UNIVERSITY OF
CAMBRIDGE

Cambridge Working Papers in Economics

With God We Trust: Religion, Trust and Cooperation in Large-Scale Societies

Julien Gagnon

CWPE 1406

With God We Trust: Religion, Trust and Cooperation in Large-Scale Societies

Julien Gagnon*

15th April 2014

Abstract

The first aim of this paper is to revisit the puzzle of cooperation in large-scale societies. It proposes a game theoretic model showing how endogenous emotion-based punishment can sustain full cooperation when interactions are not repeated, provided that players' endogenous *trust* is high enough. The model is extended to allow for players' heterogeneity, in which case multiple stable equilibria of cooperation can coexist. The second aim of this paper is to explain how certain institutions may support trust and cooperation in large societies. It builds on the example of a religious group and shows that costly religious requirements may foster trust within a community, which in turn bolsters cooperation. When players are heterogeneous, the model shows that religion may also serve as signalling device to exclude defectors. Religion is thus shown to have a twofold role of *trust coordination* and *signalling*. This paper thus extends the signalling theory of religion. Finally, the model enables clear and tractable predictions about the levels of religious affiliationand participation within a society. Evidence of the model's implications is discussed.

JEL Codes: D02; D03; D71; D81; Z12.
Keywords: Cooperation; Emotions; Psychological Game Theory; Punishment; Religion; Trust.

1 Introduction

The sustainability of cooperation in large-scale societies is puzzling. In such societies, most interactions are not repeated, and reputational systems tend to break down due to individuals' lack of information on their peers' history. Standard economic models predict in these contexts that cooperation, an individually costly but socially beneficial behaviour, cannot form a dominant strategy for rational agents. Standard theory has thus proven limited to account for humans' capacity to engage in durable cooperative relationships on a large scale.

The primary aim of this paper is to revisit the puzzle of cooperation in large-scale societies. It proposes a game theoretic model showing how endogenous, costly "rational" punishment can sustain cooperation when interactions are not repeated, provided that players' endogenous *trust* is high enough. Costly punishment has been widely and reliably found to be central to the sustainability of cooperation in experimental settings (Gächter et al., 2010; Henrich et al.,

*Trinity College and Faculty of Economics, University of Cambridge, jg653@cam.ac.uk. I wish first to express my gratitude to Sanjeev Goyal for numerous discussions and comments and continuous support throughout this work. I am also grateful to Gabriel Arsenault, Sarika Bensal, Diego Cerdeiro, P.-A. Dupuis Laflamme, Steven Durlauf, Valérie Gagnon, Edoardo Gallo, Alex Harris, Sulaiman Ijaz, Sriya Iyer, Stephen Morris, Laura Richter, Larry Samuelson and Anand Shrivastava for useful comments and discussions.

2006; Fehr and Gächter, 2000, 2002) and in ethnographic accounts (Henrich et al., 2010), even in one-shot interactions. Theoretical work has also shown that the "altruistic punishment" of defectors can proliferate and sustain cooperation in large populations (Henrich, 2004; Boyd et al., 2003; Fehr and Gächter, 2002). Models of altruistic punishment have however typically overlooked agents' "rational" motivations to punish defectors as they simply assume that the taste to punish is intrinsic to (some) agents' preferences. This approach ignores important facets of agents' decision to punish and the mechanisms underlying it (e.g. negative emotions; see e.g. Falk et al., 2005; Fehr and Gächter, 2000). In particular, evidence suggests that *trust* is key to agents' adherence to cooperative norms and readiness to punish defection, which contributes to explain why levels of cooperativeness vary across communities (Balliet and Van Lange, 2013; Fischbacher et al., 2001). To my knowledge, no attempt has been made to analyse theoretically the relationship between trust and punishment and their joint role in supporting cooperation.

The model introduced in this paper is based on three stylized facts emerging from experimental economics, game theory, psychology and evolutionary biology:

> ***Stylized Fact A***. *Cooperative norms are enforced and upheld by the punishment of defectors;*
>
> ***Stylized Fact B***. *Punishment arises primarily from negative emotions, which make cooperators willing to sacrifice their own material well-being to reduce the well-being of defectors;*
>
> ***Stylized Fact C***. *Negative emotions are prompted primarily by disappointed expectations, which depend on outcomes and players' beliefs.*

The benchmark model features a continuum of homogeneous players who are pairwise matched in a one-shot prisoners' dilemma. Unlike the canonical prisoners' dilemma, players have, in addition to their material payoffs, belief-based social preferences interpreted as negative emotions when they are "cheated on". Indeed, when a cooperator i is matched with a defector j, j's utility enters negatively into i's utility. A cooperator matched with a defector is thus willing to punish her partner even if punishment is costly (Fact B), which is made possible in a second and last stage of the game. The intensity of cooperators' negative emotions is increasing in their expectations about the proportion of cooperators in the population: the more they *trust* their coplayers, the more intense their negative emotions when they are cheated on (Fact C). The model shows that emotion-based punishment can sustain full cooperation at equilibrium, although full defection is also always an equilibrium (Fact A). I then extend the model to allow for heterogeneity in the form of an additional, individual-specific cost of cooperation. I show that multiple interior "stable" equilibria of trust and cooperation can be sustained in a population, even though full defection is again always an equilibrium.

The second aim of this paper is to explain how certain institutions, such as religious organisations, may have helped towards, or been necessary to, the achievement of higher trust and cooperation in large societies[1]. To do so, I build on the example of a religious group. I assume that any one player can form such a group and tie its membership to an observable and costly signal (e.g. costly religious requirements). Players thereafter play the game as in the benchmark model. It is shown that if religious requirements are costly enough, they permit the coordination of players' trust as players know that if their coreligionists did not intend to cooperate, they would not partake in the religious group. This makes full cooperation an equilibrium in weakly dominant strategies. Hence, if the optimal requirements are not too costly in comparison to the benefits of cooperation, a religious group arises endogenously and ensures the coordination of players' expectations and the realisation of full cooperation. Finally, I show that when players are heterogeneous, then a religious group may also serve as a signalling device to exclude those who would never cooperate. In such case, a religious group plays a twofold role (*signalling* and *coordination*), *necessary* for cooperation to be achievable. In that respect, my model extends the signalling theory of religion (e.g. Levy and Razin, 2012, forthcoming; Berman, 2000; Iannacconne, 1992, 1994).

Lastly, the model enables clear and tractable predictions about the levels of religious affiliation (the proportion of players partaking in the religious group) and religious participation (the intensity of religious practice) within a society. It thus proposes a theory of the factors influencing the "size" of religion in a society. It shows notably that increasing benefits of cooperation, decreasing benefits of defection or decreasing cost of punishment all entail an increase in religious affiliation and a decrease in religious participation (and vice-versa). Evidence supporting the implications of the model is discussed.

The rest of this paper is divided as follows. As the three stylized facts underpinning the model are central to the arguments of this paper, I devote the next section to reviewing the literature and evidence on which they rest. In the third section, I introduce and develop the benchmark model. I discuss its implications and how it relates to the existing literature. In the fourth section, I allow players to partake in a religious group. I analyse how such possibility changes the results of the model, and how these results relate to existing findings in the literature. The fifth section discusses some of the predictions of the model in terms of religious affiliation and participation within a society. The sixth section concludes.

[1] It is commonly accepted that the evolution of large-scale societies "required norms and institutions that sustain fairness *in ephemeral exchanges*" (Henrich et al., 2010: 1481; italics added), and a dominant view in sociology and evolutionary anthropology regards religion as one of these institutions (Henrich et al., 2010; Norenzayan and Shariff, 2008; Bulbulia et al., 2008).

2 Stylized Facts: Punishment, Emotions and Cooperation

2.1 Stylized Fact A: Cooperative norms are enforced and upheld by the punishment of defectors

Costly punishment has been widely and reliably found to be central to the sustainability of cooperation in experimental settings (Gächter et al., 2010; Henrich et al., 2006; Fehr and Gächter, 2000, 2002) and in ethnographic accounts (Henrich et al., 2010), even in one-shot interactions. It is also regarded as central to the evolution of cooperation (Boyd et al. 2010; Fehr and Fischbacher, 2004; Boyd et al., 2003).

In economics, the theory of repeated games has widely been used to account for costly punishment and cooperation in situations where they seem a priori inconsistent with selfish behaviour (see e.g. Kandori, 1992; Axelrod and Hamilton, 1981). Simply put, this theory contends that sufficiently patient players may at equilibrium choose strategies that credibly deter their coplayers from deviating from the cooperative behaviour.[2] Kandori's (1992) seminal paper on informal community enforcement uses this approach. In his model, agents are randomly pair-wise matched at every period and cooperate until they are matched with a defector, after which they defect forever.[3] Hence, if players are sufficiently patient, they have a strong incentive to cooperate since a single defection, in the long run, induces the complete unravelling of cooperation within the community.

