"A Game-Theoretic Framework to Study the Influence of Globalisation on Social Norms of Co-operation"

G Grimalda

CSGR Working Paper No. 151/04

November 2004

Centre for the Study of Globalisation and Regionalisation (CSGR), University of Warwick, Coventry, CV4 7AL, United Kingdom. URL: http://www.csgr.org

A Game-Theoretic Framework to Study the Influence of Globalisation on Social Norms of Co-operation

Gianluca Grimalda[1]
CSGR, University of Warwick
CSGR Working Paper No. 151/04
November 2004

Abstract

A game-theoretic framework is developed to study the evolution of social norms in a society. The two main theoretical assumptions underpinning the model are, first, that agents have some kind of "social" preferences, in addition to standard "self-interested" preferences. Second, individuals modify their behaviour over time in accordance to the "imitation of the most successful agent" paradigm. A stylised model of social interactions is developed, along with concepts of static and dynamic equilibria. After social preferences are specified in accordance with the normative expectation theory, an analysis of the type of equilibria in public goods interactions is provided. Finally, the impact on co-operation of a change in a society's modes of behaviour, which may be seen as a result of migration or the impact of global communication media, is studied.

Keywords: Co-operation, social norms, evolution

JEL Classification: H41, D02, C62

Address for correspondence:
Dr Gianluca Grimalda
CSGR, University of Warwick
Coventry, CV4 7AL, UK
Email: g.f.Grimalda@warwick.ac.uk

[1] **Acknowledgments**: I would like to thank Nancy Buchan for the extensive discussions that have led to the setting of the agenda behind this paper, and John Gerry for assistance with proof-reading. All errors and omissions are my sole responsibility.

1. INTRODUCTION

Concepts such as social capital, trust and co-operation are now seen as key resources for a socio-economic system to progress (for a champion of each of the above notions see Putnam (2000), Fukuyama (1996), and Taylor (1987), respectively). Although the unit of analysis of most of this research has generally been local communities or nation-states, the far-reaching process of globalisation has also pointed to the importance of trans-national cooperation. The provision of global public goods, i.e. those public goods that transcend national borders, such as the environment, international justice, and international financial stability (e.g. Kaul et al. (2003)), has thus been seen as crucial to the growth of global prosperity (see for instance the agenda set by the Millennium Development Goals). However, not only has globalisation brought to centre stage such a new form of cross-national co-operation, but also it has posed new challenges to the other – more traditional - forms of co-operation, that which takes place at the local level. Though the three concepts mentioned above are obviously linked with each other, the focus of this paper will in particular be on co-operation, and on the conditions whereby social norms favouring co-operation can become established and endure over time in a (global) society.

As for the local aspect of co-operation, an account of co-operation that has attracted consensus is that based on the notion of reciprocity (see e.g. Axelrod (1984) for the game-theoretic treatment of this notion). The underlying idea is that in those situations that can be characterised as 'social dilemmas' (see e.g. Hardin (1982)), i.e. those in which the predicaments of individual rationality part away from social rationality, a (nearly) universal co-operative outcome can all the same be upheld. For this to be the case, interactions need to be frequent and personalised, and some forms of punishment of deviant behaviour must be put in place. For if such conditions are satisfied, then the long-term benefits of abiding by the co-operative norm may outstrip the short-term incentives to 'free ride' on others' contribution. There may exist several ways in which this is possible, but the bottom line is that each individual anticipates that her selfish behaviour – once recognized and then reciprocated by other community members – will lead to the progressive disruption of the co-operative norm (for a formal treatment of this account, see e.g. Kandori, 1992). In game-theoretic parlance, this leads to the notion of a tit-for-tat equilibrium in a repeated Prisoner's Dilemma game.

On the grounds of this analysis, a strong argument may be put forward that the influence of globalisation on this type of co-operation is in fact negative (see e.g. North, 1990). In fact, it has to be noted that globalisation impinges upon the very nature of social relations, transforming what were personal, small-scale and frequent interactions within close-knit 'traditional' communities into anonymous, large-scale and rare social exchanges. A well-known effect of globalisation is in fact to enlarge the scale of interactions, that is, to increase the number of individuals involved in them. This is what has been referred to as the 'de-territorialisation' of social relationships by scholars of globalisation (Scholte (2000)). This aspect will of course make it more difficult to sustain a reciprocity-based system of co-operation, because the frequency of interactions among individuals will be reduced, interactions will become more impersonal – if not nearly anonymous – and the possibility of enforcing punishment of 'deviant' members will tend to disappear.

However, this argument may be countered by the consideration that an even more radical entrenchment to 'local' concepts of identity may be triggered as a direct reaction to the process of homogenisation that is seemingly associated with globalisation. Social identity theory from psychology suggests that identification with a group typically rests on the perceived existence of a 'stranger' to the group, that is, individuals or groups who have different social/cultural/economic characterisations then those of the group to which one feels to belong (e.g. Messick and Brewer (1983)). Applying this theory to the process of globalisation may suggest that by making the perception of the presence of a 'stranger' more vivid than before, the attachment to the group may actually increase, and thus lead to higher levels of in-group (or localized) trust and co-operation. The persistence of support for ethnically characterised political movements in several countries may be considered as evidence for this idea.

The influence of globalisation on co-operation within an international context is also ambiguous. On the one hand, globalisation widens the number of agents involved in the interaction, and this should generally act as a disincentive for co-operation, because the incentive to free ride on others' contribution is – at least for the most common settings – positively related to the number of players involved (see the seminal analysis by Olson

(1965); and also Kandori (1992))[2]. Moreover, the combination of different national identities may make the problem even more complicated, because of the lack of a common system of shared beliefs on mutual behaviour (see Ostrom (2003) for a discussion of the influence of shared norms on cooperation and trust) and the possibility of a diffused diffidence towards the foreigners in many countries (Barth (1995)). In contrast, a more optimistic view rests on the idea that the 'creolisation' of cultures (Hannerz (1992)) triggered by globalisation may be expected to reduce substantial cultural differences across countries and foster the recognition by individuals of a similar – or common – cultural framework for interpreting the environment. Furthermore, globalisation – almost by definition - makes interactions more frequent, and this should have a positive effect on co-operation, with the increased incentive to build a reputation as a "co-operator". In other words, the discount factor of future utility increases as an effect of the acceleration of the rate of encounters. The contrasting implications of these hypothesis makes rather difficult to predict the 'sign' of the influence of globalisation on global co-operation.

The arguments set out above should make it clear how globalisation may be a relevant factor in affecting social norms of co-operation, both at the local and the international level. It is the purpose of this paper to develop an analytical framework that makes it possible to study the evolution of social norms, as well as the result of a change in some of a society's structural factors, such as the composition of its population. Though the model that will be developed only represents a first building block for the study of the relationship between globalisation and social norms of co-operation, I believe that its generality will make it possible to receive several applications once suitable specifications are implemented.

The theoretical framework is based on two elements, that is, a model of individual choice, and a principle that engenders the evolution of individual action and social norms. As for the former, the theoretical framework that will be adopted in modelling individual choices is that of the so-called 'other-regarding' motivations (e.g. Ben Ner and Putterman (1998); Fehr and Schmidt (2001)). The underlying assumption is that individuals have some form of concern for the others when making decisions, which may either include an altruistic attitude to further the well-being of other people as well as their own, or the disposition to comply with others' expectations, or the propensity to reciprocate the intentions perceived in others'

[2] The economic experimental literature is however cautious on this point, as the impact of increasing the

behaviour. Though the motivations just listed may not appear too unexpected as guidance to human behaviour, it is notorious how the classical rational choice approach to individual action has tended to neglect most kind of motivations that extended beyond the self. The key unanswered question of this approach is obviously who the relevant 'others' are for an individual and thus the extension of the 'group' with whom the individual identifies. One can contrast two extreme hypotheses in this respect: The first is that the group of agents on which the subject bases her judgments is relatively 'local', i.e. it takes as the main reference the views, interests, and modes of assessment of the community to which the individual is physically close. The alternative hypothesis is instead that the 'others' to which a subject refers to is, in some sense, 'global', i.e. it is not constrained by geographical, or even cultural and socio-economic barriers. This latter hypothesis then leads to a model of individual where s/he possesses multiple identities, and these are created taking a national and/or global perspective (see Sen (1999)). The underlying idea of this paper is that the social identity of an individual is a crucial factor in determining her attitudes towards co-operation, and that globalisation may significantly impinge on the latter through reshaping an individual's perception of her social identity.

As for the second element of the framework, the dynamic of the model is driven by the so-called replicator dynamics. Though its original application has been in biology, the basic aim being to create a model for species evolution exposed to natural selection, this is now a popular tool of analysis in the social sciences, too. The main idea here is that individuals' objective function is a measure of their 'success' in the social environment, alike 'fitness' and ability to survive in a biological environment. The basic engine of social norms evolution is then the assumption that individuals desire to imitate the most successful agents in a society, though they are subject to limited information and bounded rationality. This leads to actions that are conducive to economic and 'social' success to spread with higher frequency among the population, thus causing social norms to evolve.

Section 2 develops a model of individual choice where a *comprehensive* utility function is broken down into *self-interested* and *other-regarding* utility. A model of social interaction is also put forward; in typical game-theoretic fashion, social interactions are seen as pairwise 'encounters' where two agents drawn at random from two different 'populations' of

---

numbers of players does not seem to change significantly the degree of co-operation (see Ledyard (1995)).

individuals are matched to play a 'game'. Section 3 provides a static and two dynamic notions of equilibria, which are adapted from the concept of Psychological Nash equilibrium. These will form the basic analytical tools of the study. Section 4 puts forward a particular specification of the other-regarding component of utility, which draws on the theory of normative expectations. Although it is not the purpose of this paper to comment on this particular theory, it will be offered as an example of how the different concepts of equilibria can be applied. Section 5 proposes a preliminary application of this framework to the study of the impact of globalisation on social norms. In particular, the change in the composition of the population, with one section of the population now bearing different attitudes to comply with social norms than the other, is studied. This type of change may be interpreted as the result of migratory forces, or as a change in individual mode of behaviour resulting from the 'exposure' to global media of communication. Section 6 concludes and puts forward possible developments of this line of enquiry.

2. INTERACTION BETWEEN MULTIPLE-MOTIVATIONS-BASED POPULATIONS

2.1 The Stage Game with Comprehensive Utility Functions

As customary in economic analysis of social interactions, I shall draw on the tools of game theory in order to give a formal representation of a general situation of interaction. The situation is structured so as to involve *pairs* of individuals at a time. The pair of individuals can best be thought of as having different *roles* in the interaction, which makes it possible to distinguish among different groups – or, in game-theoretic jargon, *populations* - of agents. Examples of roles may be gender, the direction from which two drivers approach a crossroads, or people's cultural/ethnic belonging. The framework may be easily generalised to situations involving more than two roles, and may be also carried over to situations involving interactions *between* agents belonging to the same population as well as different populations.

