' in subsequent periods because it improves their reputation. In particular, individuals with $W_i < 0$ and $\varepsilon_i > 0.5$ may opt to engage in 'X' if, by doing so, they are able to improve their reputation sufficiently to escape ostracism in subsequent periods.

Proposition 4 *Suppose that there are two moral codes, as described by (5) and (14), with initial beliefs given, respectively, by $\delta, \delta' \in (0,1)$, $\delta < \delta'$, the distribution of W_i is described by the c.d.f. $F(.)$, $F(\underline{W}) \in (0,1)$ and $\varepsilon_i = \varepsilon > \underline{\varepsilon}$ for all i; where $\underline{W}$ and $\underline{\varepsilon}$ are the values for which the conditions in (11) and (15), respectively, hold with equality. Then individuals with $W_i \geq \underline{W}$ will engage in act 'X' in each period that event e^i_w occurs and be subject to ostracism in subsequent periods. When the number of distinct individuals who engage in the act reach a finite number $l_1 > 0$, individuals with $W_i \in [0, \underline{W})$ will also engage in act 'X' in each period that event e^i_w occurs and no ostracism will take place in subsequent periods. When the number of distinct individuals who engage in the act reach a finite number $l_2 > l_1$, individuals with $W_i \in (-\infty, 0)$ will also engage in act 'X' in each period that event e^i_w occurs.*

5 Application: Female Labour Market Participation

In this section, we present an application of the theory on changing social norms to the phenomenon of low female labour market participation in Bangladesh. Traditionally, women's economic role in rural Bangladesh has been limited to activities that can be carried out within the household, for example the rearing of livestock and work on the family farm. Women's employment activities appear to be constrained by two sets of social norms: first, the practice of 'purdah' which restrict the presence of women in public spaces (Paul 1992, White 1992); and second, the traditional division of labour by gender, which assigns men the role of breadwinner, and women responsibility for domestic work (Amin 1997, Kabeer 2001).

During the 1980's and 1990's, two market-related phenomena began to change the scope of women's participation in economic activities, particularly in rural areas. First, the growth of microfinance, with loan-products targeted at women, gave them direct access to credit from the market, and increased their scope for engaging in smale-scale entrepreneurial activities. Significantly, these entrepreneurial activities could be carried out from within the household, with assistance from male household members in marketing the produce, and thus without violating the social norm against women's direct participation in market-activities (Kabeer 1998, 2001).

Second, the same period saw the emergence and growth of the export-oriented ready-made garments (RMG) sector which employed large numbers of women. In 2014, about 4

million workers were employed in this sector, growing from just 40,000 in 1983[3] and 80% of the workforce is female (Khatun et al. 2007). The sector presently accounts for 79% of exports and 14% of GDP for Bangladesh (Bangladesh Bureau of Statistics 2013). According to Heath and Mobarak (2015), women employed in the RMG sector earn 13.65% more compared to those with the same education and experience in other industries. Unlike the entrepreneurial activities stimulated by microfinance loans, participation in the RMG sector would require women to step out of the home, and go into factories, mostly located in urban areas – a routine which is contrary to social norms regarding their presence in public spaces and engagement in market-related activities.

Despite the large numbers of women employed in the RMG sector in Bangladesh today, women's employment outside of the home remains very low – 10% according to the 2005 Household Income and Expenditures Survey. Recent qualitative evidence shows that there is a strong social stigma associated with female employment in this sector. Asadullah (2014) conducted interviews with women resident in two sites in Bangladesh with the highest concentration of RMG factories – in the districts of Gazipur and Narayanganj. The pattern which emerged from these interviews is that although women from impoverished backgrounds find the wages available in the RMG sector attractive, they also associate such employment with loss in social standing and would avoid it as much as possible. We present three cases below to illustrate these attitudes:

> Case 1: "Rozina didn't study beyond grade 4 becase her mother was unwell and she had to provide support at home... She had managed an [RMG] job within 1 month of arriving in Narayanganj (an area with a high concentration of RMG factories). However after her marriage, her husband stopped her working as he disliked any outside work. This is despite the fact that Rozina used to enjoy RMG work. If her husband allows, she is keen to resume work ... I asked her plans about her daughter and the response was '*I will educate her as far as God grants ... [If possible] I will find work myself but I will not make her do any work.*'" .
>
> Case 2: "Mukta is a local resident of Narayanganj. Her father was in RMG work but left because of health problems. Then he put her into RMG work. So she dropped out of school after grade 5 and subsequently worked for 5-6 years. There she met [her husband] Shumon ... [After their marriage] Shumon started objecting to her work even though they were in the same factory. Her father-in-law also opposed her work. So she left work soon after marriage. I asked her to reflect on that decision and she said: '*I will go again if I really need the money; everyone who works there do it because they badly need the money.*'" .