The repeated game approach nevertheless suffers empirical shortcomings. First, it cannot account for many stylized facts characterizing punishment and cooperation enforcement (see Sobel, 2005, for an extensive presentation of this argument). In particular, it falls short of explaining punishment in large groups where players interact "with people they will never meet again, and where reputation gains are small or absent" (Fehr and Gächter, 2002: 137). Punishment in such contexts is yet well evidenced. In fact, Henrich and colleagues' (2010) experimental findings across different societies support the argument that punishment becomes *more* important in large groups: "as reputational systems break down in larger populations, *increasing* levels of diffuse costly punishment are required to sustain large harmonious communities" (2010: 1483; italics added).

[2] Alternatively, in n-players settings, players may want to cooperate to build a reputation of cooperation and thus induce their coplayers to cooperate with them. Reputation-based models operate slightly differently as some learning takes place but rest on the same general intuition (see Sobel, 2005, for a review).

[3] Punishment is thus undirected in Kandori's approach: a cooperator matched with a defector responds by defecting thereafter, which punishes all players including cooperators. In my model, punishment is directed only at defectors after their defection, while players' decision to cooperate or not depends only on their expectations about all other players' propensity to cooperate. This formulation is more tightly linked to the voluminous empirical literature regarding both punishment and conditional cooperation (see e.g. Fischbacher et al., 2001).

2.2 Stylized Fact B: Punishment arises primarily from negative emotions, which make cooperators willing to sacrifice their own material well-being to reduce the well-being of defectors

Defection in public good games or offers perceived as unfair or derisory in ultimatum, dictator or trust games, to name a few examples, trigger strong negative emotions that have been shown to motivate players to punish selfish players and defectors, even if punishment is materially costly (Falk et al., 2005; Fehr and Falk, 2002; Fehr and Gächter, 2000, 2002). In fact, a reliable finding in experimental economics and psychology is that anger and related negative emotions are the best predictors of one's decision to punish, even when punishment is costly (see e.g. Roberts et al., 2013; Hopfensitz and Reuben, 2009).[4] De Quervin and colleagues (2004) even find that punishment, although stemming from negative emotions, can be intrinsically pleasurable (c.f. Huettel and Kranton, 2012). Anger is also viewed as having had a determinant role in the evolution of cooperation through punishment (Jensen, 2010; Fessler, 2010).

The most straightforward way to incorporate individuals' negative emotions and emotion-based punishment in a formal framework is to assume that agents hold negative other-regarding preferences when they are "cheated on" (see e.g. Falk and Fischbacher, 2005, 2006; Falk et al., 2005, 2008; Rabin, 1993). Models or intrinsic reciprocity or interdependent preferences provide "clearer and more intuitive explanations" for costly punishment and cooperation than models that abstract from emotions (Sobel, 2005; 393). These models typically adopt a utility function of the form $u_i(s_i,\ s_j) = \pi_i(s_i,\ s_j) + a_{ij}(s_i,\ s_j) \cdot \pi_j(s_j,\ s_i)$ for any player i with coplayer j, where π_i and π_j are respectively i's and j's material payoffs, s_i and s_j are respectively i's and j's chosen strategies, and $a_{ij}(.)$ is the *social preferences weight*, that is the weight i attaches to j's material payoffs (see Sobel, 2005, for a review). The specific form of $a_{ij}(.)$ depends on the emotion or nature of the social preference at play. In the context of cooperation enforcement and emotion-based punishment, it is natural to conceive of $a_{ij}(.)$ as reflecting the intensity of i's negative emotions vis-à-vis j when the latter defects.

2.3 Stylized Fact C: Negative emotions are prompted primarily by disappointed expectations, which depend on outcomes and players' beliefs

The third stylized fact states that the magnitude and sign of one's emotions vis-à-vis someone else's actions depend on one's *ex ante* expectations about one's payoffs.[5] Empirical evidence

[4] Interestingly, Roberts et al. (2013) recently found that when controlled for anger, players' willingness to punish defectors in public good games remained unaffected by the "probability of future interaction". This provides further evidence that the repeated games theory fails to account for mechanisms that appear central to cooperation and cooperation enforcement.

[5] In particular, this implies that individuals' incentives to enforce cooperative norms depend on their expectations about how

underpinning this fact is abundant. Bosman and van Winden (2002), Bosman et al. (2005) and Reuben and van Winden (2008), in a series of experiments of power-to-take games (PTTGs),[6] find that receivers' expectations regarding proposers' take rate predict their level of anger (and punishment through destruction of their endowments) far more accurately then their opinion about the "fair" take rate. In the same vein, Charness and Dufwenberg (2006) and Dufwenberg and Gneezy (2000) show experimentally that players strive to avoid "guilt", which is an increasing function of players' expectations about their coplayer's expectations. In other words, they show that players display a desire to live up to their coplayer's expectations and the higher they perceive these expectations to be, the higher their guilt when they fail. A natural explanation is that players anticipate their coplayer's negative emotions (e.g. disappointment and anger) to be higher when they think she expects more.

The modelling of belief-dependent preferences and emotions lends itself well to *psychological games* (PGs), which allow players' utility to depend on their beliefs as well as on their material payoffs. Rabin (1993) pioneered the application of PGs with a model of intent-based reciprocity in which he adopts a utility function of the form $u_i(s_i,\ s_j) = \pi_i(s_i,\ s_j) + a_{ij}(s_i,\ s_j) \cdot \pi_j(s_j,\ s_i)$, introduced earlier, but allows the social preferences weight $a_{ij}(.)$ to vary with players' *ex ante* expectations. In his model, player i values positively player j's utility if i believes that j behaved "kindly", and vice-versa. In turn, player i is more disposed to positive reciprocity if she believes that player j behaved kindly, and vice-versa.[7] To model anger and anger-based punishment, I adapt Smith's (2009) psychological game model of frustrated anger. Smith uses a utility function similar in spirit to that used by Rabin. His specification of $a_{ij}(.)$, when translated to a one-shot non-sequential game, takes the form $\max\{0,\ \pi_{i,0} - \pi_i^*(s_i,\ s_j)\}$, where $\pi_{i,0}$ is player i' *ex ante expected* material payoffs and $\pi_i^*(s_i,\ s_j)$ is i's *realised* material payoffs resulting from i's and j's actions. Hence, higher expectations lead to greater intensity of anger when expectations are not met. I adopt the same specification but, in addition, allow cooperators who have been cheated on (i.e. players for whom $a_{ij}(.) < 0$) to directly inflict, at a certain cost, a punishment on their coplayer in a second stage.[8]

much these norms are actually observed by their peers. When a norm is well observed, expectations are high, and deviations will typically trigger strong nevative emotions. In other words, "violations of social norms do not make us angry if we are used to them" (Dubreuil, 2010a: 217).

[6] In PTTGs, two players are given the same amount of money. One subject (proposer) is then allowed to take a fraction of the second subject's (receiver) endowment. The receiver can destroy a fraction of her own endowment in order to "punish" the proposer by reducing her/his payoffs (so doing, she also destroys the same fraction of her own final payoffs).

[7] This "believed kindness" depends on the difference between i's expected payoffs from the outcome stemming from i and j's actions and some "focal" payoff. He assumes this focal payoff to be the "equitable" payoff, itself measured as an average of i's highest and lowest possible payoffs.

[8] Battigalli and Dufwenberg (2007, 2009) exploit an intuition very similar to Smith's. They assume that players derive negative utility from guilt, which they feel if their actions fail to live up to a coplayer's expectations. Hence, the more player i expects j to expect, the guiltier i will feel if she does not live up to j's expectations. As it will be shown, to allow players to punish a coplayer directly in a second stage permits to incorporate both anger and guilt in the model. My model thus allows incorporating the two approaches in a unified fashion in the context of a specific application.

3 The Model

3.1 The Benchmark Model: Homogeneous Players

3.1.1 Players, Strategies and Payoffs

Consider a continuum of measure one of homogenous players who are pairwise randomly matched in a prisoners' dilemma (PD) game. The one-shot nature of the game, the random matching and the continuous population altogether enable the model to capture short-term interactions. These interactions, as discussed earlier, are typical of large-scale societies. Furthermore, in the context of punishment which is central to the model, bilateral interactions permit a starker and clearer analysis of players' incentives. Many-player games such as public-good games could nevertheless be analysed in the framework of the model.

The interactions modeled here have two stages. In the first stage, players play the PD game. In the second stage, players may punish their coplayer. Such punishment takes place because players who cooperated in the first stage display *negative social preferences* towards their coplayer if the latter defected. Such negative social preferences spark a desire to alter a coplayer's utility through punishment. I will now detail the two stages of the game.