Let us start from introducing the notation relative to the basic situation of interaction between two agents, whereas the rules about how agents are matched to play will be illustrated in the next section. The stage game G is made up as conventional by a triplet of elements: a set L of players, a set of strategies S_i and a utility function U_i for each agent. Formally, $G = \{L, S, U\}$, where $S = \underset{l \in L}{\times} S_l$ defines the set of feasible strategies profiles, and likewise U is the set of vectors of utilities. Since I shall only be dealing with two-person games, the sets L and U are

two-dimensional, and the two players are labelled *i* and *j*. Allowing for the use of mixed strategies by the agents, we can further introduce the operator Δ(X) to express the randomisations over a set of elements X. We can thus define the set of possible randomisations over the strategy sets of the agents: $\Sigma_l := \Delta(S_l)$; finally, we can consider the vector including a randomisation for each agent: $\Sigma := \underset{l \in L}{\times} \Sigma_l$, where the generic element is indicated with $\sigma \in \Sigma$.

In the game G, the payoffs are taken to represent a measure of the *self-interest* of the agents involved. They are defined, as customary, firstly over the outcomes of the games, as represented by a pure strategies profile: $\overline{U}_l(S)$. Furthermore, taking on standard assumptions regarding expected utility, we introduce Von Neumann-Morgestern utility functions defined over mixed strategies profiles:

$$U_l(\sigma) := \sum_{s \in S} P_\sigma(s) \overline{U}_i(s) \qquad (1)$$

$P_\sigma(s)$ represents the probability that the pure strategy profile *s* is played according to the mixed strategy profile σ.

So far the analytical apparatus is common to many game-theoretic models. A major deviation is instead introduced in that agents' preferences are allowed to depend on *beliefs* over each other behaviour as well as on the 'material' outcomes of the game. This innovation makes it possible to add a wide-ranging set of motivations to self-interest, as called upon by scholars of individual choice (e.g. Ben-Ner and Putterman (1998)). The introduction of beliefs into the utility function requires an extension of the notation. A first order belief for, say, player *i* is a probability measure over the other players' mixed strategy set, namely $B_i^1 := \Delta(\Sigma_{-i})$; thus the generic element $b_i^1 \in B_i^1$ defines the probability with which *i* believes that the other players are going to implement the profile of strategies σ_{-i}. In the same fashion we can define $B_{-i}^1 := \times_{j \neq i}(B_j)$. Obviously, when there are just two active players, we have $B_i^1 := \Delta(\Sigma_j)$ and $B_{-i}^1 := B_j$. A second order belief for player *i* is a conjecture over the belief of *j* over *i*'s strategies. Therefore, it consists of a probability measure over the Cartesian of other players' beliefs of first order: $B_i^2 := \Delta(B_{-i}^1)$. Thus the generic element of this set, $b_i^2 \in B_i^2$, represents

i's probability that the belief of *j* over *i*'s strategies is b_j^1[3]. We shall indicate with $b_i = \left(b_i^1, b_i^2, ...\right)$ the infinite-dimension vector collecting the beliefs of each order for player *i*.

A concept that will prove to be useful throughout the analysis is the notion of *coherence* of beliefs. Suppose it is common knowledge that a certain mixed strategy profile $\bar{\sigma}$ is going to be played. In order for formation of expectations to be rational, a basic requirement would obviously be that an agent adjusts her vector of beliefs of any order in accordance with such information. In particular, she will assign probability one on that her counterpart will play strategy $\bar{\sigma}_j$. She will also attach probability one to her counterpart having a single-point distribution assigning probability one to her playing $\bar{\sigma}_i$. Iterating this reasoning to any higher-order belief, we have that these will be given by single-point distributions consistent with the playing of $\bar{\sigma}$. We shall call $\beta_i(\sigma)$ the distribution of beliefs *coherent* with assigning probability 1 to the strategy σ by an *i*-player, and with $\beta(\sigma) = (\beta_1(\sigma), ...\beta_n(\sigma)) \in B$ the profile of such beliefs for the *n* players.

This treatment enables us to consider a *comprehensive* utility function, where 'non-self-regarding' motivations are also considered, where these may include emotions such as surprise, anger, willingness to retaliate over actions perceived as 'wrong', and more generally other motivations as moral commitments or desire to live up to others' expectations. From the formal point of view, I define a comprehensive utility function as a function $V_l(\sigma; b)$, where beliefs are arguments of the function along with outcomes, defined through mixed strategies. Assuming that $V_l(\sigma; b)$ can be broken down into these two arguments, as will be the case throughout the paper, then we can see $V_l(\sigma; b)$ as an 'extension' of the utility function previously defined in (1). That is, $V_l(\sigma; \cdot) = U_l(\sigma)$.

---

[3] Although beliefs are probability distributions iteratively defined over probability distributions, the associated probabilities over pure strategies can be easily obtained by means of the following formulas: $P_{b_i^1}(s_j) = \int_{\Sigma_j} P_\sigma(s_j) P_{b_i^1}(\sigma_j) d\sigma_j$; $P_{b_i^2}(s_i) = \int_{B_j^1} P_{b_j^1}(s_i) P_{b_i^2}(b_j^1) db_j^1$.

Thus the first formula indicates the overall probability that player *j* is going to play s_j, according to the belief b_i^1 held by player *i*, and the second the overall probability that player *j* holds about *i*'s performing s_i, according to the second order belief b_i^2.

The distinction between self-regarding and other-regarding motives to action can be made more explicit by a further specification of the utility function. In particular, I shall assume throughout the analysis that the other-regarding motives rest upon the notion of a *normative principle* used to appraise social states of affairs, which embodies the relevant notion of fairness – or, in more general terms, or morality – that an agent adopts[4]. This generates a ranking of the strategy combinations made on the grounds of such a normative principle. This is formally analogous to an *individualistic* social welfare function in that it is dependent on the material utilities of the agents involved in the interaction and establishes a certain formal property of the material utilities' distribution among the agents themselves:

$$\overline{T} := \underset{i \in I^*}{\times} \overline{U}_i(S) \rightarrow R \qquad (2)$$

Therefore, such a normative principle permits the creation of an ordering over the possible states of affairs, which represents the assessment that an impartial spectator would give to the different social situations on the basis of the relevant normative criterion of distribution. A higher value of the function T, defined over outcomes, implies that the associated social state of affairs satisfies to a higher degree the normative criterion.

Taking the structure of the game as granted, it is possible to make the function directly dependent on the pure strategy profile set S, and, also, on the mixed strategies of the game: $T(\sigma) := \sum_{s \in S} P_\sigma(s) \overline{T}\left[\overline{U}(s)\right]$.

In analogy with individual expected utility, the expected normative function is simply a weighted sum of the indexes of welfare distribution under all possible pure strategies profiles, with weights given by the probabilities that each outcome is actually played.

The comprehensive utility function will then have the following form:

$$V_i(\sigma, b) = U_i(\sigma) + \lambda_i f_i\left[T(\sigma); b\right] \quad i \in I \qquad (3)$$

The first term U_i represents the self-interested source of utility, whereas the second term reflects the agent's concern with other-regarding motivations. This is expressed as a function

[4] For a more extensive exposition of the underpinnings of this particular version of the model, see Grimalda and Sacconi (2005).

f, shared by all agents belonging to the same population, of the social normative criterion *T*. Such a function also depends on the beliefs *b*, as the reciprocal expectations on each other's behaviour may matter in the compliance with the normative criterion *T*. For simplicity, the two components enter the function additively, and the parameters λ_i, possibly differing across populations of agents, measure the weight attributed to the other-regarding vis-à-vis self-interested utility. The function *f* may be specified in different ways in order to account for various possible forms of the morality-grounded motive to action.

2.2 The Random Matching Process

As mentioned above, I assume there exist two populations of agents, labelled with *i* and *j*, each defined on a continuum. As customary in Evolutionary Game Theory, I assume that a member from each population is drawn at random and enter a stage-game in a fixed position, i.e. *i*-players always occupy the role of the Row-player in the game, and *j*-players that of the Column-player. I also assume that each player can play a mixed strategy, rather than solely a pure strategy as is generally the case in Evolutionary Game Theory. I denote with p_i and p_j the vectors of *average play* for the two populations. That is, for a given $l=i,j$:

$$p_l(s_l) = \int_{\sigma_l \in \Sigma_l} P_{\sigma_l}(s_l)\vartheta(\sigma_l)d\sigma_l \qquad (4)$$

where $P_{\sigma_l}(s_l)$ is, as stated above, the probability of playing the pure strategy $s_l \in S_l$ according to the mixed strategy σ_l, and $\vartheta(\sigma_l)$ is the *density* of players using the mixed strategy σ_l, which satisfies the condition $\int_{\sigma_l \in \Sigma_l} \vartheta(\sigma_l)d\sigma_l = 1$; that is, the integral over all the strategies densities exhausts the Leabesgue-measure of the whole population, which has been conventionally set equal to 1.

I assume that p_i and p_j are common knowledge among players of both populations, so that any player called to play the game can compute her own expected payoff and that of her counterpart. Using the notation introduced earlier, $U_j(p_i, p_j)$ is the material payoff that an *i*-player gauges a *j*-player is expecting, given the common knowledge on average plays. $U_j(\sigma_i; p_j)$ is instead the *actual* expected payoff accrued to a *j*-player by the *actual* play by agent *i*, i.e. σ_i. Since the average plays p_i and p_j are common knowledge among players, I assume that individual beliefs are consistent with them; that is, $b_i = \beta(p_i, p_j)$. This also

permits a simplification of the notation: the comprehensive utility function will generally be indicated as a function of average plays, rather than beliefs:

$$V_i(\sigma; b) = V_i(\sigma; \beta(p_i, p_j)) \equiv V_i(\sigma; p_i, p_j) \quad (5)$$

Hence, the first argument of *V(. ; .)* refers to consequences of actions, whereas the second refers to expectations over actions.

3. NOTIONS OF EQUILIBRIA

3.1 Static Notion of Psychological Nash Equilibrium

In their seminal paper on psychological games, Geanakoplos *et al.* (1989, GPS henceforth) elaborated a concept of equilibrium for this particular setting, which is a generalisation of the Nash concept for standard games. In fact, they required two conditions to hold in equilibrium. The first is analogous to the standard Nash optimality condition, i.e. no other strategy exists giving a player a higher payoff than the equilibrium one. In other words, agents do not have an incentive to deviate from the prescribed equilibrium behaviour. The second condition concerns beliefs, and requires them to be coherent with the equilibrium play. The rationale of this second condition is quite obvious in the light of the discussion of the present section: by definition equilibrium implies that the corresponding strategy are common knowledge among players, thus it seems reasonable that beliefs should be set accordingly.

As this notion was originally put forward to address two-person games, it needs to be amended here because of the two-population setting we are dealing with. In particular, since in section 2.2 I assumed average play to be common knowledge and individual beliefs to be coherent with them, such a notion is now redundant. In other words, coherence of beliefs with equilibrium play is assumed even off-equilibrium.