[3]Figures provided by the Bangladesh Garment Manufacturers and Exporters Association at http://www.bgmea.com.bd/

Case 3: "Then I asked about her about her sister-in-law, Moni, who was present at that time. She used to work as a private tutor for a grade 4 student. The student's mother then fixed her the job of a kindergarten school teacher where she teaches in grades 2-5 ... Her salary is less than that of an operator in the RMG sector. I asked why she didn't consider taking up the better-paid RMG job. She said it didn't have social respect." (Asadullah 2014).

In the first two cases, we find examples of men reluctant to have their wives/daughter-in-law engage in factory work even when they themselves hold a similar type of employment. This illustrates the social stigma associated with female factory work, extending to their immediate families. We also find women concurring with the view that such work leads to a loss of social standing for women – Moni makes this point explicitly and Rozina is adamantly opposed to factory work for her daughter although she found her own experience of such work enjoyable and is keen to resume it. Particularly telling is the second interview, where Mukta – who left her work in an RMG factory because of opposition by her husband and father-in-law – regards it as a last resort for earning money. This qualitative evidence raises the question why there remains a social stigma regarding women's employment in RMG factories, despite the fact that the sector has been employing large numbers of women over a long period of time.

5.1 Explaining the Phenomenon of Low Participation

Here, we provide an explanation to this question, as well as insights about how social norms may evolve in the future, using the theory developed earlier in this paper. Let us assume that there is a maxim that "Women who take up factory jobs are of bad character" and, initially, there is strong belief in this maxim; i.e. δ is small. Act 'X' is the decision to take up work in the RMG sector, and e_w^i is the event that an opportunity for such employment become available. For simplicity, we model only the decisions of female agents in the population. The act of 'associating' and choosing not to 'associate' with other agents, as modelled, can be regarded as a stylised representation of social interactions in a traditional community.

During the 1990's, women who typically chose to work in the RMG sector were divorced, abandoned, or had fled their marital homes because of domestic violence (Kabeer, 2000). In traditional Bangladeshi society, these women would be social outcastes, considered to be of dubious moral character. In terms of the model, they would have a high initial value of ε. If social outcastes take up factory jobs, this would have little effect on δ, since their ε is large relative to δ. Therefore, the norm where the average woman typically avoid taking up these jobs would persist.

In recent years, the profile of female workers in the RMG sector appears to have changed. In Table 1, we present descriptive statistics on women employed in the RMG sector and

in the wider population using data from the 2014 Bangladesh Women's Life Choices and Attitudes Survey (WiLCAS). The WiLCAS includes a nationally representative sample of women aged 20-39 years, as well as an over-sampling of women in the urban areas with the highest concentration of RMG factories. The RMG workers are younger and less likely to be married. They are also more much more likely to be divorced or separated from their husbands than the typical women in the sample but the majority of them – nearly 70% – are not. They also have, on average, similar levels of schooling to women not employed in the RMG sector. A notable difference is found in their parental background: their fathers had less land, had less schooling and were more likely to be an artisan or day-labourer (low-paying occupations), with all these differences being statistically significant.

This evidence suggests that the primary driver for women's entry into RMG work today is poverty and economic hardship. In terms of the theoretical model presented above, this would correspond to a situation where the gains from act 'X' is heterogeneous, with the poorest women gaining the most. But if poverty, per se, is not believed to be correlated with moral character, then women who opt to work in the RMG sector should, initially, have a relatively low ε. Their choice of work would cause their ε to decline and they will be subject to social ostracism. But, if ε is initially low relative to δ, then δ should start to increase. In other words, as more of these women opt for employment in RMG factories, the belief in the maxim that "women who take up factory jobs are of bad character" in the general population should weaken.