First-stage strategies and payoffs. In the first stage, players' set of possible strategies is defined as $S_i = S_j = \{C,\ D\}\ \forall i,\ j \in [0,1]$, where C and D stand for "cooperation" and "defection", respectively. Players opting for C (D) are called "cooperators" ("defectors"). Let $\pi_i\ :\ S_i \times S_j \to \mathbb{R}$ denote player i's material payoffs. Throughout, I shall use the notation $s_i \in S_i$ and $s_j \in S_j$ to denote players i and j's chosen strategies. The *material payoffs matrix* for player i is written as follows:

$$\begin{array}{ccc} & s_j = C & s_j = D \\ s_i = C & \pi_c & 0 \\ s_i = D & \pi_d & 0 \end{array} \tag{1}$$

with $\pi_d > \pi_c > 0$. Note that defection is thus materially weakly dominant. I assume $\pi_i(s_i = D,\ s_j = D) = \pi_i(s_i = C,\ s_j = D) = 0$ only for simplicity and without altering the intuition behind the results.

Social preferences. From the "material game" above, I derive a "psychological game" based on the stylized facts (A), (B) and (C). I base players' utility function on the following equation, introduced earlier:

$$u_i\left(s_i,\ s_j\right) = \pi_i\left(s_i,\ s_j\right) + a_{ij}\left(s_i,\ s_j\right) \cdot \pi_j\left(s_j,\ s_i\right) \tag{2}$$

I assume that a player i has negative social preferences towards her coplayer j only when the latter "cheats on her", i.e. when i cooperates while j defects.[9] These negative social preferences are interpreted as *negative emotions* (e.g. anger) towards a defecting coplayer. Furthermore, as implied by the stylized fact (C), I assume that the magnitude of a cooperator's negative emotions towards a defecting coplayer depends on the former's *ex ante* expectations. Hence, the higher the material payoffs a player expects *ex ante*, the higher the negative emotions she will feel if her coplayer defects. To incorporate these assumptions into a specification of the social preferences weight, $a_{ij}(.)$, consider the following definition.

Definition 1. *Let $q_i \in [0,1]$ denote player i's* trust*, which corresponds to i's* ex ante *belief about the probability of being matched with a cooperator.*

Using Definition 1, players *ex ante expected material payoffs* can be written as $E_i(\pi_i(s_i,\ \tilde{s_j})) = q_i[\pi_i(s_i,\ C)] + (1-q_i)[\pi_i(s_i,\ D)]$. Let $\tilde{Q}(q_i)$ denote player i's *disappointed expectations,* defined as the difference between i's *ex ante* expected material payoffs and realised payoffs. Formally:

$$\tilde{Q}(q_i) \equiv\ E_i(\pi_i(s_i,\ \tilde{s_j})) - \pi_i(s_i,\ s_j) \tag{3}$$

Using (3), the full specification of the social preferences weight $a_{ij}(.)$ can be written as follows:

$$a_{ij}(s_i,\ s_j,\ q_i) \equiv \begin{cases} -\gamma\tilde{Q}(q_i) & if\ s_i = C \neq s_j \\ 0 & otherwise \end{cases} \tag{4}$$

with $\gamma > 0$, an exogenous parameter. This specification neatly incorporates the stylized fact (C) in that players' negative emotions, through their disappointed expectations, depend on both outcomes and players' trust. In particular, the more a cooperator i trusts her coplayer j, the higher will i's negative emotions be when j fails to cooperate. Likewise, the higher the loss incurred on i by j's defection, the stronger will i's negative emotions be.

Using equations (2) and (4), players' utility after the first stage of the game can be fully written as follows:

$$u_i(s_i,\ s_j,\ q_i) = \pi_i(s_i,\ s_j) + a_{ij}(s_i,\ s_j,\ q_i) \cdot u_j \tag{5}$$

Second-stage strategies. A cooperator i can impose a *punishment loss* p on a defector j at a cost $c(p)$, with $c'(.) > 0$, $c''(.) > 0$ and $c(0) = 0$. The set of possible strategies for players in this second stage is defined as $P_i = P_j = \mathbb{R}_+$. To understand players' incentive to punish,

[9] Rabin (1993) and Falk and Fischbacher (2006) allow $a_{ij}(.)$ to be positive when two players cooperate or are "kind" to each other. Battigalli and Dufwenberg (2007) and Smith (2009), in contrast, focus solely on negative emotions. Since I focus on the role of negative emotions in punishment and the inclusion of "positive reciprocity" would change little to the general intuition of this model, I opt for the latter approach.

note that players i and j's utility functions, upon i being cheated on by j, can be written, respectively, as:

$$u_i\left(C,\ D,\ q_i\right) = -\gamma\tilde{Q}\left(q_i\right)u_j \tag{6}$$

$$u_j\left(D,\ C,\ q_j\right) = \pi_d,\ \forall q_j \in [0,1] \tag{7}$$

Incorporating player i's possibility to inflict a punishment loss of $p > 0$ on j at a cost $c\left(p\right)$ in the second stage and substituting (7) into (6), i's utility can be written as:

$$u_i\left(C,\ D,\ q_i\right) = -\gamma\tilde{Q}\left(q_i\right)\left[\pi_d - p\right] - c\left(p\right) \tag{8}$$

As shown by equation (8), player i can clearly gain some utility by reducing j's utility through punishment. Indeed, punishment reduces her negative emotions, $-\tilde{Q}\left(q_i\right)u_j$, and can thus be regarded as an "anger-relief" good that cooperators are willing to "buy". This modeling thus captures the stylized fact (B), stating that punishment arises from negative emotions making people willing to sacrifice their own well-being to reduce the well-being of defectors.

Consider now the subgame starting after j cheated on i in the first stage. Player i then solves the following decision problem (8):

$$\max_{p>0}\left\{-\gamma\tilde{Q}\left(q_i\right)\left[\pi_d - p\right] - c\left(p\right)\right\} \tag{9}$$

This maximisation problem yields the following first order necessary condition assuming an interior equilibrium:

$$c'\left(p^*\right) = \gamma\tilde{Q}\left(q_i\right) \tag{10}$$

which admits a unique interior equilibrium for i's punishment, denoted $p^*\left(q_i\right)$. Note that since $a_{ij}\left(s_i,\ s_j,\ q_i\right) = 0$ whenever the strategy profile $\left(s_i,\ s_j\right) \neq \left(C,\ D\right)$, clearly only cooperators matched with defectors will inflict a punishment loss on their coplayer.

Remark 1.

Cooperator i's optimal level of punishment on defector j is increasing in both i's trust, q_i, and i's material loss from j's defection, π_c. It is also decreasing in the marginal cost of punishment.

The proof to Remark 1 follows directly from the properties of the punishment cost function, $c(.)$. As noted earlier, the higher i's trust (q_i) and the higher the payoffs to cooperation (π_c), the higher i's potential disappointed expectations $(\tilde{Q}\left(q_i\right))$, and the stronger i's negative emotions when j defects. In turn, the stronger i's negative emotions, the stronger i's punishment on j. Hence, a corollary of Remark 1 is that more trusting players punish defecting coplayers more.

Let us now consider players' *ex ante* expected utility value function,[10] expressed as follows:

$$E_i\left[u_i\left(s_i,\ \tilde{s_j},\ q_i,\ \tilde{q_j}\right)\right] = E_i\left[\pi_i(s_i,\ \tilde{s_j})\right] - \mathbf{1}_{\boldsymbol{s_i=D\neq s_j}}\left(E_i\left[p^*\left(\tilde{q_j}\right)\right]\right)$$
$$-\mathbf{1}_{\boldsymbol{s_i=C\neq s_j}}\left[\gamma\tilde{Q}\left(q_i\right)\left[\pi_d - p^*\left(q_i\right)\right] - c\left[p^*\left(q_i\right)\right]\right] \tag{11}$$

Equation (11) shows how player i's utility depends on (i) the material outcomes of the game; (ii) i's *trust* or *ex ante* expectations; and (iii) i's expectations about j's level of trust. This dependence on beliefs and second-order beliefs is a defining characteristic of psychological games.[11] Incorporating this utility function into the normal-form game, the *final payoffs matrix* for player i is as follows:

	$s_j = C$	$s_j = D$
$s_i = C$	π_c	$-\gamma\tilde{Q}\left(q_i\right)\left[\pi_d - p^*\left(q_i\right)\right] - c\left[p^*\left(q_i\right)\right]$
$s_i = D$	$\pi_d - E_i\left[p^*\left(\tilde{q_j}\right)\right]$	0

(12)

3.1.2 Equilibrium Analysis

Because the game described above forms a psychological game, I use the definition of psychological Nash equilibrium, imposing that all beliefs and higher-order beliefs about behaviour conform to actual behaviour at equilibrium. Since payoffs are *ex ante* uncertain, I adapt this definition to Bayesian games. Hereinafter, for simplicity, I shall refer to any solution of the game as an "equilibrium".