On the other hand, with respect to the GPS original version, it seems natural to add a further condition, which requires that in equilibrium individual behaviour coincide with the average play within the populations. In fact, if this condition did not hold, individuals would have an incentive to perform a behaviour differing from the average, and this would gradually cause average play to change. In other words, an average behaviour that did not reflect optimal behaviour at the individual level would be likely to be swept out by a process of adjustment of

players toward optimality, which is likely to take place, though at a relatively *slow* rate, even within a setting implying bounded rationality.

Taking account of the notation introduced in the previous section, a Nash Psychological equilibrium can be restated as follows: it will be given by a pair of average plays $(\hat{p}_i, \hat{p}_j)$ such that:

i) $\hat{b} = \beta(\hat{p}_i, \hat{p}_j)$

ii) for each $l \in L$, $\hat{\sigma}_l$ that satisfies $V_l(\sigma_l; \hat{b}) \leq V_l(\hat{\sigma}_l; \hat{b})$ for every $\sigma_l \in \Sigma_l$, (6)

is such that $\hat{\sigma}_l = \hat{p}_l$

Notice that the subscript *l* refers to a generic player in either population. Condition (i) imposes coherence of beliefs with average play, which is in any case assumed even off-equilibrium. Condition (ii) ensures that individual behaviour is optimal and that average behaviour coincides with individually optimal behaviour.

3.2 Dynamic Notions of Equilibrium

3.2.1 Is the Replicator Dynamics Suitable?

I now propose what I call 'dynamic' notions of equilibria, which can be seen as refinements of the static concept previously put forward. This requires defining two conceptual tools. The first is a plausible model of dynamic evolution of the agents' behaviour. The second is a concept of equilibrium, and of stability, in the dynamic setting.

As for the first, I shall adopt the replicator dynamics as a rule of motion of agents' behaviour. Given the extensive studies carried out on the properties of replicator dynamics, all the pros and cons of its application are indeed well known (see Weibull (1995)). However, I should at least spend some words on its suitability for the case under study. In fact, the application of replicator dynamics to social interactions is usually justified on the grounds of the paradigm of the *imitation of most successful agents*. That is, individuals adopt the strategies used by other agents once they realise that these bring about better results than the strategies they are currently using. The adjustment to the currently more profitable strategies is not immediate, as information does not spread instantaneously through the system, and because agents are not

always able to process that information in the most profitable way. This is why replicator dynamics can be considered an aggregate model of evolution responding to the behaviour of boundedly rational agents. Behind this general justification for the employment of replicator dynamics, there lie some more specific underpinnings. First, better micro-founded accounts of this dynamic process can be offered. Second, other processes of evolution can be shown to lead, under some conditions, to the same results in the long run (Weibull (1995)).

The more controversial issue concerning *any* evolutionary criteria, not merely replicator dynamics, regards *what* is to be understood as 'success'. In many contexts this has a clear connotation, e.g. profit for firms involved in a competitive market. In other settings, however, especially those involving choices made by individuals, defining individual success is quite problematic. First comes the issue of identifying the *individual* notion of success. Obviously, there is no universal consensus as to which notion has to be adopted, as critics waver between a subjective and an objective notion of value. This issue is further aggravated in the present case, as the very idea that individuals *imitate* others' behaviour when they see it as more successful requires that individual notion of success are, in principle, comparable. This would call for an objective notion of value, but on the other hand the 'consumers' sovereignty' principle that characterises modern economics seems to foster a subjective account. Unfortunately the lack of a consensus in the theory of individual choice prevents me from reaching a satisfactory argument on this point, thus I shall assume that the model applies to sufficiently homogenous communities such that preferences can be taken to be the same across different individuals[5].

What seemingly makes this issue even more complicated in the present context is the presence of the other-regarding component within individual comprehensive utility. In fact, at first sight this is an even less tangible element than individual self-interest. A strategy that some scholars adopt is to apply replicator dynamics to the self-regarding rather than to the other-regarding component; that is, individuals' behaviour carrying greater 'material' or 'economic' success diffuse more rapidly across the population, unlike the fulfilment of their other-regarding motivations (Fershtman and Weiss (1998)). This account seems consistent with the biological idea of 'success' as 'fitness with respect to the environment', which in a social context would find its more direct counterpart in some economic standards. However,

those scholars' argument seems in some way to beg the question as they assume the possibility of recognizing one another's disposition to co-operate, thus indirectly making the socially rewarded behaviour the most successful one in 'fitness' terms.

However, in my view the issue of the comparability of 'success' on the other-regarding account is no more complicated than that concerning the self-interested component. For an individual is faced with the same basic problem in both spheres, i.e. that of comparing how her action fares with respect to the average in the population in terms of some shared standard of assessment of individual behaviour. Take in particular the case in which other-regarding motivations are somehow associated with social status, e.g. because people abiding by the normative criterion of assessment (2) can enjoy higher social status than others. If this is so, then it is arguable that community-members will have a clear-cut way to assess how they fare with respect to the rest of the population. For social status is almost by definition related to an inter-subjective source of value, which makes it relatively easy to effect interpersonal comparisons. To be sure, it could be argued that moral values are entrenched in an individual's system of choice in a deeper way than self-regarding preferences are, and thus they are more difficult to change over time. Nevertheless, it would be technically possible to assume that self-regarding and other-regarding evolve at different speed in this framework, but this would only complicate the analysis more than necessary. Moreover, the presence of the term λ in the comprehensive utility function represented in (3) is already a way to grasp how individuals attach different importance to the two motivational sources.

Another, apparently more technical, issue concerns the use of mixed strategies at the individual level, as I assumed in the previous analysis, despite most works have been carried out under the assumption of agents performing only pure strategy. As will be immediately clear, this latter choice makes the analysis easier under many respects. However, as highlighted by Fudenberg and Levine (1998), this is not a neutral choice as dynamics based on pure strategy seem to have a 'stabilising' effect in some cases with respect to a mixed strategy dynamic mechanism. In what follows I will still put forward a basic definition allowing for agents using mixed strategies, thus making the analysis comparable to that carried out in the static context. I apply a qualitative investigation of the properties of the

[5] Another line of defence of the present approach is that to define ex-post as a 'population' thus gathering individuals with sufficiently homogenous preferences.

equilibria in the rest of the section. Notwithstanding all these *caveats*, then, in the following analysis I shall still adopt the replicator dynamics as the basic evolutionary mechanism.

3.2.2 Deviations with Steady State-Consistent Beliefs: The GPS Replicator Steady State

The original notion of Nash Psychological equilibrium presented by GPS (1989) only holds in a static context. Besides their basic definition, they also put forward some refinements of this concept with the purpose of carrying over notions such as that of *(trembling-hand) perfect* equilibria to the new setting. The key characteristic of this type of refinement is that equilibrium strategies are slightly perturbed, thus allowing for any other strategies to be played with an arbitrary small probability (Myerson (1991)). A static equilibrium is then said to be *trembling-hand perfect* if it is still an equilibrium for all the 'perturbed' games as the perturbation becomes increasingly small. One can then interpret such a concept as making the equilibrium *robust* to small changes in the related strategy, where such changes, in some sense, 'converge' to it; hence, some unsophisticated conception of dynamic stability can be said to be embedded in this concept[6]. Therefore, it is possible to start from here in order to develop a notion of stability in a dynamic setting.

In the Nash Psychological equilibrium, the main characteristic of these refinements is that 'off-equilibrium' beliefs are required to be coherent with the equilibrium strategy. That is, even on off-equilibrium paths it is common knowledge that average play is consistent with that played under the GPS Nash equilibrium. In fact, once this notion is carried over to the present dynamic setting, its rationale is that what is being tested is whether the behaviour of players whose *Leabesgue-measure* is negligible with respect to the whole population, will converge or not, once a set of players whose Leabesgue measure is equal to 1 – namely, to the measure of the entire set - are actually playing the static *equilibrium* strategy. Only in this case would it be plausible to assume common knowledge of the would-be equilibrium strategies when analysing the situation *off* the equilibrium. In other words, this notion of dynamic equilibrium investigates the robustness of the equilibrium as changes by very 'few' mutants within the population occur, while the bulk of the population stick to the 'candidate-

[6] In reality, what still makes this notion a static one is that the perturbed games are at any rate considered in isolation from each other; that is, even if any equilibria of a 'succession' held separately from each other, it still would not imply that there was a 'tendency' for the play to become 'attracted' by the equilibrium play. One could conclude that in this case there exist an analogous relation to that between evolutionary stable strategies and stable steady states of a replicator dynamics.

to-equilibrium' strategy. In the next section, I shall discuss a stronger notion of stability, where deviations by sub-sets of the population that have positive measure are allowed.

Since we are dealing with mixed strategies, the replicator equation needs some amendments with respect to its standard version. Recalling notation introduced in section 4.1.1, its application to each density yields:

$$\frac{\dot{\vartheta}(\sigma_l)}{\vartheta(\sigma_l)} = V_l(\sigma_l; p_l) - \overline{V}_l(p_l) \tag{7}$$

where $\overline{V}$ is the average payoff obtained in population *l*:

$$\overline{V}_l = \int_{\sigma_l \in \Sigma_l} V(\sigma_l)\vartheta(\sigma_l)d\sigma_l \tag{8}$$

If one wanted to calculate the change in the play of a pure strategy, then, one should keep track of the changes in every density:

$$\dot{p}(s_l) = \int_{\sigma_l \in \Sigma_l} P_{\sigma_l}(s_l)\dot{\vartheta}(\sigma_l)d\sigma_l \tag{9}$$

A GPS replicator steady state can then be defined as a vector $\hat{p}_l$ such that

(i) $\hat{p}_l$ is a solution to the system $\dot{p}(s_l) = 0$

(ii) In the system of equations (7) $V_l(\sigma_l; b_l) = V_l(\sigma_l; \beta(p_l))$ (10)

Condition (10i) is the standard notion required for a steady state. Condition (10ii) requires that beliefs be consistent with $\hat{p}_l$ itself. However, in the two-strategy case with which I shall be dealing in the following sections, it is easier to look for the solution to the system of differential equations (7) instead of that formed by (9):

(i') $\hat{p}_l$ is a solution to the system $\frac{\dot{\vartheta}(\sigma_l)}{\vartheta(\sigma_l)} = 0$ *for any* $\sigma_l \in \Sigma_l$ (11)

In fact, this is a *more* restrictive condition than the previous one. It requires that in equilibrium there is no tendency for any mixed strategy to change its frequency, as they all fare the same as the average play given by $\hat{p}_l$.

That players have no incentive to change their mixed strategies does not necessarily imply that the associated steady state is stable; indeed, stability requires the tendency of the system

to converge on, or not to move far away from, the steady state position, after some variables have been perturbed. This usually straightforward notion now requires some qualifications as we have two types of 'variables' that are qualitatively different: strategies and beliefs. In other words, we need a condition telling us how beliefs are shaped *off-equilibrium*. I provide two different answers to such question, which build on the two main theoretical contributions on the topic of Psychological Games.