As per Proposition 2, the decline in beliefs in the maxim should initially have no impact on the *propensity* of women to take up employment in the RMG sector or the tendency to ostracise those who do. But when a sufficiently large number of poor women have been *seen* to take up work in RMG factories, the social stigma against women who opt to work in the RMG sector would unravel and the propensity to do so should increase suddenly.

6 Application: School Enrollment of Girls

In this section, we consider the phenomenon of rising female school enrollment in Bangladesh to present a second application of the theoretical model. Historically, girls were much less likely to enroll in primary or secondary school than boys. Given the expectation that a daughter would marry early and that her primary responsibilities would involve childcare and household work, her schooling was deemed less important than that of a son (White, 1992). More widely, in South Asia, female schooling was constrained by concerns with the management of female sexuality, which meant girls would drop-out of school at the time of puberty; to the extent that they received any education, the emphasis was "on domestic skills and feminine social accomplishments" (Dube, 1997).

The literature indicates that attitudes began to shift during the 1980's. This is a period

which saw a number of government initiatives to lower costs and improve access to schooling for girls (Schurmann 2009, Asadullah and Chaudhury 2009), as well as improved labour market opportunities as discussed in the preceding section. But the evidence suggests that, in rural areas, this change in perception went beyond the perceived economic benefits of female schooling. Programmes to increase the number of female school teachers provided alternative role models for girls in rural communities where female education had been lacking (Schuler 2007). Rural parents observed examples of educated women in the middle-class, whose manners and lifestyle signalled good upbring. In other words, education was evolving into a signal of good character whereas, in the past, it had been the opposite. And this changing perception potentially accelerated the increase in female school enrollment.

However, this reasoning raises the question why, in the case of schooling, educated women were regarded as positive role models which accelerated the process whereas, in the case of market employment, working women were not. We can apply the theoretical model to obtain a precise answer to this question. Both in the case of employment and schooling, there was heterogeneity in the gains from making the choice in question. But in the case of employment, the first women to opt for work were social outcastes, whereas the visible instances of educated women belonged to the middle-class, whom the rural population looked up to. Therefore, female labour market participation did not change the moral code regarding female employment, while female schooling did.

7 Conclusion

In this paper, we develop a theory of changing social norms which combines elements of both the sociological and economic approaches to the subject. We formally model a 'moral code' or maxim that provides a mapping from the behaviour of individuals to their 'character', and people derive utility from ostracising those they believe to have 'bad character'. The threat of ostracism discourages individuals from engaging in certain type of actions that, according to the moral code, would reflect badly on them. The key mechanism for the evolution of social norms in the model is that the individual actions lead others to not only update their beliefs about the character of these individuals, but also lead them to update their beliefs regarding the (accuracy of the) 'moral code'. Consequently, the threat of ostracism against the actions in question changes over time.

The theory provides a coherent framework for thinking about how social taboos against certain types of behaviour may be sustained over time, and how they may evolve. The model generates its own dynamics, even in the absence of exogenous shocks or changes to the environment and we show that it can generate both sudden and gradual changes in social norms.

In the second part of the paper, we apply the model to explain the phenomenon of

persistent low labour market participation of women in Bangladesh. The existing evidence suggests that there is a social stigma against women's work in the manufacturing sector despite the dramatic growth,in the last three decades, of the ready-made garments sector, which employs 4 million women in the country today. The model suggests reasons why the social stigma against female employment outside of the home has persisted in spite of these changes. Nevertheless, it argues that the present conditions (and employment patterns) are ripe for an evolution in the social norms in the near future.

The theoretical framework developed in this paper can provide a systematic way for analysing evolving social norms in a variety of different contexts.

8 Theoretical Appendix

Proof. of Proposition 1: If a person j has not previously engaged in act 'X', then beliefs regarding j's character will correspond to prior beliefs. If condition (12) holds, j will not be ostracised by others. If, given history h_t, j has engaged in act 'X' in some period $\tau < t$, then, it follows from the reasoning in Section 2, that beliefs regarding j's character is given by

$$\Pr\left(c_j = 1 | \left(e_w^j, 1\right) \in h_t\right) = \frac{(1-\varepsilon)\,\delta}{\varepsilon + (1-\varepsilon)\,\delta}$$