Definition 2. *A strategy profile* $s^* = \left[\left(s_i^*,\ p_i^*\right),\ \left(s_{-i}^*,\ p_{-i}^*\right)\right]$ *is an equilibrium if,* $\forall i$ *and* $s_i^* \neq s_i',\ \forall s_i^*,\ s_i' \in S_i$*:*

1. $p^*\left(q_i\right)$ *obeys equation (10);*
2. $E_i\left[u_i\left(s_i^*,\ \tilde{s_{-i}},\ p_i^*,\ \tilde{p_{-i}^*},\ q_i,\ \tilde{q_{-i}}\right)\right] \geq E_i\left[u_i\left(s_i',\ \tilde{s_{-i}},\ p_i^*,\ \tilde{p_{-i}^*},\ q_i,\ \tilde{q_{-i}}\right)\right]$*;*
3. $q_i = q_{-i} = E_i\left(\tilde{q_{-i}}\right) = E_{-i}\left(\tilde{q_i}\right) = q^*$*;*

where q^* *corresponds to the actual proportion of cooperators in the population.*

[10] This is a value function because the optimal punishment, which solves equation (10) for all players conditional on their level of trust, is incorporated directly in players' utility.

[11] Note that the model permits to incorporate explicitly both Smith's (2009) modelling of frustrated anger and Battigali and Dufwenberg's (2007) modelling of simple guilt, even though these authors focus on sequential games. Indeed, imposing $c\left(p\right) = \frac{1}{2}p^2$ generates the exact same negative psychological payoffs for players as those proposed in Smith's (2009) model of frustrated anger. This specification also yields the result $p^*\left(q_i\right) = \tilde{Q}\left(q_i\right)$, which entails that any defector i expects a punishment loss of $E_i\left[\tilde{Q}\left(q_j\right)\right]$ if matched with a cooperator j. When this punishment cost is interpreted as "guilt", it corresponds exactly to Battigali and Dufwenberg's (2007) modelling of simple guilt.

Definition 2 states that any equilibrium must form a subgame perfect equilibrium. Conditional on players' beliefs and strategy in the first stage, players' punishment strategies must maximise their utility in the second stage. In turn, players' optimal strategy in the first stage must be consistent with backward induction and maximise their expected utility conditional on their level of trust. This level of trust must be rational at equilibrium and correspond to the actual probability of being matched with a cooperator.

Proposition 1.

Let p^*_{max} denote a cooperator's optimal punishment when her level of trust is maximal (i.e. when $q_i = 1$).

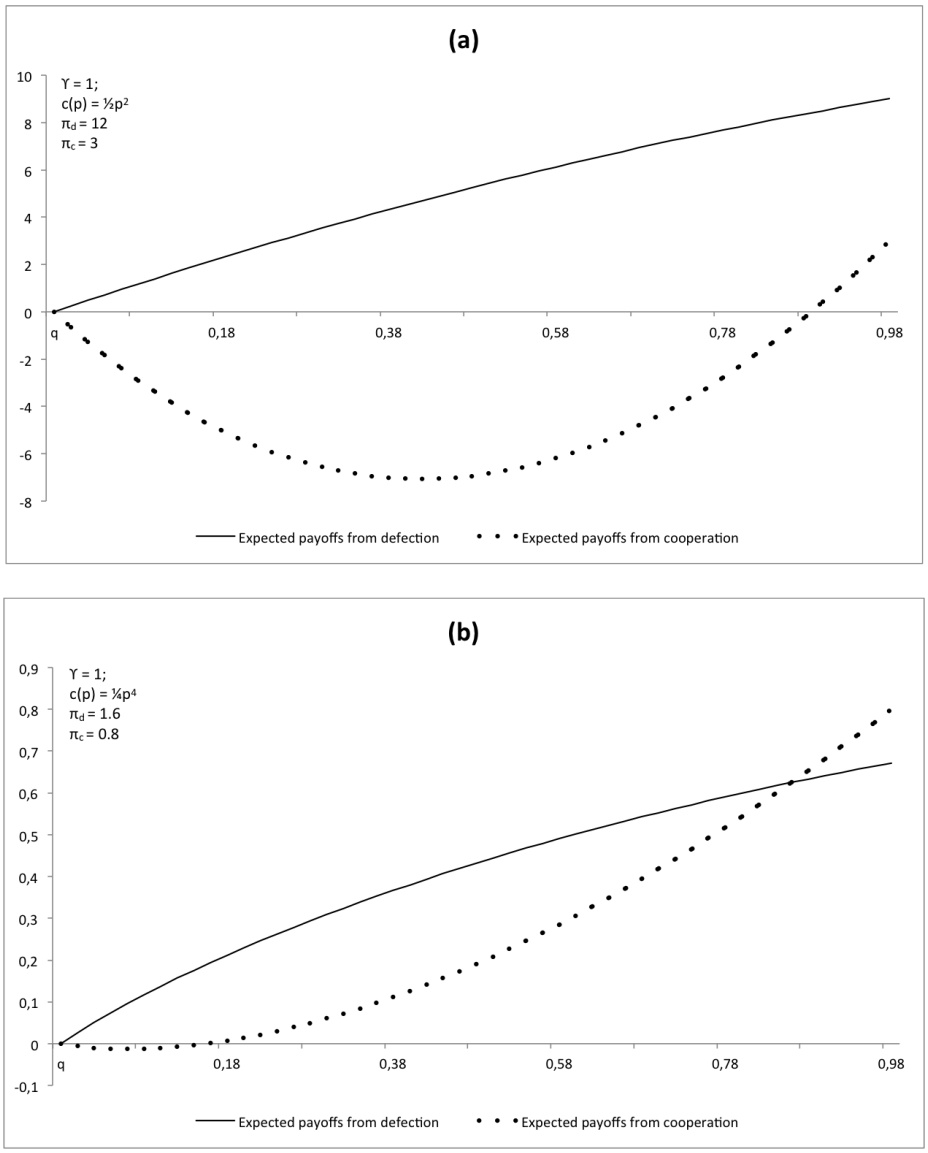
1. *If and only if the benefits of cooperation are lower than the benefits of defection when cooperation is maximal (i.e. if and only if$\pi_c < \pi_d - p^*_{max}$), then there exists a unique low equilibrium with $q^* = 0$;*
2. *If and only if $\pi_c \geq \pi_d - p^*_{max}$, then there exists three equilibria, namely a low equilibrium ($q^* = 0$,), a high equilibrium ($q^* = 1$), and an intermediate equilibrium ($q^* \in (0, 1)$).*

Proof. All proofs are given in the Appendices.

Figure 1 (a) presents a calibration generating a unique low equilibrium. Figure 1 (b) provides an example of calibration with multiple equilibria. The curves represent the expected payoffs (vertical axis) of cooperation and defection as a function of q, the proportion of cooperators, which correspond to players' trust level at equilibrium. Graphically, an equilibrium obtains when the curve of "expected payoffs from cooperation" crosses the curve of "expected payoffs from defection" (except for the possible corner solution of full cooperation).

As illustrated on the graphs, a direct implication of Proposition 1 is that the game always displays a low equilibrium, that is an equilibrium characterised by full defection, no trust and no punishment. This result is intuitive. Suppose that all players expect full defection, and expect all other players to expect full defection too. Then, players' payoffs consist solely of material payoffs as psychological payoffs are equal to zero. Full defection is then an equilibrium in weakly dominant strategies.

Figure 1: Trust-based Equilibria



However, Proposition 1 also shows how the model captures the stylized fact (A), stating that punishment of anti-cooperative behaviour can make cooperation prevail. If punishing defectors is sufficiently cheap (i.e. if p^*_{max} is sufficiently high), the model displays an efficient, Pareto-optimal high equilibrium characterised by full cooperation, full trust and maximal (latent) punishment. Indeed, suppose that all players expect full cooperation, and expect all other players to expect the same. If $\pi_c > \pi_d - p^*_{max}$, then clearly the *ex ante* expected payoffs from cooperation are greater than the *ex ante* expected payoffs from defection, entailing that all players are then actually better off cooperating. This in turn rationalises players' expectations and makes full cooperation an equilibrium.

The third "intermediate" equilibrium lends itself to different possible interpretations. Let $\hat{q} \in (0,1)$ denote the proportion of cooperators at such an equilibrium. First, since players are homogenous, $\hat{q}$ represents the probability of any player choosing cooperation in a symmetric mixed-strategy equilibrium. Second, $\hat{q}$ also represents the minimal trust level for a player to choose cooperation as a pure strategy. In that respect, $\hat{q}$ reflects the attraction power of the low equilibrium. Indeed, consider the following remark.

Remark 2. *Suppose that the game is played a finite number of times.*[12] *Suppose that players' initial trust is exogenous. Suppose finally that players update their trust level after observing the outcome of the game at the first period;*[13] *formally, suppose that* $q_{i,t} = q_{j,t} = q_{t-1},\ \forall t > 1$ *and* $\forall i, j$*, where* q_t *denotes the proportion of cooperators at period* t*. Then, a proportion of players of* at least $\hat{q}$ *must have an initial level of trust of* at least $\hat{q}$ *for the high equilibrium to be achieved in the second period and in every subsequent period.*

Remark 2 states that if the game is adapted to this simple dynamic setting, then $\hat{q}$ can be seen as the *basin of attraction* of the low equilibrium. The higher $\hat{q}$, the higher players' initial trust must be for cooperation to be sustainable dynamically. In that respect, Proposition 1 and Remark 2 highlight the potential "self-fulfilling" dimension of trust. *Ceteris paribus,* when trust is low, players tend to prefer defection to cooperation, which incidentally justifies the low level of trust. The opposite holds true too.