The answer that seems in line with GPS original paper is possibly the simplest one: beliefs are consistent with the steady state average play. No argument, other than analytical simplicity, is offered in GPS to underpin this hypothesis. As suggested earlier, this specification is coherent with the idea that *deviations* from the steady state equilibrium are performed by a set of agents whose Lebeasgue-measure is zero.

Rather than considering the mathematical notion of *local* stability of a steady state based on the theory of linear systems of differential equations, I will find it easier, and also more appealing from the intuitive point of view, to deal with the following *analytical* notion, especially in the two-strategy case, to which the following condition refers:

A GPS replicator steady state $\hat{p}_l$ is *Liapunov-stable* if, besides satisfying (11) and (10ii), it also fulfils the following condition[7]:

$$\exists\ \omega > 0 \quad \text{s.t.} \quad \forall\ \sigma_l \in \left|\sigma_l - \hat{p}_l\right| < \omega \quad \left\{V\left(\sigma_l;\hat{p}_l\right) - \overline{V}\left(\hat{p}_l\right)\right\}\left(\sigma_l - \hat{p}_l\right) \leq 0 \qquad (12)$$

Notice that the first term of the last inequality is that determining the growth rate in the frequency of a strategy σ_l. Therefore, this condition implies that strategies *above* $\hat{p}_l$ are characterised by payoffs no greater than the average, so that the relative frequency will not increase over time, and *vice versa*. Overall, then, frequencies are such that they will *not diverge* with respect to the steady state frequency $\hat{p}_l$. In particular, the fact that the main inequality of (12) can also be satisfied with equality means that it suffices that the system is not *led away* from the steady state, but it cannot guarantee that the system comes closer to it either. This is why I have labelled the previous concept 'Liapunov' stability, as such a concept indeed only requires the system "not to depart" from the steady state (see Hirsch and Smale (1974)).

If instead we wanted to add the strongest condition that the system *does* converge toward the steady state, then the main condition of (12) should hold with *strict* inequality. In this case, I shall talk of *local asymptotical stability:*

A GPS replicator steady state $\hat{p}_l$ is said to be *locally asymptotically stable* if, besides satisfying (11) and (10ii), it also fulfils the following condition[8]:

$$\exists\ \omega > 0 \quad \text{s.t.} \quad \forall\ \sigma_l \in \left|\sigma_l - \hat{p}_l\right| < \omega \quad \left\{V\left(\sigma_l;\hat{p}_l\right) - \overline{V}\left(\hat{p}_l\right)\right\} < 0 \tag{13}$$

Now, strategies *above* $\hat{p}_l$ are characterised by payoffs strictly greater than the average, so that the relative frequency will decrease over time, and *vice versa*. Overall, then, frequencies are such that they will indeed *converge* to the steady state frequency $\hat{p}_l$. Obviously, local asymptotic stability implies Liapunov stability. *Global* asymptotical stability would hold when the basin of attraction of a steady state coincides with the whole region on which variables exist; that is, there would exist only one local stable steady state.

3.2.3 Deviations with Off-Steady State-Consistent Beliefs: The VK Replicator Steady State

The dynamic notion of stable dynamic equilibrium put forward in the previous section was based on the idea that deviant agents have beliefs consistent with the strategies played in the static equilibrium. This is tantamount to assuming that, whereas some deviant agents are performing a different behaviour from that carried out in equilibrium, the bulk of the population is already performing the steady state behaviour and this is common knowledge to deviants as well. There seems to be some ground to argue that such a concept of dynamic equilibrium actually requires *too little*, in that only the tendency of some negligible-size cohorts of agents to converge to the equilibrium is investigated, neglecting the question of whether there is the tendency for the *whole* population to converge, at least when starting within a suitably defined neighbourhood of the equilibrium. In other words, the GPS replicator steady state only studies the stability with respect to mutations by *0-measure*

---

[7] The generalisation of this condition for the n-strategy case would be as follows:
$\exists\ \omega > 0$ s.t. $\forall\ \sigma_l \in \left\|\sigma_l - \hat{p}_l\right\| < \omega$, $\forall\ k = 1..n,\ \left\{V\left(\sigma_k^l, \hat{p}_{-k}^l;\hat{p}_l\right) - \overline{V}\left(\hat{p}_l\right)\right\}\left(\sigma_k^l - \hat{p}_{-k}^l\right) \leq 0$

[8] The generalisation of this condition for the n-dimension case would be as follows:
$\exists\ \omega > 0$ s.t. $\forall\ \sigma_l \in \left\|\sigma_l - \hat{p}_l\right\| < \omega$, $\forall\ k = 1..n,\ \left\{V\left(\sigma_k^l, \hat{p}_{-k}^l;\hat{p}_l\right) - \overline{V}\left(\hat{p}_l\right)\right\} \leq 0$

subsets of agents, but it does not deal with mutations of sets of agents with positive measure, thus falling short of some of the properties that a dynamic concept would be required to fulfil. These considerations echo those put forward by Van Kolpin with regard to the original paper of GPS (Van Kolpin (1992), VK henceforth). In fact, some of the refinements put forward by GPS, such as those of trembling-hand perfect equilibria, though not still dynamic in a strict sense, imply the study of optimal behaviour *outside* the equilibrium. Then, so Van Kolpin argues, beliefs should be designed to be consistent with the *actual* average play, rather than assuming consistency with the steady state. This makes the analysis of behaviour probably more complicated, but surely more coherent with its own premises.

Building on these considerations, I shall propose a refinement of the previous concept of GPS stable steady state, which allows for the fact of significant deviations from the steady state behaviour, and beliefs that are built consistently with such deviations. On more practical grounds, this approach implies studying the rule of motion of deviant strategy when the average play differs from the steady state, *and* beliefs are consistent with such averages. Moreover, a similar distinction to that between stability in the Liapunov sense and in the local asymptotic sense that was put forward in relation to the GPS steady state, will also be proposed here.

A VK replicator steady state $\hat{p}_l$ is *stable in the sense of Liapunov* if, besides satisfying (11) and (10ii), it also fulfils the following condition:

$$\exists\ \omega > 0 \quad \text{s.t.} \quad \forall\ \sigma_l \in \left|\sigma_l - \hat{p}_l\right| < \omega \ \ \textit{and} \ \forall\ \widetilde{p}_l \in \left|\widetilde{p}_l - \hat{p}_l\right| < \omega\ ,$$
$$\left\{V\left(\sigma_l;\widetilde{p}_l\right) - \overline{V}\left(\widetilde{p}_l\right)\right\}\left(\sigma_l - \hat{p}_l\right)\left(\sigma_l - \widetilde{p}_l\right) \le 0 \qquad (14)$$

Notice that this condition applies to the two-strategy case[9]. The main difference with respect to (12) is that the beliefs of deviant agents are now consistent with some average play $\widetilde{p}_l$ lying in a neighbourhood of the steady state $\hat{p}_l$, rather than being coherent with $\hat{p}_l$ itself as in the GPS case. Local asymptotic stability requires the main inequality to hold strictly:

A VK replicator steady state $\hat{p}_l$ is *locally asymptotically stable* if, besides satisfying (11) and (10ii), it also fulfils the following condition:

$$\exists\ \omega > 0 \ \text{ s.t. } \ \forall\ \sigma_l \in \left|\sigma_l - \hat{p}_l\right| < \omega \ \textit{ and } \forall\ \widetilde{p}_l \in \left|\widetilde{p}_l - \hat{p}_l\right| < \omega\ ,$$
$$\left\{V\left(\sigma_l;\widetilde{p}_l\right) - \overline{V}\left(\widetilde{p}_l\right)\right\}\left(\sigma_l - \hat{p}_l\right)\left(\sigma_l - \widetilde{p}_l\right) < 0 \qquad (15)$$

4. THE THEORY OF NORMATIVE EXPECTATIONS AT TEST

4.1 Sugden's Model of Normative Expectations

In the present section I shall take the model of individual choice put forward in Sugden (2000) as illustrative of the normative expectations theory[10]. This model fits the general version of a utility function separable into a self-interested and an other-regarding motivation put forward in expression (3) above. The latter component is grounded on the so-called *resentment hypothesis*, which implies that a fundamental component of human action is the willingness to avoid others' resentment when executing an action that is socially disapproved. The first systematic representation of this hypothesis is probably that offered in Adam Smith's "Theory of Moral Sentiments". In his own words, *"What reward is most proper for promoting the practise of truth, justice and humanity? The confidence, esteem and love of those we live with. Humanity does not desire to be great, but to be beloved."* (Smith, 1759/1982, p. 166). *"We are pleased to think that we have rendered ourselves the natural objects of approbation, though no approbation should ever actually be bestowed upon us: and we are mortified to reflect that we have justly merited the blame of those we live with, though that sentiment should never actually be exerted against us"*(Smith, 1759/1982, p. 116).

It is worth noting that Smith appends importance to both the willingness to avoid others' resentment and the motivation to elicit others' approval. However, Sugden takes a narrower version of this formulation, and explicitly rules out from his notion the latter aspect. The motivation he offers for doing so is that the inclusion of the 'positive' feeling of having elicited the social approval would lead to 'unnecessary' forms of altruism (Sugden (2000)). This aspect is in fact consistent with the idea that what really assigns a normative character to social norms is not so much the *approval* in the case of conformity, but rather the *disapproval* in the case of violation (Pettit (1990)). Furthermore, there is a second, perhaps more subtle

---

[9] The generalisation of this condition for the n-strategy case would be as follows:
$\exists\ \omega > 0$ s.t. $\forall\ \sigma_l \in \left\|\sigma_l - \hat{p}_l\right\| < \omega$, $\forall\ k = 1..n,\ \left\{V\left(\sigma_k^l, \widetilde{p}_{-k}^l;\widetilde{p}_l\right) - \overline{V}\left(\widetilde{p}_l\right)\right\}\left(\sigma_k^l - \hat{p}_k^l\right)\left(\sigma_k^l - \widetilde{p}_k^l\right) \leq 0$

specification of Smith's hypothesis that Sugden does, which leads him to link the social disapproval to the *expectations* of the community over an agent's actions. In this version, an agent would elicit resentment when failing to conform with the expectations that other members of the community can 'reasonably' hold on his behaviour. Sugden is also clear in asserting that a 'reasonable' expectation is one that is grounded on history, i.e. on the past occurrences of the situation. For instance, if agents have in the past successfully co-ordinated on driving on the left-hand side of the road, then each agent may hold a 'reasonable' expectation that the same will occur in the next occurrence of the interaction. Although co-ordination games are quite peculiar forms of interaction, as their equilibria are mutually beneficial, Sugden is also clear in stating that this interpretation of the 'reasonableness' of an expectations can also be carried over to situations more general than co-ordination games.