Then, under condition (13), j will be ostracised in subsequent periods whenever another individual has the opportunity to do so. It follows that, if the condition in (11) holds, then no-one will engage in act 'X' when they have an opportunity to do so. Thus, under conditions (11)-(12), we have a unique prediction of each person's actions following each possible history as follows: no-one engages in act 'X' when they have the opportunity to do so, and a person is ostracised by others if and only if they have previously engaged in act 'X'. ■

Proof. of Proposition 2: Suppose that $W_j > \underline{W}$. Then (11) does not hold for j. Therefore, j will engage in act 'X' in each period that $e_w^j = 1$. Suppose that such an opportunity arises for the first time in period t following history h_t. As per the reasoning in Section 2, this will lead to Bayesian updating of belief in the moral code as follows:

$$\Pr\left(\mu = 1 | h_t\right) = 1 - \delta_t = \frac{\varepsilon\left(1 - \delta_{t-1}\right)}{\delta_{t-1} + \varepsilon\left(1 - \delta_{t-1}\right)}$$

where δ_t denotes belief in the moral code in period t. If $\delta_{t-1} > 0$ and $\varepsilon < 1$, then $1 - \delta_t < 1 - \delta_{t-1}$, i.e. $\delta_t > \delta_{t-1}$. If the condition in (13) is satisfied for δ_t, then j will be subject to ostracism in subsequent periods.

If there are l such individuals who have engaged in act 'X' in history h_t, we have

$$\Pr\left(\mu = 1 | h_t\right) = 1 - \delta_t = \frac{\varepsilon^l\left(1 - \delta\right)}{\delta + \varepsilon^l\left(1 - \delta\right)}$$

and

$$\lim_{l \longrightarrow \infty} \Pr(\mu = 1 | h_t) = 0$$

It follows that for l sufficiently large, $\Pr(\mu = 1|h_t) < \frac{\varepsilon}{(1-\varepsilon)}$; i.e. the updated beliefs regarding the moral code following history h_t would not satisfy the condition in (13). Therefore, individuals who engage in act 'X' after history h_t has occurred will not face ostracism. Therefore, any individual j for whom $W_j > 0$ will engage in act 'X' whenever event e_w^j occurs following history h_t.

While $\Pr(\mu = 1|h_t) \geq \frac{\varepsilon}{(1-\varepsilon)}$, an individual j will engage in act 'X' if and only if $W_j > \underline{W}$. Therefore, the probability of anyone engaging in act 'X' in period t is given by

$$\Pr\left(h_t = \left(h_{t-1}, e_w^j\right), W_j > \underline{W}\right) = n\delta_w \left\{1 - F\left(\underline{W}\right)\right\}$$

which is constant. ■

Proof. of Proposition 3: Consider some history h_{t-1} in which no-one has engaged in act 'X' and suppose that event e_w^i occurs in period t where $\varepsilon_i < \underline{\varepsilon}$. Then (13) does not hold for i. Therefore, i will not be ostracised for engaging in act 'X'. Therefore, i will engage in act 'X'. As per the reasoning in Section 2, this will lead to Bayesian updating of belief in the moral code as follows:

$$\Pr(\mu = 1|h_t) = 1 - \delta_t(h_t) = \frac{\varepsilon_i(1-\delta)}{\delta + \varepsilon_i(1-\delta)}$$

where $\delta_t(h_t)$ denotes belief in the moral code in period t following history h_t. If $\delta > 0$ and $\varepsilon_i < 1$, then $1 - \delta_t(h_t) < 1 - \delta$ i.e. $\delta_t(h_t) > \delta$. Following the same reasoning, we see that $\delta_t(h_t)$ increases each time that an individual engages in act 'X'. Therefore, given a sequence $\{h_t\}_{t=1}^T$ where each h_t denotes the first t-period history in h_T, we have $\delta_t(h_t)$ weakly increasing over time. Therefore, the subset of individuals for whom the equivalent of (13) does not hold – i.e. $\left\{j \in \mathcal{I} : \delta_t(h_t) \geq \frac{\varepsilon_j}{(1-\varepsilon_j)}\right\}$ – and the probability that an individual engages in act 'X', are also weakly increasing over time.