3.2 Heterogeneous Population

In this subsection, I relax the assumption of players' homogeneity. This assumption is stringent notably because the opportunity cost and/or the benefits of cooperation may vary across the population. There are at least three main reasons to think so:

1. Some individuals are naturally more inclined towards cooperation than others. For instance, if some individuals believe that cooperative behaviours are rewarded in the after-life while others don't, *ceteris paribus*, it is reasonable to expect the former to be more prone to cooperate than the latter;

2. What constitutes cooperation and how it should be performed is context-specific. Individuals may learn this information at different costs. For instance, what constitutes a "fair trade practice" may vary across trading parties from different cultural or religious

[12] Note that for finitely repeated games, Proposition 1 holds at every period.

[13] This beliefs updating rule is deliberately simplistic and is only instrumental in expressing the intuition that the higher $\hat{q}$, the more trustful players will have to be to coordinate on the high equilibrium. Of course, more complex and realistic beliefs updating rules could be designed; this is however not the purpose of this discussion.

backgrounds. This may hinder cooperation between socially distant traders (see e.g. Leeson, 2008; Leeson and Coyne, 2012);

3. Cooperation may serve the interests of some individuals in particular within a society. In other words, some individuals may benefit more than others from cooperation.

The simplest way to capture this heterogeneity is to assume the presence of an additional, heterogeneous cost to cooperation for each player. Assume that this cost, denoted $\tau_i \in [0,\ \tau_{max}]$, is private information and drawn according to a continuous, twice-differentiable distribution function with cumulative distribution denoted $\Phi : [0,\ \tau_{max}] \to [0,1]$.[14] The final payoffs matrix for player i (12) can be modified as follows to account for players' heterogeneity:

$$\begin{array}{lll} & s_j = C & s_j = D \\ s_i = C & \pi_c - \tau_i & -\gamma \tilde{Q}_i\left(q_i\right)\left[\pi_d - p^*\left(q_i\right)\right] - c\left[p^*\left(q_i\right)\right] - \tau_i \\ s_i = D & \pi_d - E_i\left[p^*\left(\tilde{q}_j\right)\right] & 0 \end{array} \tag{13}$$

Note that $\tilde{Q}\left(q_i\right)$, player i's disappointed expectations, remains unaffected by τ_i. Optimal punishment thus also remains unchanged. Finally, the definition of equilibrium remains the same, with the exception that players' expected utility now displays the additional argument τ_i.

Proposition 1* is analogous to Proposition 1 but generalises it to account for players' individual cost of cooperation.

Proposition 1*.

1. *If and only if* Φ *is weakly convex and:*

   (a) $\pi_c - \tau_{max} < \pi_d - p^*_{max}$, *then there exists a unique low equilibrium characterised by full defection (i.e.* $q^* = 0$*);*[15]

[14] In the literature on this economics of religion, this cost is usually taken as reflecting players' individual "religiosity" or beliefs in after-life rewards or punishments, which in turn influence their individual cooperativeness. Note that I consider only cases with $\tau_i \geq 0$. Hence, the individual cost to cooperation is always a *cost*, strictly speaking. I do not consider cases where faith would, for instance, instill beliefs of supranatural *rewards* in case of cooperation, which would translate into additional *positive* payoffs to cooperation. The implicit assumption here is that even players with stronger religious beliefs (and thus with lower τ_i) require some cooperation enforcement to opt for cooperation. In contrast, Levy and Razin (forthcoming, 2012) consider in their models that religiosity *per se* is sufficient to induce certain players to cooperate. Such assumption, which in my model would amount to the inequality $\pi_c - \tau_i > \pi_d$ for some τ_i, would prevent the realisation of a "low" equilibrium.

[15] As explained in the Appendices, it is possible to have 3 equilibria (one low equilibrium and two intermediate equilibria) with this condition when $\pi_d - p^*_{max} < 0$ for some q. However, the resulting higher intermediate equilibrium always yields negative payoffs to all players. Such an equilibrium is not very interesting. In particular, to impose $\pi_d - p^*_{max} \geq 0$ sufficient to ensure that this condition yields a unique low equilibrium.

(b) $\pi_c - \tau_{max} \geq \pi_d - p^*_{max}$, *then there exists three equilibria, namely a low equilibrium* ($q^* = 0$,), *a high equilibrium* ($q^* = 1$), *and an intermediate equilibrium* ($q^* \in (0,1)$).

2. *If* Φ *is not weakly convex and:*

(a) $\pi_c - \tau_{max} < \pi_d - p^*_{max}$, *then there exists one low equilibrium* ($q^* = 0$), *never displays a high equilibrium* ($q^* = 1$) *but may display an even number of intermediate equilibria* ($q^* \in (0,1)$);

(b) $\pi_c - \tau_{max} \geq \pi_d - p^*_{max}$, *then there exists* at least *3 equilibria, namely one low equilibrium* ($q^* = 0$,), *one high equilibrium* ($q^* = 1$), *and one intermediate equilibrium* ($q^* \in (0,1)$).

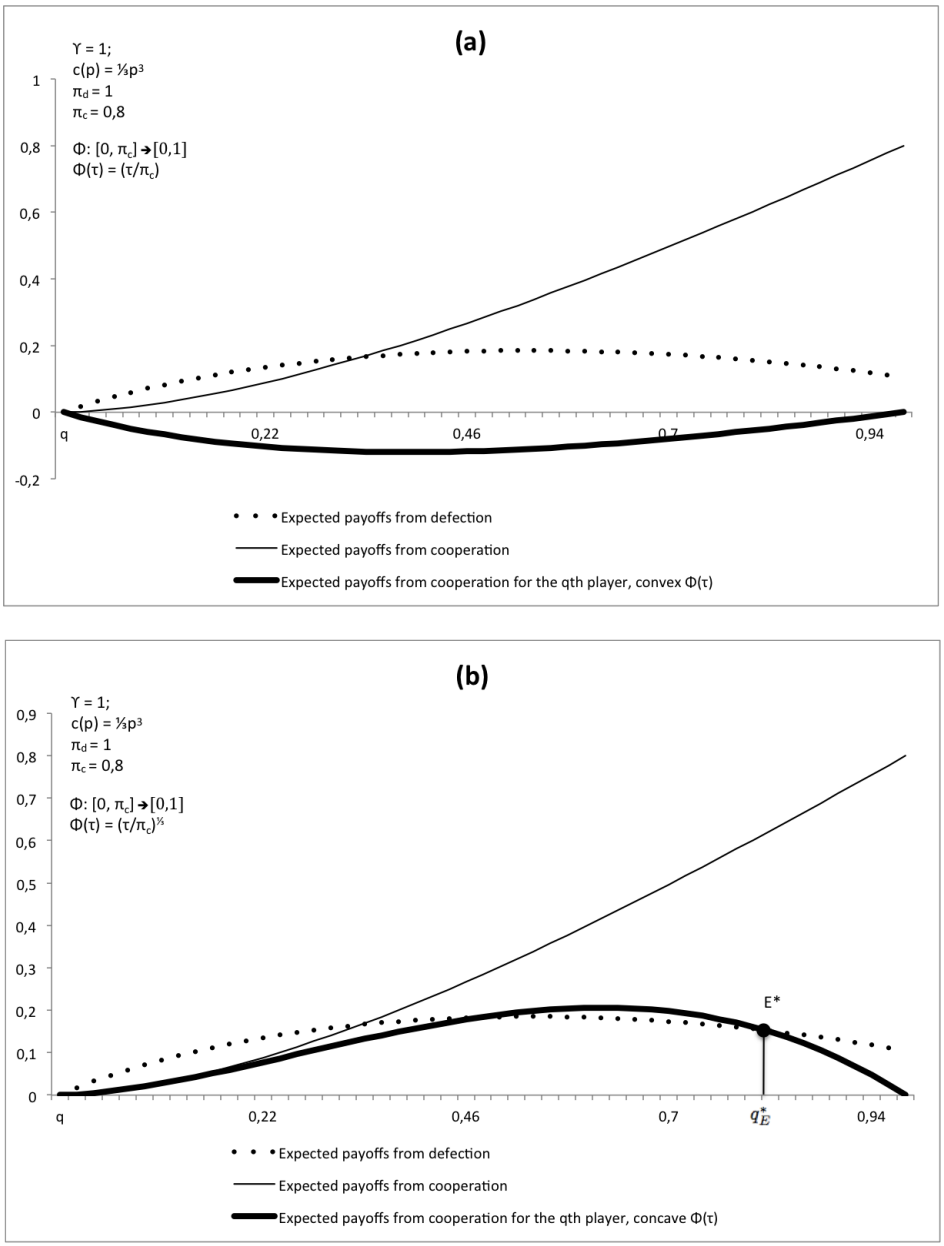
Equilibria when Φ is weakly convex and not weakly convex are illustrated on Figures 2 (a) and (b), respectively. The two graphs plot (i) the curve of "expected payoffs from cooperation with homogeneous players"; (ii) the curve of "expected payoffs from defection"; and (iii) the curve of "expected payoffs from cooperation *for the* q^{th} *player*" when players are ordered in increasing order of individual cost of cooperation (τ_i). Curve (iii) is constructed by subtracting $\Phi^{-1}(.)$ from curve (i).With heterogeneous players, an equilibrium occurs whenever curve (ii) crosses curve (iii).