The final shift of Sugden's argument is to associate community-members' expectations with *payoffs* expectations. That is, an agent will trigger the resentment of other members of the community when inflicting a loss in their payoffs with respect to the level they expect on the basis of the past occurrences of the game. These considerations lead to the following specification. Firstly, one has to specify what Sugden calls an impact function, that is, the loss in an agent's opponent brought about by her actions:

$$m(\sigma_i; p_i, p_j) = U_j(\sigma_i; p_j) - U_j(p_i; p_j) \tag{16}$$

Recall that p_i and p_j are the average play within the *i*-player and the *j*-player population respectively, which are common knowledge across the players. Hence, an *i*-player who is playing against a generic *j*-player will expect that the *j*-player expects a payoff equal to $U_j(p_i; p_j)$. However, the *i*-player will expect that the *actual* payoff accrued to player *j* is instead given by the first factor of the left-hand side of (16). Hence, the difference between these two terms is the extra gain (loss) assigned to *j* with respect to what expected by *i*'s action. More precisely, when $m(\sigma_i; p_i, p_j) < 0$ an *i*-player is failing to conform to the normative expectations of the community of agents, as agent *j* obtains a payoff lower than expected. Conversely, if $m(\sigma_i; p_i, p_j) > 0$ agent *i* is performing an action that rewards agent *j* with an extra-payoff with respect to what expected; in Pettit's (1990) words, *i* is performing a *super-erogatory* action. However, only the former of these two aspects is relevant for

[10] In Sugden (1998) a different account of normative expectations is developed. However, the merely qualitative treatment of the dynamics makes this model unsuitable to a comparison with the present approach.

Sugden's version of the resentment hypothesis, as an action *benefiting* the counterpart with respect to the initial expectation does not bring about any psychological reward to the agent performing it – or at least, such a reward is not considered as a relevant component of the model. This is the reason of the discontinuity of the function (16) at its zero. On the basis of these considerations, the other-regarding component within the comprehensive utility function will take the following form[11]:

$$f_i(\sigma;b)=\begin{cases}0 & \text{if} \quad m(\sigma_i;p_i,p_j)\geq 0\\ m(\sigma_i;p_i,p_j) & \text{if} \quad m(\sigma_i;p_i,p_j)<0\end{cases} \tag{17}$$

4.2 Static Equilibria in a PD

In testing the implications of the previous model, I shall focus on the following general version of the Prisoner's Dilemma, where the limitations that $\beta>\gamma>\alpha>\delta$ ensures the fulfilment of the usual properties of the interaction:

	Co-operation	Defection
Co-operation	γ,γ	δ,β
Defection	$\beta,\ \delta$	α,α

Figure 1

For the purpose of the analysis, it is key whether the quantity $(\beta-\gamma)$ exceeds $(\alpha-\delta)$. Let us first assume that

$$\eta \equiv (\beta-\gamma)-(\alpha-\delta)>0 \tag{18}$$

Making use of a definition put forward in the literature (Fershtman and Weiss (1998)), under condition (18) individual strategies can be called *substitutes*, as the 'disincentive' to co-operate is larger when the other party is Co-operating than when she is Defecting.

In what follows I report the main results of the analysis and the graphical illustration of the equilibria that can be found in the game. In the Appendix one can find more detailed computations. The first insight in the game is that it is never optimal for agent *i* to perform a

[11] The dependence on the difference between actual and expected payoff has here been assumed linear for simplicity, despite Sugden only constrains overall utility to be monotonically decreasing in *m* when this is

'super-erogatory' action. This depends on the fact that normative expectations do not reward actions that accrue *greater* utility than expected to the opponent with a positive extra utility.

Therefore, the only strategies that are feasible equilibria will be those such that $(\sigma_i - p_i) \leq 0$. By solving the optimisation problem for a generic *i*-player, the following inequality obtains:

$$\frac{\partial V_i(\sigma_i; p_i, p_j)}{\partial \sigma_i} \geq 0 \Leftrightarrow p_j \leq \overline{p}_j \tag{19}$$

where

$$\overline{p}_j = \frac{\lambda_i(\beta - \alpha) - (\alpha - \delta)}{(1 + \lambda_i)\eta} \tag{20}$$

The implication of inequality (19) is as follows: provided that the average play within the *j*-player population is (strictly) *below* the threshold level given by (20), then increasing the probability of Co-operation increases the overall payoff of an *i*-player. This is of course true for all *i*-players who are co-operating with probability less than the average p_i within the *i*-population. In order to appreciate the intuition behind this result, we first have to notice that inequality (19) is meaningful only insofar as $\overline{p}_j$ lie between zero and one. This implies the following condition on the parameter λ_l:

$$\lambda_{\min} = \frac{(\alpha - \delta)}{(\beta - \alpha)} < \lambda_l < \frac{(\beta - \gamma)}{(\gamma - \delta)} = \lambda_{\max} \tag{21}$$

In fact, if λ_l did not lie at 'intermediate' levels, then it would make either unconditioned Co-operation (when λ_j is relatively high) or unconditioned Defection (when λ_l is relatively low) the dominant strategies for the agent. Throughout the paper, instead, I shall focus on those cases that are strategically more interesting and that do not prescribe an unconditioned behaviour to an agent. More precisely, conditions (21) concern the *inclination to resentment* of an individual when failing to live up to others' expectations; overall, they state that resentment will be the prevailing motivation only in the context that is *less* costly in terms of self-interest. Since in the present context of substitute individual strategies, Co-operation is *more* costly when the other party is *co-operating* rather than when she is defecting,

---

negative.

resentment will *permit* Defection when the counterpart is co-operating, and will *impede* Defection when the other party is defecting[12].

This explanation should also make it clear the rationale of condition (21); under the substitute strategies assumption, the probability with which the opponent, on average, co-operates must not be too high in order to spur the co-operation of agent *i*; in fact, were it too high the individual would start to defect, as in that case the self-interested motivation overcomes resentment considerations. Conversely, if the opponent co-operates with a sufficiently low probability, the inclination to resentment will trigger a co-operative behaviour. Obviously, given the symmetry of the game, an analogous condition holds for *j*-players. To be sure, such a behaviour may appear paradoxical, but is a consequence of the resentment hypothesis. If this is a genuine prompt to action, it must prevail over self-interest in at least some occasions; however, in the context of a PD it can prescribe a submissive behaviour in the face of an opportunistic one.

Diagrammatical analysis shows that a large number of equilibria are possible. In Figure 2, I have depicted the best reply functions for two generic players belonging to the *i*-population and the *j*-population. Notice that the two threshold levels are not necessarily the same, as they could differ for a different value of λ_l, i.e. the two populations may be different because of the weight attributed to other-regarding utility. Moreover, the shape of the function is such that it is never optimal to co-operate with higher probability than the average of the population; that is, the best reply function for player *l* is constrained to lie below p_l. A preliminary condition to find an equilibrium is that, as usual, the two best reply functions intersect. However, this is not enough, as condition (10ii) also states that individual optimal play must coincide with average play in a population. Therefore, none of the three candidates for equilibrium circled in Figure 2 can be considered equilibria of the game.

---

[12] In fact, the first inequality can be rearranged to yield: $\lambda(\beta-\alpha)>(\alpha-\delta)$
The first term is the loss, due to resentment, of other-regarding utility, whereas the second is the benefit in terms of self-regarding utility stemming from a drop in the probability of co-operation, provided that the other party is defecting. Therefore, this condition ensures that the resentment cost outstrips the self-interested benefit under defection from the other party. Analogous considerations hold for the second inequality, which can be so re-expressed: $\lambda_l(\gamma-\delta)<(\beta-\gamma)$

Here, the first term represents the resentment for failing to co-operate and the second the self-interested gain, provided that the other party is co-operating.

Instead, outcomes in which the population average play is below the threshold level are equilibria. Such are the configurations belonging to the set: $E_1 = \{(\hat{p}_i; \hat{p}_j): \hat{p}_i \leq \overline{p}; \hat{p}_j \leq \overline{p}\}$. Figure 3 shows one of such equilibria. Point E_1 in Figure 3 is a mixed strategy equilibrium, where the probability of Co-operation is bounded from above by the two threshold levels. This makes the corresponding outcome overall *inefficient,* in the usual sense in which mutual Defection is inefficient in a PD. Moreover, since no agent is required to produce a super-erogatory action when the other agent is not, we may qualify this set of outcomes as *reciprocal*. In fact, the probability of Co-operation is low because expectations on each other population's co-operation is low, which fails to trigger the resentment mechanism. Hence, such a set can be called an *inefficient reciprocal* type of equilibria. Notice that it also includes as a particular case the standard Nash equilibrium of the game $(\hat{p}_i = 0; \hat{p}_j = 0)$.