For each j such that $\varepsilon_j \geq \underline{\varepsilon}$, there is a threshold value $\underline{\delta}_j \in (\delta, 1)$ such that (13) does not hold for $\delta \geq \underline{\delta}_j$. Since, $\delta_t(h_t)$ is weakly increasing in t for each given sequence $\{h_t\}_{t=1}^T$, it follows that the unconditional probability $\Pr\left(\delta_t(h_t) \geq \underline{\delta}_j\right)$ is weakly increasing in t. Therefore, the probability that j engages in act 'X', conditional on the occurrence of event e_w^j in period t, is increasing in t.

If k has better initial reputation than j – i.e. $\varepsilon_j > \varepsilon_k$ – then $\underline{\delta}_k < \underline{\delta}_j$. Therefore, $\Pr\left(\delta_t(h_t) \geq \underline{\delta}_k\right) > \Pr\left(\delta_t(h_t) \geq \underline{\delta}_j\right)$. Therefore, the probability that k engages in act 'X', conditional on the occurrence of event e_w^j in period t, is higher than that for j. In other words, the probability that an individual engages in act 'X' is increasing in his or her initial reputation.

Since (11) is assumed to hold, no individual will engage in act 'X' if this leads to ostracism in subsequent periods. As reasoned above, an individual engages in act 'X' when belief in the

moral code – as represented by $\delta_t(h_t)$ – is sufficiently weak for him/her to escape punishment. Therefore, ostracism does not occur in equilibrium. ■

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] Asadullah, M.N. (2014) Fieldnotes for Bangladesh Women's Life Choices and Attitudes Survey. mimeo, University of Kent.

[3] Bangladesh Bureau of Statisics (2013). *2012 Statistical Yearbook.* Statistics and Information Division. Ministry of Planning. Government of Bangladesh.

[4] Bénabou, Roland, and Jean Tirole (2006). "Incentives and Prosocial Behavior." *The American Economic Review,* Volume 96(5), pp. 1652-1678.

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

[6] Bernheim, B. (1994) "A Theory of Confirmity", *Journal of Political Economy*, Vol. 102(4), pp. 841-877.

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

[8] Bowles, S., and H. Gintis (2011). *A Cooperative Species: Human Reciprocity and its Evolution*, Princeton University Press, 2011.

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

[10] Ellickson, R.C. (1998). "The Market for Social Norms", *American Law and Economics Review*, Volume 3(1), pp. 1-49.

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

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

[13] Fehr, E. and K. M. Schmidt (2006). "The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories", *Handbook of the Economics of Giving, Altruism and Reciprocity*, Vol. 1, edited by Serge-Christophe. Kolm, Jean Mercier Ythier, Elsevier.

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

[15] Heath, R., & Mobarak, A. M. (2015). "Manufacturing growth and the lives of Bangladeshi women", *Journal of Development Economics*, Volume 115, pages 1-15.

[16] Kabeer, Naila (1998). "'Money can't buy me love'? Re-evaluating gender, credit and empowerment in rural Bangladesh". *Discussion Paper-Institute of Development Studies*, University of Sussex.

[17] Kabeer, Naila (2000). *The power to choose. Bangladeshi women and labour market decisions in London and Dhaka*, London.

[18] Kabeer, Naila. (2001). "Conflicts over credit: re-evaluating the empowerment potential of loans to women in rural Bangladesh", *World Development*, Volumne 29(1), pages 63-84.

[19] Kazianga, Harounan and Zaki Wahhaj (2013) "Gender, Social Norms, and Household Production in Burkina Faso", *Economic Development and Cultural Change*, Vol. 61, No. 3, pp. 539-576.

[20] Khatun, Fahmida, Rahman, Mustafizur, Bhattacharya, Debapriya, Moazzem, Khondker G. & Shahrin, Afifa. (2007). *Gender and Trade Liberalization in Bangladesh: The Case of Ready-made Garments*. USAID.

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

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

[23] Paul, Bimal K. (1992). "Female activity space in rural Bangladesh", *Geographical Review*, Voume 82(1), pages 1-12..

[24] Roland, Gerard (2004). "Fast-moving and Slow-moving Institutions". *Studies in Comparative International Development*, Volume 38, No. 4, pp. 109-131.

[25] Wahhaj, Zaki (2012). "Social Norms, Higher Order Beliefs and the Emperor's New Clothes", *ThReD Working Paper 2012-023.*

[26] White, S. (1992). *Arguing with the Crocodile: Gender and Class in Bangladesh*, Zed Books.