Proposition 1* states that when Φ is weakly convex, Proposition 1 holds with a minor change to the condition for the existence of a high and an intermediate equilibria (the condition must hold *for the player with the highest* τ_i). Graphically, curves (ii) and (iii) have then a similar shape. When Φ is not weakly convex, however, results may change substantially. Stable, interior equilibria may exist as shown by point E^* on Figure 2 (b). At equilibrium E^*, a fraction q^*_E of players (those with the lowest individual cooperation cost) cooperate, while the rest defect. Finally, little else can be said without a more specified form for Φ when it is not weakly convex, except that a low equilibrium again always exist and that a high equilibrium exists whenever cooperation is more profitable than defection for the individual with the highest individual cooperation cost when $q = 1$ (i.e. when $\pi_c - \tau_{max} > \pi_d - p^*_{max}$).

3.3 Discussion: Trust, Punishment and Cooperation

This paper relates to the literature on the enforcement of cooperative norms in large scale societies. In particular, it is closely linked to the theoretical work in evolutionary game theory investigating the role of altruistic punishment in cooperation (Henrich, 2004; Boyd

Figure 2: Equilibria with Heterogeneous Players



et al., 2003; Fehr and Gächter, 2002). These models typically build on a cooperative game (e.g. prisoners' dilemma, public good game) played by different types of players, namely "selfish players", "cooperative players" and "cooperative altruistic punishers". In absence of the latter, selfish players always enjoy greater payoffs than cooperative players. Since players with higher fitness (measured by payoffs) reproduce more, the "selfish player" type thus takes over the whole population in the long run. However, the appearance of altruistic punishers in the population can change the structure of payoffs in favor of cooperators. Indeed, if punishment incurs high enough costs on defectors, then selfish players' fitness falls below

that of cooperators. The proportion of cooperators increases over time as a result, which in turn fosters altruistic punishers' fitness relative to selfish players. Hence, under certain parametric conditions, these models show that an altruistic punisher type can proliferate and sustain cooperation in large populations where interactions are not repeated.

Models of altruistic punishment typically overlook agents' "rational" motivations to punish defectors as they simply assume that the taste to punish is intrinsic to (some) agents' preferences[16]. Indeed, altruistic punishers are deemed to be biologically wired to punish selfish players. They are truly "altruistic" in that they passively accept to bear the costs of a public good (punishment) while not individually benefiting from it. Hence, the altruistic punishment approach is difficult to conciliate with rational behaviour. In particular, it ignores important facets of agents' decision to punish and the mechanisms underlying it, such as negative emotions and trust.

This paper reconciles costly punishment in one-shot interactions with rational behaviour. It introduces a model where agents may punish their coplayer to reduce the negative emotions they feel upon being "cheated on". The intensity of cooperators' negative emotions is increasing in their expectations about the proportion of cooperators in the population: the more they trust their coplayers, the more intense their negative emotions when they are cheated on, and the harder they punish. At equilibrium, cooperation can be sustained by players' expectations (trust), which ensure a level of cooperation enforcement (punishment) large enough for players to opt for cooperation.

A direct implication of the model is that punishment, trust and cooperation should in general be positively correlated. In this regard, Balliet and Van Lange (2013) present a quantitative meta-analysis of 83 studies in 18 different societies bearing on the links between cooperation, trust and punishment. They define "trust" as individuals' beliefs about others' benevolence and propensity to cooperate (e.g. contribute to a public good, "cooperate" in prisoners' dilemma), which is in direct line with the definition of trust offered in Definition 1. They report empirical evidence strongly supporting the implications of the model:

> [T]he present findings unpack the puzzle of punishment even further by providing novel support for the perspective that societal levels of trust and the enforcement of social norms are mutually reinforcing... The present research provides evidence that effective norm enforcement for cooperative behavior, which results in greater

[16] Gintis and colleages' (2001) evolutionary game theoretic model of punishment as a signal of individual quality is a notable exception.

> [cooperation], positively relates to a society's level of trust and norms of cooperation. (373-74).

Players' trust is key to making cooperation possible as players cooperate only insofar as they expect their coplayers to cooperate too (Fischbacher et al., 2001). The more individuals trust those they interact with, the more confident they are that defectors will be punished, which reduces their expected payoffs to defection compared to the payoffs to cooperation.

Lastly, trust, cooperation and the effective enforcement of norms can better be approached as interrelated dimensions of social capital. While "social capital" has lent itself to a myriad of uses and definitions since it became a commonplace concept in social sciences in the 1990s, these three dimensions seem indeed to be largely agreed upon. Bowles and Gintis (2002, in Durlauf and Fafchamps, 2005: 1643), for instance, define social capital as referring to "trust... a willingness to live by the norms of one's community and to punish those who do not." Ostrom (2000:176) adopts a similar definition: "social capital is the shared knowledge, understandings, norms, rules and expectations about patterns of interactions that groups of individuals bring to a recurrent activity" (see Durlauf and Fafchamps, 2005, for an extensive review). In that respect, the model presented in this paper can be seen as the first attempt to formally integrate these three elements of social capital in a single framework.

4 Religion, Trust, and Cooperation

4.1 Religious Groups

This section addresses how certain institutions, such as religious groups, may arise endogenously and raise trust and cooperation within a population. To address this question, I build on the example of a religious group. This example fits the features of the model particularly well. First, religious groups are typically large-scale organisations in which members meet infrequently or not frequently enough for "reputation-sensitivity... [to be] sufficient to explain the features of strong prosocial tendencies" (Norenzayan and Shariff, 2008: 58)[17]. Second, the enforcement of cooperation through emotion-based punishment (Stylized Facts A and B) also fits well the nature of religious groups. Indeed, "religions encourage compliance with codes of conduct" and systems of beliefs that are conducive of cooperation through promises

[17] In fact, it is commonly viewed that religion may have evolved precisely to permit the rise of stable and large societies and the prevalence of cooperation in ephemeral exchanges (see Bulbulia et al., 2008 for an extensive review).

of supernatural sanctions and rewards (McBride and Richardson, 2012: 123). These supernatural payoffs "are used to support and strengthen material punishment" and to construct complex social systems in which cooperation is possible (Sosis, 2005: 19). Furthermore, the very emotions involved in norms enforcement are reflected and made salient in the content of religious rituals and practices. Indeed, "in societies lacking a central political authority... intense and negatively valenced religious rituals address the inherent free-rider problems of collective action" as they provide a "reliable emotionally anchored mechanism for the subordination of immediate individual interest to cooperative group goals" (Sosis and Alcorta 2004: 339). Incidentally, an almost universally recurrent characteristic of gods or supernatural beings is their propensity to feel angry and manifest their wrath in the face of disloyalty and disobedience. Even though religious laws and codes of conduct often aim at controlling anger as a potentially destructive force, anger directed at the upholding of social order is often viewed as desirable and justified (Potegal and Novaco, 2010). Hence, religions groups form a natural and intuitive application of the model.

4.2 The Model with a Religious Group

As discussed earlier, players face a coordination problem whenever a high equilibrium and a low equilibrium coexist. The question is thus whether religion can operate as a mechanism ensuring the realisation of the high equilibrium. To tackle this question, assume that without religion, players coordinate on the low equilibrium. In other words, witout religion, players have a utility of zero.[18]

Suppose that the two game stages introduced earlier are preceded by a primitive stage. During this primitive stage, assume without loss of generality that a player, denoted player L, can decide to form a religious group. She also decides of a non-negative cost of religious requirements, denoted r, that any player choosing to become a member of L's religious group has to bear. Hence, during this first stage, all players choose a *membership strategy* $m_i \in \mathcal{M}_i = \mathcal{M}_j = \{M,\ N\}\ \forall i,\ j \in [0,1]$, where M and N stand for "membership" and "non-membership", respectively. Players choosing $m_i = M$ shall hereinafter be referred to as "members" and players choosing $m_i = N$, as "seculars". As soon as one or more players join L's religious group, all members pay the cost r and then restrict their second-stage interactions to the in-group. Since players coordinate on the low equilibrium when there is no religion, seculars have a reservation utility of 0. To study the equilibria of the modified game, I adapt the definition of an equilibrium as follows.

[18] This assumption may also reflect the "risk dominance" of the low equilibrium over the high equilibrium. Coordination may be increasingly risky and difficult to achieve as societies grow, rendering "no coordination" equilibria more attracting.