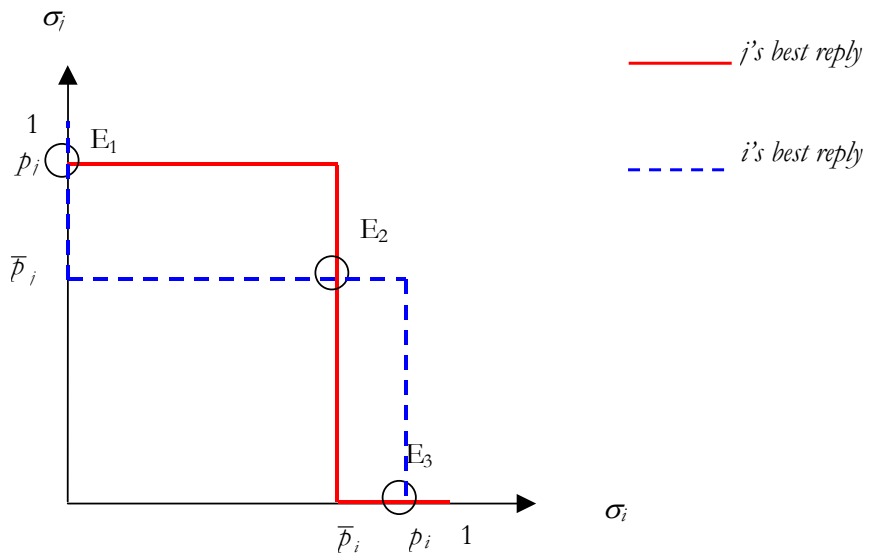

Figure 2

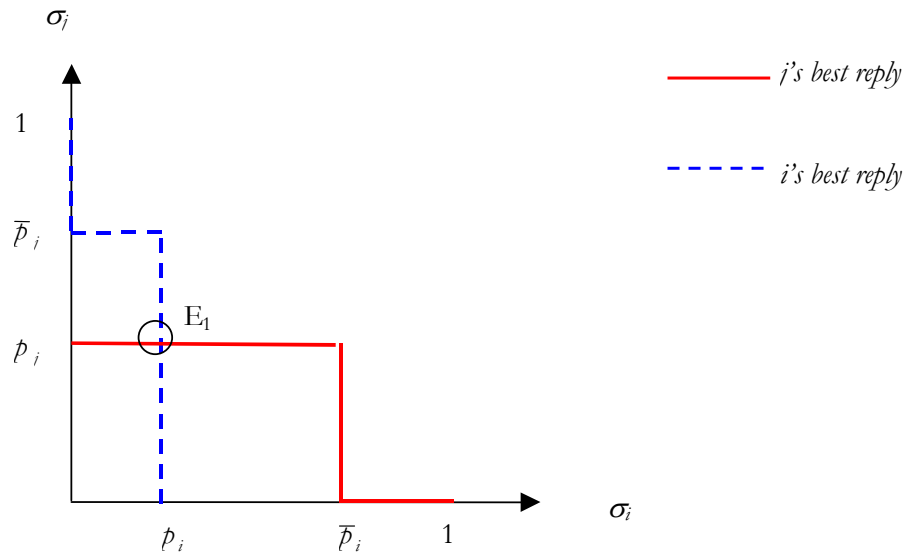

Figure 3

Figure 4 shows a type of equilibrium where all the agents of a population act submissively – namely, they co-operate with high probability - whereas all of the others act exploitatively. As the picture shows, all the outcomes such that $E_2 = \{(\hat{p}_i; \hat{p}_j): \hat{p}_i = 0; \hat{p}_j \geq \overline{p}\}$ are equilibria (of course, symmetrical outcomes are equilibria as well). To mark the contrast of this set of equilibria with the other, I shall call this type *anti-reciprocal*, or *exploitative*, in that one group of individuals is prompted to co-operate by the very fact of others' Defection: on the one hand, resentment-inclined individuals will feel obliged to live up to *i*-players expectations, demanding as these may be. On the other hand, the very low level of expectations set on *i*-players in relation to their co-operation, justified by their population's general opportunistic behaviour, suffices to avoid the resentment of their opponents. The seemingly paradoxical character of this equilibrium lies in that it is sustained by expectations that may be deemed as *empirical*, but not *causal*; that is, general conformity to the Co-operative norm by *j*-players is not triggered by considerations in terms of self-interest, but from the mere past conformity of individuals in that population (see Sugden (2000: 107-112)).

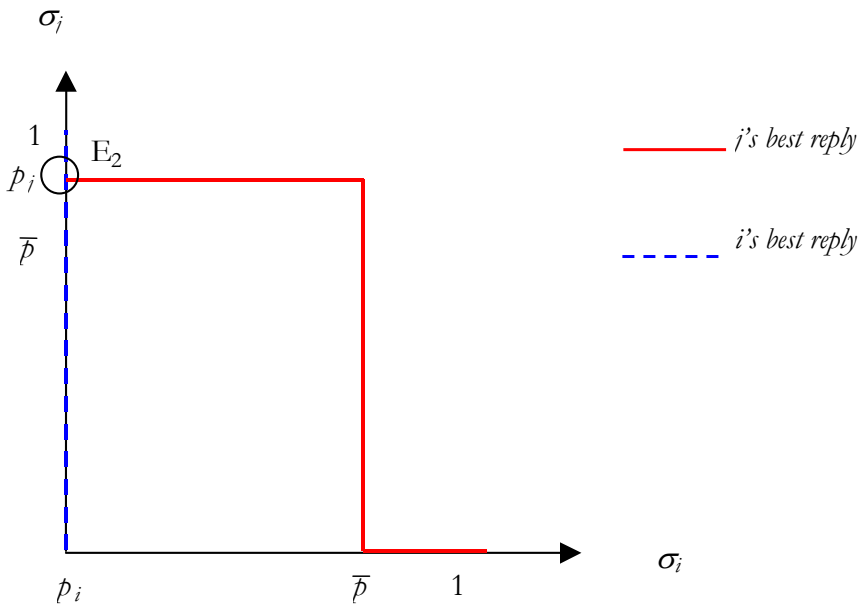

Figure 4

I now take on the case of Complementary Strategies, where $\eta<0$. In this case, a set of *(almost)-efficient* equilibria is possible. In fact, the previous optimality inequality (17) is now reversed:

$$\frac{\partial V_i(\sigma_i; p_i, p_j)}{\partial \sigma_i} \geq 0 \Leftrightarrow p_j \geq \overline{p}_j \qquad (22)$$

where the threshold value is the same as in expression (20). Now, the conditions that ensure that there is no dominant strategy are as follows:

$$\frac{(\beta-\gamma)}{(\gamma-\delta)} < \lambda_\eta < \frac{(\alpha-\delta)}{(\beta-\alpha)} \qquad (23)$$

The interpretation is the same as that outlined above; however, as individual strategies are now complements, a reversal of the terms of those inequalities occurs. This third type of equilibria is illustrated in Figure 5. This set can be given a general representation as follows: $E_3 = \{(\hat{p}_i; \hat{p}_j): \hat{p}_i \geq \overline{p}_i; \hat{p}_j \geq \overline{p}_j\}$. The economic intuition is analogous, but 'opposite in sign' with respect to that given for E_1 and E_2.

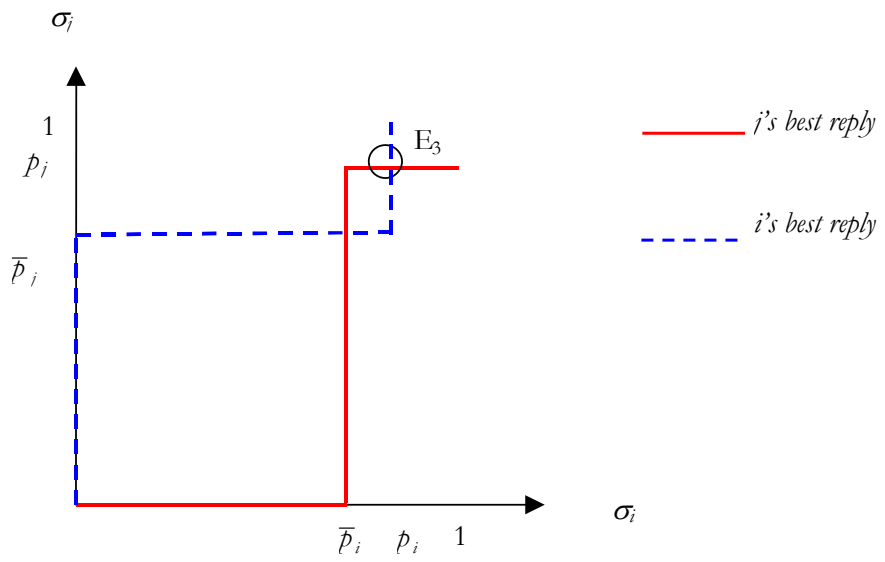

Figure 5

The inclination to resentment is now triggered when the other party co-operates, given the smaller opportunity cost, in terms of self-interested utility, borne by the individual in this situation. Therefore, each individual has sufficient incentive to co-operate when the other is Co-operating, thus bringing about this *reciprocal* equilibrium. Since the probability of Co-operation is now bounded from below, it seems natural to call this an *efficient,* or *almost-efficient*, equilibrium. In this setting, no equilibrium can be sustained such that agents co-operate with probability less than $\overline{p}$, the only exception being the standard Nash equilibrium where both populations always defect.

4.3 Dynamic Equilibria in the PD

In what follows, I illustrate how the concept of GPS stable steady state can be used to test whether the first type of solutions reported in section 4.2, i.e. *inefficient* equilibria in the substitute strategy case, can be GPS replicator stable steady states. Notice that such a static equilibrium is certainly a trivial solution to the system formed by (11). What needs to be checked is whether this steady state is stable. In order to do this, we first have to compute the average payoff of the population, which is made easier by the assumption that beliefs are consistent with the steady state strategy. The payoff of a generic *i*-player who is playing that

equilibrium strategy is thus $V_i(\hat{p}_i;\hat{p}_i,\hat{p}_j)$. Therefore, the average player in population *i* will not experience any resentment, as her behaviour coincides with that of the bulk of the population: $m_i(\hat{p}_i;\hat{p}_i,\hat{p}_j)=0$. Hence, her comprehensive payoff boils down to her self-interested one.

As for payoffs from 'deviant' behaviour, this, once again, varies in relation with whether we consider strategies 'above' or 'below' the average play level. Consider first the case of $\sigma_i > \hat{p}_i$. Here, the analysis is made easier by the shape of the resentment function: since super-erogatory actions are not rewarded with greater social approval, then the agent cannot gain any extra other-regarding utility from this type of action, thus the comparison between average payoff depends only upon the material component. But clearly the deviant agent gains an inferior payoff than the average, since Defection is the dominant strategy of the stage game. As a consequence, the density of any mixed strategy above the equilibrium level $\hat{p}_i$ is bound to decrease. Slightly more complex is the case of $\sigma_i < \hat{p}_i$, as now other-regarding utility does enter into play. However, the computation of comprehensive utility for the deviant agent in this case, shows that the same condition as (19) holds. This implies that for all $\sigma_i < \hat{p}_i$ the deviant players' frequency of play will increase (decrease) provided that $p_j < \overline{p}_j$ ($p_j > \overline{p}_j$). But this is indeed the case in regions surrounding the equilibrium, by construction of equilibria of type E_1.

Figure 6 shows the phase diagram of this case, drawn on the grounds of the foregoing analysis. Notice that the directions of the arrows signal the tendency of change of strategies within the sub-population of deviant agents. The result is clearly the *local* stability of the steady state coinciding with the static GPS equilibrium. The intuition is that there exists a tendency for deviant players to conform to the general behaviour of the majority of the population. Co-operating with higher probability than average is clearly inefficient as no gain is reaped. But also playing Defection with higher probability than average is not optimal, as the resentment induced in other-regarding utility outstrips the gain in material utility. Therefore, deviant behaviour will converge to average behaviour.

It is worth noticing that the condition determining the local stability of this steady state is the same as that which ensures that this is a Nash Psychological equilibrium of the game. This is

not surprising, as the coherence between individual and average behaviour that we had imposed for the static concept of equilibrium (6) is clearly reminiscent of a dynamic notion of convergence. Moreover, the relationship between static Nash Psychological equilibria and stable GPS replicator steady states seems analogous to that between Nash equilibria and stable replicator steady states (see Weibull (1995); Fudenberg and Levine (1998)). In fact, since expectations are bound to be consistent with the equilibrium, the other-regarding component of utility will not be relevant in the comparisons between the payoffs, so that these can be carried out in terms of standard self-regarding utility functions. Though this appears a general result, a formal proof will not be provided here.