Definition 3. *A strategy profile* $s^* = \left[\left(m_i^*,\ s_i^*,\ p_i^*\right),\ \left(m_{-i}^*,\ s_{-i}^*,\ p_{-i}^*\right)\right]$ *is an equilibrium if,* $\forall i$ *and* $s_i^* \neq s_i',\ \forall s_i^*,\ s_i' \in S_i,\ m_i^* \neq m_i',\ \forall m_i^*,\ m_i' \in \mathcal{M}_i$, *and* $\forall \tau_i \in [0,\ \tau_{max}]$*:*

1. $p^*\left(q_i\right)$ *obeys equation (10);*
2. $E_i\left[u_i\left(m_i,\ m_{-i},\ s_i^*,\ \tilde{s_{-i}},\ p_i^*,\ \tilde{p_{-i}^*},\ q_i,\ \tilde{q_{-i}}|m_i,\ m_{-i},\ \tau_i\right)\right] \geq$

$$E_i\left[u_i\left(m_i,\ m_{-i},\ s_i',\ \tilde{s_{-i}},\ p_i^*,\ \tilde{p_{-i}^*},\ q_i,\ \tilde{q_{-i}}|m_i,\ m_{-i},\ \tau_i\right)\right]$$

3. $E_i\left[u_i\left(m_i^*,\ s_i^*,\ \tilde{s_{-i}^*},\ p_i^*,\ \tilde{p_{-i}^*},\ q_i,\ \tilde{q_{-i}},\ \tau_i\right)\right] \geq E_i\left[u_i\left(m_i',\ s^*{}_i,\ \tilde{s_{-i}^*},\ p_i^*,\ \tilde{p_{-i}^*},\ q_i,\ \tilde{q_{-i}},\ \tau_i\right)\right]$*;*
4. $q_i = q_{-i} = E_i\left(\tilde{q_{-i}}\right) = E_{-i}\left(\tilde{q}_i\right) = q^*$*;*

where q^* *corresponds to the actual proportion of cooperators within the religious group.*

Definition 3 is analogous to Definition 2 but states in addition that players' action strategy in the second stage must maximise their expected utility conditional on players' membership strategy. In turn, players' optimal membership strategy must be consistent with backward induction.

4.3 Religion with Homogeneous Players

The question addressed now is whether a religious group, as described above, can enhance trust and cooperation. To achieve higher cooperation, religion must somehow deter defection. Let $\Pi_d\left(q\right)$ denote the expected utility from defection for a level of q when $q_i = E_i\left(q_j\right) = q,\ \forall i,\ j$. Note that $\Pi_d\left(q\right)$ is represented by the curve of "expected payoffs from defection" on Figure 1 and can be written at length as $\Pi_d\left(q\right) = q\left(\pi_d - p^*\left(q\right)\right)$. Suppose first that players are homogeneous, as in Section 3.1. Consider the following definition.

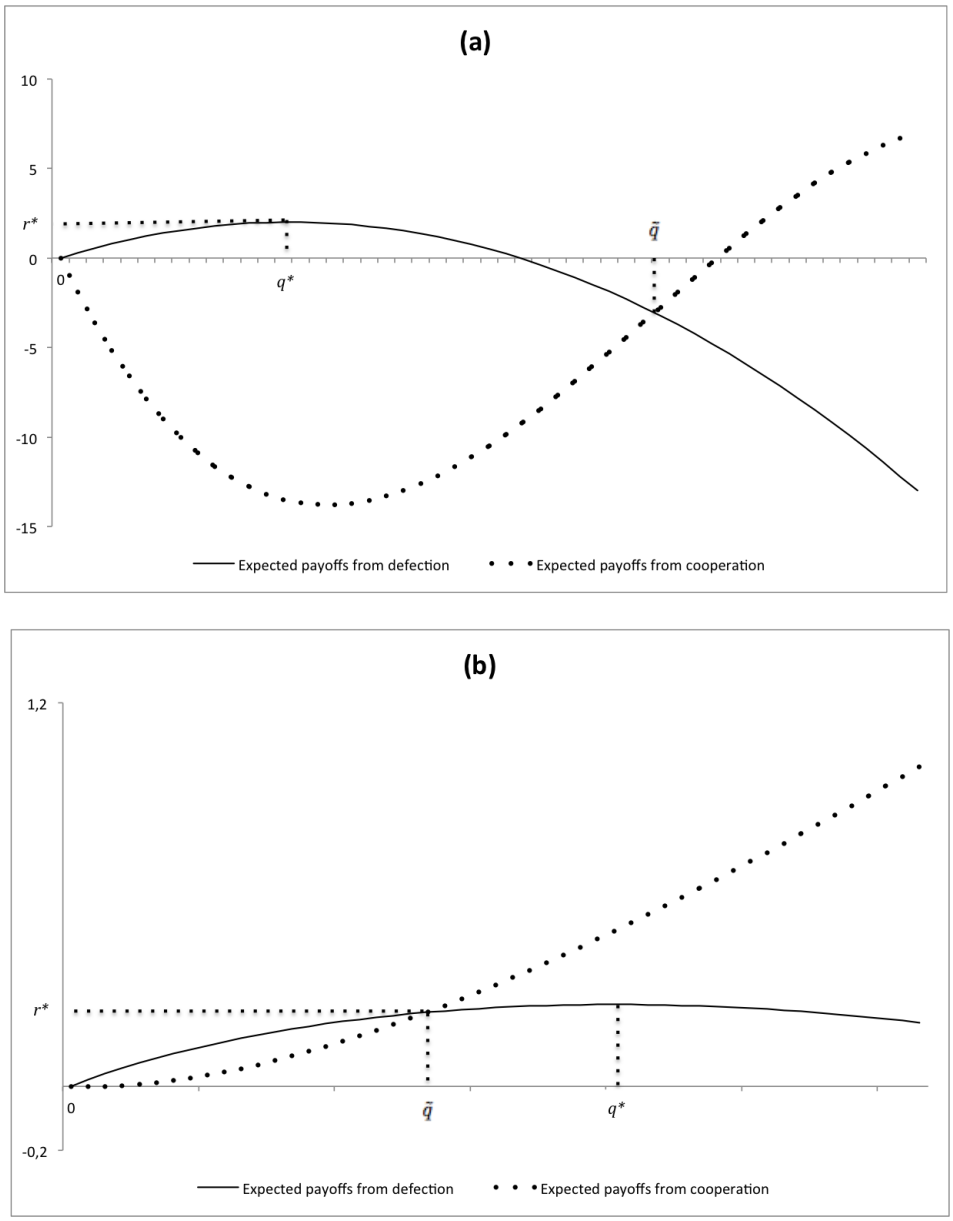
Definition 4. *Denote by* r^* the minimum cost of religious requirements ensuring coordination, *written as*:

$$r^* = \begin{cases} \Pi_d\left(\hat{q}\right) & if\ \hat{q} < q^* \\ \Pi_d\left(q^*\right) & otherwise \end{cases}$$

with $\tilde{q}$*, the proportion of cooperators at the intermediate equilibrium and* $q^* = \arg\max_q \left\{\Pi_d\left(q\right)\right\}$.

Figure 3 (a) and (b) provide graphical illustrations of $r^*$when $\hat{q} > q^*$ and when $\hat{q} < q^*$, respectively. The intuition as to why r^* can achieve full cooperation is straightforward.

Figure 3: Minimum Cost of Religious Requirements Ensuring Coordination



Suppose first, as on Figure 3 (a), that $\hat{q} > q^*$. With a cost of religious requirements of $\Pi_d(q^*)$, a member clearly cannot achieve positive payoffs is she chooses to defect. Indeed, defection can yield payoffs of *at most* $\Pi_d(q^*)$. A player choosing "membership and defection" would thus always be better off remaining secular. *Ex ante*, all players know that and can thus infer that if they join the religious group, they will not be matched with a defector. Second, suppose that $\hat{q} < q^*$. Then, the only way for a member to achieve net positive expected payoffs by defecting is if she expects $q > \hat{q}$. However, as it is clear on Figure 3 (b), if she expects $q > \hat{q}$, then she will strictly prefer to cooperate. Again, all players know that

fact *ex ante* and will thus infer that once in the religious group, they will not be matched with a defector.

Let us now go back to player L's decision of whether or not to form a religious group. If she does so, clearly she won't choose a level of requirements different than r^*. Indeed, requirements cheaper than r^* would not be sufficient to coordinate beliefs and ensure full cooperation; the low equilibrium would be effectively replaced by an equilibrium of full secularity, with no impact on players' final payoffs. In contrast, requirements greater than r^* would only be more costly without achieving anything more than r^*. Therefore, player L's decision boils down to determining whether the benefits of full cooperation are worth the cost of the religious group born by each member, r^*. Hence the following proposition.

Proposition 2.

1. *If and only if* $\pi_c \leq r^*$, *then no religious group arises endogenously within the population.*
2. *If and only if*$\pi_c > r^*$, *then a religious group arises endogenously with membership cost* r^*. *Full membership and full cooperation then constitute an equilibrium in weakly dominant strategies.*

Proposition 2 implies that if the benefits to cooperation are high enough, then a religious group with costly religious requirements will emerge endogenously and achieve the coordination of players' trust. This makes "membership and cooperation" an equilibrium in (weakly) dominant strategies.[19] If we assume that player L forms the religious group if she is indifferent, then secularity is strictly dominated for all players by membership followed by cooperation. In that respect, religion arises endogenously whenever $\pi_c > r^*$.