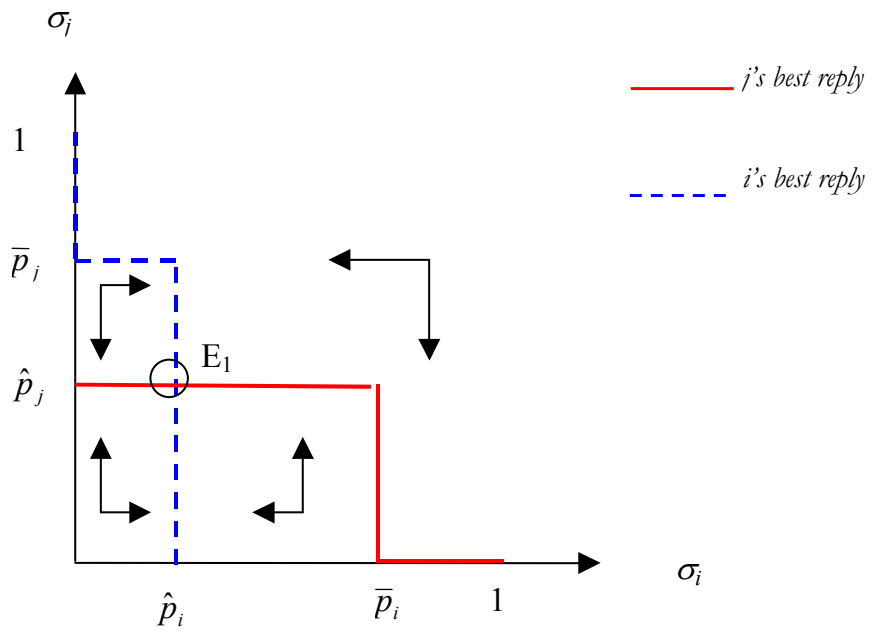

Figure 6

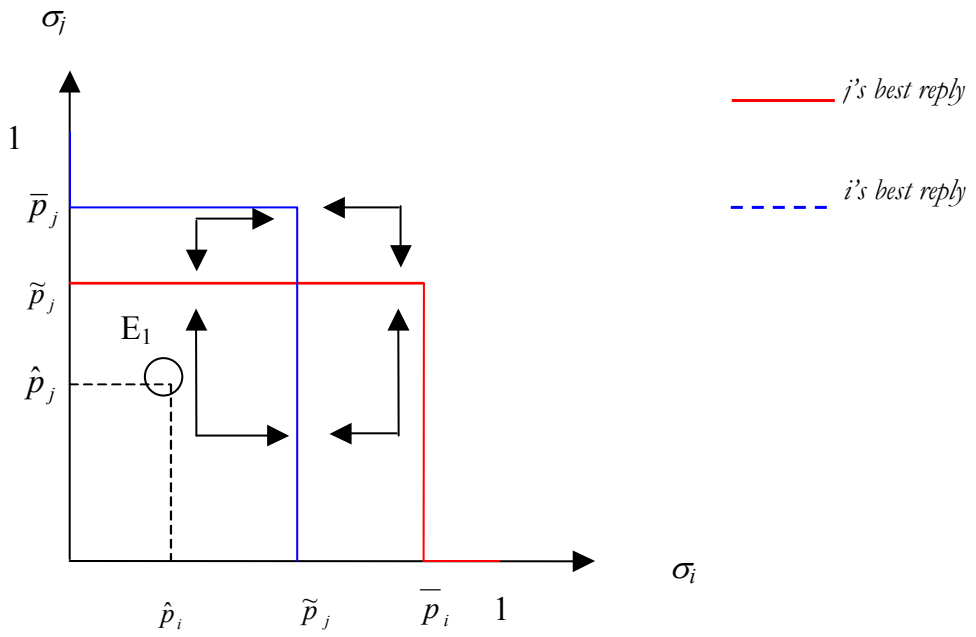

Figure 7

Applying the VK concept of dynamic equilibrium to the analysis of the Prisoner's Dilemma seen in the previous section does imply a substantial difference, as Figure 7 shows. In fact, the same reasoning developed to analyse the previous case, now implies that the system will tend to orbit around the *actual* average play $\tilde{p}_l = (\tilde{p}_i, \tilde{p}_j)$, provided that $\tilde{p}_l < \overline{p}_l$, $l = i,j$. In other words, there is no tendency for the system to move away from the current position and reach the 'designated' steady state $(\hat{p}_i, \hat{p}_j)$. In the light of the definitions of stability just put forward, we can conclude that $(\hat{p}_i, \hat{p}_j)$ is *stable* in the *Liapunov* sense, but not in the local *asymptotical* sense: given a steady state, the system will not depart away from a neighbourhood of the steady state, but it will not converge toward it either.

The reason for this result is that every sub-set of deviant agents will find it convenient to abide by what the bulk of the population is already doing: those who are Co-operating with higher probability than the average do not gain any reward for this, thus they will find it worthwhile to decrease their level of co-operation; those who co-operate with smaller probability than the average, provided that the *j*-player population is expected to perform a not too high amount of co-operation that elicits co-operation to an *i*-player, will experience

resentment for causing a loss in utility to the opponent with respect to what expected, and this will outstrip the gain in material utility. Then, they will be prompted to increase their probability of Co-operation.

Therefore, although the VK criterion does not rule out steady states as *unstable*, it qualifies their stability as a Liapunov one, thus it implies that the system will lack a tendency to move away from its current position. This appears to be a general characteristic of this version of the normative expectations theory, which also carries over to the other types of equilibria that we have found, i.e. *anti-reciprocal*, or *exploitative*, and *efficient* ones.

5. AN APPLICATION: THE IMPACT OF HETEROGENEOUS POPULATION ON CO-OPERATION LEVELS

As a way of illustration of the possible applications of this game-theoretic framework, the case of a change in the composition of the population will be analysed in the present section. This model will form a first basic building block to address some of the questions that have been laid out in the introduction as regards the influence of globalisation on social norms of co-operation. In the concluding section, some possible extensions of the present analysis will be presented.

One of the channels whereby globalisation has reshaped social interactions is through the proposition of alternative modes of behaviour than those previously existing. Generally speaking, there have been two ways in which this has happened: migration and the diffusion of global means of communication, such as the television and the Internet. To be sure, both channels have been active for a very long period of time, which surely spans a longer phase than the one most commonly associated with globalisation. However, even if this was the case, the analysis of their bearing on social norms would not for this reason be less interesting; moreover, it is not the purpose of this paper to argue when the globalisation's clock has started ticking, but a credible approach to globalisation issues is that most of the factors underlying globalisation have been in place for a long time, but it is only when the scope of these factors has become overarching that scholars have started adopting this concept extensively. What migration and global means of communication has triggered in what were more "homogenous" and less differentiated societies is the introduction of different cultures,

moral values, which have ultimately led to different modes of behaviour. New modes of behaviour live alongside the previously existing ones in societies that favour a multicultural approach, or they become blended in a "melting pot" in societies that favour integration of different cultures. Both phenomena are indeed worth attention, but in the present section we shall deal with the former, and only analyse a particular aspect of the complex relationship between new and "ascending" social norms and traditional and long-established ones. The question I want to address in this section is simply what is the impact on co-operation levels of the introduction of different moral values in a section of a society.

I model this change in the society's overall moral disposition in typical economic fashion, by studying how an equilibrium is perturbed after a "shock" in some of the main parameters of the model occurs. In this particular setting, I assume that the existing "equilibrium" where a society was located is characterised by a homogenous population. That is, all of the individuals in this society share the same moral values, and have the same disposition to apply this in practice. In terms of the model previously developed, this society is characterised by two populations of *i*-players and *j*-players who are actually drawn from the *same* population, so that their being labelled as *i*-players or *j*-players is purely conventional. As a consequence, the two populations shares of individuals who are disposed to co-operate will coincide. That is, $p_i = p_j$. Moreover, individuals who are part of a sub-population will have a common moral criterion *T* for assessing states of affairs, and an identical disposition to comply with such moral prescriptions within their own objective functions, i.e. $\lambda i = \lambda j$. For the sake of simplicity, let us suppose that the moral criterion coincides with the normative expectations theory illustrated above. That is, the moral criterion *T* is given by the material utility of one's counterpart in pairwise interactions as represented in (17). For the purposes of this section, it would not matter if different moral criteria were chosen. Moreover, let us focus on the complementary strategies case (see section 4.2). According to the static equilibria analysis, nearly efficient equilibria can emerge in this case, provided that the actual share of population who co-operates exceeds the threshold values $\overline{p}_i$ and $\overline{p}_j$ in both populations. Given the two sub-populations come from a homogenous pool, $\overline{p}_i = \overline{p}_j$. Suppose then that this is the case.

Now, suppose that the *j*-players sub-population suddenly change their attitudes toward moral values. In particular, although they keep on sharing a common moral criterion with *i*-players, they now attach less weight to this factor vis-à-vis the self-interested ones within their

objective function. The reason may either be that a different population of agents have migrated to the previously homogenous society, bringing in different types of behaviour than those previously existing. Or that a part of the previously homogenous population has taken on a different stance on the extent to which they should abide by moral values, thus determining a change in the weight they attribute to other-regarding motivations. More specifically, suppose this change goes in the direction of *reducing* the weight attached to moral values. That is, the new *j*-players population is characterised by a weight for other-regarding utility $\hat{\lambda}j < \lambda j$. As a result, the threshold value that separates the Co-operation region from the Defection region will shift downwards. We denote such a new value with $\hat{\bar{p}}_i$. The intuition for this change is simple. Since *j*-players are now less concerned than the previous population of *j*-players with other-regarding utility, they will be less resented when breaching the moral norm imposing to comply with others' expectations. Consequently, only if *i*-player's expectation on *j*-player's probability of co-operation is higher than before will the resentment of failing to comply with the moral norm become overriding in pushing the individual to co-operate rather than following her self-interest. Note that a value of the threshold $\hat{\bar{p}}_i$ higher than before denotes a higher proportion of *i*-players co-operating, and thus a higher expectation that a *j*-player *should* co-operate. In fact, in this model, the motivational strength of normative expectations depends on the loss in terms of material utility inflicted on the counterpart, thus a higher proportion of co-operation in the opponents' populations means that the loss in case of defection is also higher.

The impact of this change can best be seen in the following diagram. If the shift upwards in the *j*-player population's best strategy is relatively large, and it exceeds the actual percentage of the *i*-player population who is co-operating, then the co-operative equilibrium will collapse. The reason for this result is a direct consequence of what just illustrated. Given the change in disposition of a *j*-player in terms of compliance with the moral norm, the normative expectation must be higher than before to elicit a co-operative behaviour from a *j*-player. If this is not the case, then a *j*-player, given her now more selfish-oriented attitude, will switch to defection in instances where a former *j*-player still found it overall convenient to co-operate. Hence, looking at the dynamic evolution of the interaction, more and more *j*-players will gradually switch to defection. As the proportion of *j*-players co-operating shrinks, *i*-players will find that expectations on their own level of co-operation has diminished, too, because the loss inflicted on their counterparts when failing to co-operate is now on average

smaller than before. Hence, when the actual proportion of *j*-players actually co-operating falls below of the threshold level $\overline{p}_j$, even *i*-players will start switching to defection. As a result, the system will converge towards the socially inefficient equilibrium where everybody defects.

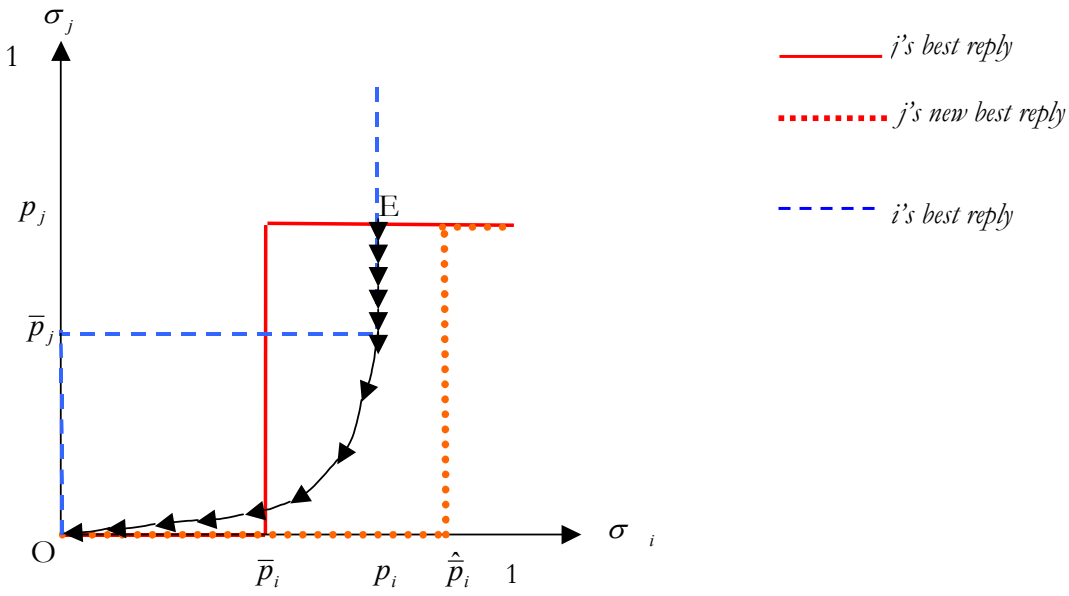

Figure 8

This result is admittedly very sketchy and only based on a rather gross generalisation of the complexity of phenomena related to globalisation and migration. Its main purpose was to show how this framework makes it possible to address some interesting aspects related with the evolution of social norms when a society is exposed to global forces that reshape the composition of the population in terms of attitudes towards compliance with moral norms. By no means I wish to generalise the results by saying that the presence of heterogenous cultures and/or moral values is always detrimental to the provision of public goods, or that multi-culturalism will lead to the disruption of a society's ethos. What I am claiming is that the influence of these phenomena is not obvious and they may have not clear-cut consequences on co-operation levels. Therefore, a framework like that developed in this paper may help shed some light on the issue.

6. CONCLUSIONS

The purpose of the present paper has been to lay out the foundations of a theoretical framework to study the evolution of social norms in a society. Given the generality of this approach, the present model seems in particular suited to study the influence of globalisation on social norms of co-operation within a given society.

The initial section has set out the broader picture of the possible causal links between globalisation and co-operation, the main idea being that globalisation alters the individual's perception of the social distance within and between different societies, as well as the frequency and the nature of social relationships. Section 2 and 3 have developed a general game-theoretic framework to study the evolution on social norms in a society. Section 3 and 4 have further specified the theoretical aspects of the model. Section 5 has aimed to show a possible application of this model to an aspect of the globalisation process, that is, the impact of a change in the disposition to comply with shared moral customs by a segment of society, which may be deemed as an effect of either migratory influx, or the receipt of different modes of behaviour through the media of global communication.

This latter application, albeit very stylised and far from grasping the whole complexity of the globalisation process, nevertheless shows the relevance of framing the problem of the influence of globalisation on social norms of co-operation in a dynamic context such as the one developed in the present paper. Since the main thrust of the model is the social outcome of interactions involving people with different social habits, moral values, or cultural traits, the model may be applied to the study of whether multi-culturalism is a better model than cultural integration with respect to the generation of public goods in a society. More specifically, the model may study the possible outcomes and contrast the relative welfare levels of two different models of social inter-relation; one would see population segmented in two culturally distinct sub-population, both strongly cohesive internally, but less oriented to co-operation vis-à-vis members of the other sub-population; the other model would model the society as overall more integrated, but with less intense social bounds, which may lead to feebler co-operative attitudes.

7. APPENDIX: DETERMINATION OF EQUILIBRIUM IN A PD

Let us consider the situation of a generic player *i*, who knows (and knows that it is common knowledge) that the percentage of plays in either population is given by the pair (p_i, p_j). First, she has to compute the expected payoff for a generic *j*-player, on the grounds of the first and second order beliefs consistent with the pair (p_i, p_j). This will be given by the following expression:

$$E_i\left[U_j(p_i, p_j)\right] = (\gamma + \alpha - \beta - \delta)p_i p_j + (\beta - \alpha)p_i - (\alpha - \delta)p_j + \alpha \quad (24)$$

Consequently, the impact function for player *i* by playing σ_i is:

$$m_i(\sigma_i; p_i, p_j) = \left[(\gamma - \delta)p_j + (\beta - \alpha)(1 - p_j)\right](\sigma_i - p_i) \quad (25)$$

Notice that the sign of m_i only depends on the sign of the expression $(\sigma_i - p_i)$. Other-regarding utility can thus be rewritten as:

$$f(\sigma_i; p_i, p_j) = \begin{cases} 0 & \text{if} \quad \sigma_i \geq p_i \\ \left[(\gamma - \delta)p_j + (\beta - \alpha)(1 - p_j)\right](\sigma_i - p_i) & \text{if} \quad \sigma_i < p_i \end{cases} \quad (26)$$

This expression is consistent with the resentment hypothesis as modelled by Sugden (2000), in that implementing a co-operative action with higher probability than the average does not provide a higher payoff; the opposite is true when a less co-operative action is performed.

The overall extended utility for agent *i* is then given by:

$$V_i(\sigma_i; p_i, p_j) = (\gamma + \alpha - \beta - \delta)\sigma_i p_j + (\beta - \alpha)p_j - (\alpha - \delta)\sigma_i + \alpha + \lambda_i f(\sigma_i; p_i, p_j) \quad (27)$$

We now have to work out what is the optimal action for agent *i*. This can be done by differentiating expression (4.8) with respect to σ_i, which leads to:

$$\frac{\partial V_i(\sigma_i; p_i, p_j)}{\partial \sigma_i} = (\gamma + \alpha - \beta - \delta)p_j - (\alpha - \delta) + \lambda_i\left[(\gamma - \delta)p_j + (\beta - \alpha)(1 - p_j)\right]Ind(\sigma_i < p_i) \quad (28)$$

where

$$Ind(\sigma_i < p_i) = \begin{cases} 1 & \text{if} \quad \sigma_i < p_i \\ 0 & \text{otherwise} \end{cases} \quad (29)$$

One can the notice that if $(\sigma_i - p_i) > 0$, then the latter term of the differential is nil, whereas the first two are both negative. This implies that it will never be optimal to perform 'super-erogatory' actions.

References:

- Axelrod, R. (1984). *The Evolution of Cooperation*, New York: Basic Books
- Barth, F. (1995). *Ethnicity and the Concept of Culture*, Harvard: Harvard University Press
- Ben Ner, A. and L. Putternam (eds.) (1998), *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 3-69.
- Fehr, E. and Schmidt, K. (2001): *"Theories of Fairness and Reciprocity – Evidence and Economic Applications",* Institute for Empirical Research in Economics, University of Zurich, WP N. 75
- Fershtman, C. and Weiss, Y. (1998). "Why do we care what others think about us?", in Ben-Ner, A. and Putterman. L. (eds): *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 133-50
- Fudenberg, D. and Levine, D. (1998). *The Theory of Learning in Games*, Cambridge (MA): MIT Press
- Fukuyama, F. (1996). *Trust: the social virtues and the creation of prosperity*, London: Penguin
- Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989), "Psychological Games and Sequential Rationality", *Games and Economic Behavior*, Vol. 1, pp. 60-79
- Grimalda, G. and Sacconi, L. (2005). "The Constitution of the Not-For-Profit Organisation: Reciprocal Conformity to Morality", *Constitutional Political Economy*, Vol. 16, forthcoming
- Hannerz, U. (1992). *Cultural Complexity: Studies in the Cultural Organisation of Meaning*, New York: Columbia University Press
- Hardin, R. (1982). *Collective Action*, Baltimore: Johns Hopkins University
- Kandori, M. (1992). "Social norms and Community Enforcement", *Review of Economic Studies*, Vol. 59, pp. 63 - 80
- Kaul, I., Conceicao, P., Le Goulven, K. and Mendoza, R. (2003). *Providing Global Public Goods*, Oxford University Press
- Kolpin, V. (1992): "Equilibrium Refinements in Psychological Games", *Games and Economic Behavior*, Vol. 4 N. 2, p. 218-228
- Ledyard, J. O. (1995). "Public Goods", in Kagel, J. and Roth, A. (eds.), (1995). *Handbook of Experimental Economics,* Princeton: Princeton University Press, pp. 112-194
- Lewis, D. (1969), *Convention: A Philosophical study*, Cambridge, MA: Harvard University Press.
- Messick, David M. and Marilynn B. Brewer (1983) "Solving Social Dilemmas: A Review." In L. Wheeler and P. Shaver (Eds.), *Review of Personality and Social Psychology*, (Vol. 4, pp. 11-44). Beverly Hills, CA: Sage.
- Myerson, R; (1991), *Game Theory: Analysis of Conflict*, Cambridge, MA: Harvard University Press
- North, D.C. (1990) *Institutions, Institutional Change and Economic Performance*, Cambridge: Cambridge University Press.
- Olson, M. (1965). *The Logic of Collective Action*, Cambridge: Mass.: Harvard University Press
- Ostrom, Elinor (2003) "Toward a Behavioral Theory Linking Trust, Reciprocity, and Reputation." *Trust & Reciprocity*. New York: Russell Sage Foundation, 19-79.
- Pettit, P. (1990). "Virtus Normativa", *Ethics*, Vol. 100, pp. 725-755
- Putnam, R. (2000). *Bowling alone: the collapse and revival of American community*, New York: Simon and Schuster
- Scholte, J.A. (2000). *Globalisation: A Critical Introduction*, Basingstoke : Macmillan
- Sen, A. K. (1999). *Development as Freedom*, New York: Knopf
- Smith, A. (1759): *The Theory of Moral Sentiments*, Oxford: Clarendon Press, (reprinted: 1976)
- Sugden, R. (1998), "Normative expectations: the simultaneous evolution of institutions and norms", in Ben-Ner, A. and Putterman, L. (eds): *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 73-100.
- Sugden, R. (2000): "The motivating power of expectations", in J. Nida-Rumelin and W. Spohn, (eds). *Rationality, Rules and Structure*, Amsterdam: Kluwer, pp. 103-29
- Taylor, M. (1987). *The Possibility of Co-operation*, Cambridge: Cambridge University Press
- Weibull, J. (1995), *Evolutionary Game Theory*, Cambridge: Mit Press