4.4 Religion with Heterogeneous Players

Let us now turn our attention to the case where players are heterogeneous. Consider first the following amendment to Definition 4.

Definition 4*. *Denote by* $\hat{r^*}$ the minimum cost of religious requirements ensuring coordination *with heterogeneous players, defined as follows:*

[19] Indeed, "full non-membership" is also an equilibrium, although weakly dominated by "full membership and full cooperation."

$$\hat{r^*} = \begin{cases} \Pi_d(\hat{q}) & if\ \exists\ \hat{q} < q^* :\ \Pi_{c,\,q}(q) > \Pi_d(q)\,,\ \forall q \in (\hat{q}, 1] \\ \Pi_d(q^*) & otherwise \end{cases}$$

with $\hat{q}$, the proportion of cooperators at any intermediate equilibrium and $\Pi_{c,\,q}(q)$, the expected payoffs from cooperation for the q^{th} player.

Definition 4* is a generalisation of Definition 4 to account for the possibility of interior stable equilibria (i.e. where curve (iii) crosses curve (i) from above). Indeed, if there is such an equilibrium $\hat{q}$ with $\hat{q} < q^*$, then $\Pi_d(\hat{q})$ is not sufficient to deter free-riders from entering the group. This is only true for interior *stable* equilibria, which were not possible with homogeneous players.

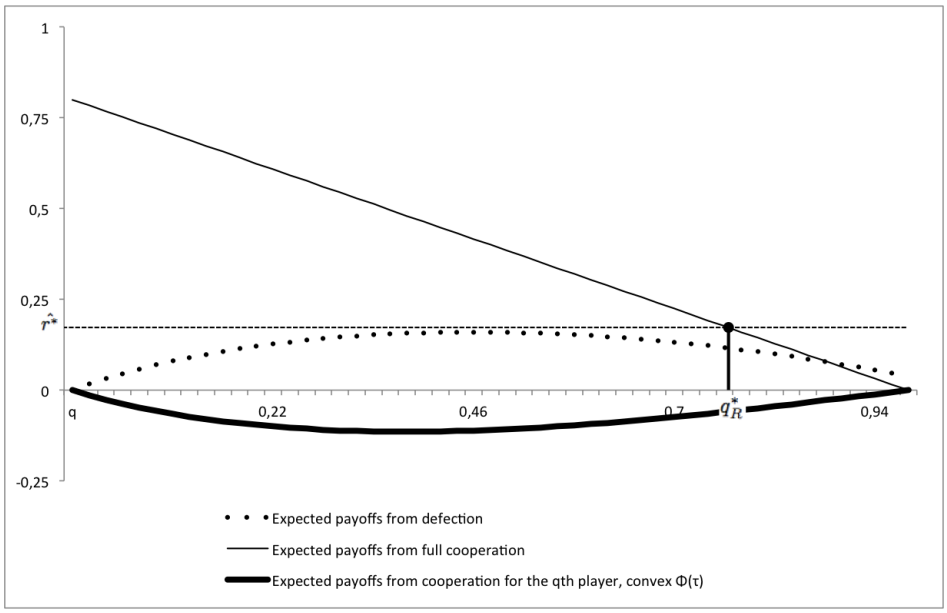
Proposition 2*.

1. *If and only if $\pi_c \leq \hat{r^*}$, then no religious group arises endogenously within the population.*
2. *If and only if $\pi_c > \hat{r^*}$, then a religious group arises endogenously with membership cost r^*. Membership and cooperation constitute an equilibrium in weakly dominant strategies for all players with $\tau_i \leq \pi_c - \hat{r^*}$, while non-membership is a strictly dominant strategy for all players with $\tau_i > \pi_c - \hat{r^*}$.*

Proposition 2*.1 mirrors Proposition 2.1. It states that if $\pi_c \leq \hat{r^*}$, then even the player the most inclined towards cooperation finds the cost of religious requirements too high for the benefits achieved. In such case, no religious group emerges. In contrast, if $\pi_c > \hat{r^*}$, then at least some players find the benefits of a religious group to be greater than its cost. Recall that when players are homogeneous, if group membership is profitable to *one* player, then it is for *all* (Proposition 2.2); this is however no longer the case with heterogeneity.

When players are heterogeneous, a religious group may thus play an additional *signalling role* by excluding the least cooperative individuals. This signalling role can better be seen on Figure 4, which presents a case where the *only* possible equilibrium without religious group when players are heterogeneous is full defection. It is so because all players know that free-riders always undermine the group to such an extent that cooperation is never a dominant strategy for any player. In such case, $\hat{r^*}$ operates first as an effective signalling device. Only players for whom cooperation is a dominant strategy when $q = 1$ (players at the left of q_R^*) are potentially willing to pay that cost, which *de facto* excludes the fraction $1 - q_R^*$ consisting of the least cooperative players. Second, $\hat{r^*}$ operates as a coordination device in the same way as with homogeneous players: the only rational reason for any player to bear the cost

Figure 4: The Signalling Role of Religion



$\hat{r}^*$ is if she expects a level of cooperation such that she is better off cooperating. This makes full cooperation the only possible outcome within the religious group, which is profitable for any player i with $\tau_i \leq \pi_c - \hat{r}^*$.

4.5 Discussion

The results of this section relate to those obtained in the literature on the economics of religion addressing the role of religious organisations in fostering in-group cooperation. Theoretical work in that field has primarily built on signalling models (see e.g. Levy and Razin, 2012, forthcoming; Berman, 2000; Iannacconne, 1992, 1994). Iannacconne (1992, 1994) famously pioneered the application of signalling theory to religious organisations to explain how costly religious requirements could increase contributions to a club good. In a seminal paper, he (1992) showed that such organisations producing club goods could tie membership to costly and easily recognisable signals in order to screen for and exclude potential free-riders. He showed that if such signals are costly enough, then a religious organisation may be able to retain only individuals with a relative advantage in the production of the club good, thereby incerasing its quality and, in turn, the welfare of its members.

In a similar vein, Levy and Razin (2012, forthcoming) propose a model in which players are randomly matched to play a symmetric PD game. In addition, players may join a religious organisation. Members of the religious organisation bear the exogenously given cost

of religious rituals and requirements. Religious participation also endows players with beliefs about the probability of a utility shock, which can be either positive or negative. This utility shock is exogenous and influences only players' *ex ante* expected utility (e.g. divine rewards or punishments). A religious player perceives the probability of a negative shock when she defects as higher than the probability of such shock when she cooperates. However, her perception of the difference between these two probabilities depends on her type, which is drawn from an exogenous random distribution. This type reflects a player's propensity to be influenced by religious participation. In that setting, costly religious requirements function as a signalling device as only the more easily influenced players partake in the religious group, and these players are more inclined to cooperate due to their beliefs.[20]

The key proposition of signalling models of religion is that religious groups, through costly religious requirements, achieve higher cooperation through the separation of different types of individuals. Indeed, these studies typically rely on the existence of two types of agents, namely "pious agents" and "defectors". As the former are *assumed* to be disposed to cooperate naturally, costly signals ensure the exclusion of the latter to enhance cooperation amongst the pious. This feature of signalling models is problematic for two reasons. First, it assumes away the puzzle of cooperation for a fraction of the population, suggesting that religious beliefs are sufficient to account for cooperation. However, Bulbulia (2012:15) surveys ample evidence that religious beliefs are "neither necessary nor sufficient to assure [cooperation]." Hence, what makes cooperation possible in the first place cannot be explained by signalling theory alone. Second, the hypothesis that beliefs *per se* induce cooperative behaviours overlooks the normative mechanisms underlying the enforcement of cooperation. Humans' capacity to enforce cooperative norms through punishment is yet widely regarded as crucial in the evolution and sustainability of cooperation throughout human history (see Dubreuil, 2010a,b, for an extensive review of this argument). Consequently, signalling models of religion have so far *de facto* ignored the interaction between religion, cooperative norms and their enforcement.

[20] Note that in Levy and Razin's model, exogenous religious requirements have an impact on cooperation only insofar as religion instils exogenous, prosocial beliefs in players. In my model, in contrast, both these requirements and players' beliefs are *endogenous*. In particular, players' beliefs reflect players' expectations about their social environment, not about hypothetical divine rewards and punishment. Even though the settings are similar, Levy and Razin's model tackles primarily the link between religious beliefs and cooperation while I focus on the link between religious practice, social norms and enforcement of cooperation. The models thus complement one another.

The model introduced in this paper extends the signalling approach to religion. It first presents religious organisations as devices coordinating players' trust. In turn, trust reinforces the mechanisms of cooperation enforcement (punishment). Players heterogeneity is not necessary for a religious organisation to fulfill this coordination role. Finally, when players are heterogeneous, a religious group may in addition serve as a signalling device to exclude those who would never cooperate, which is in line with signalling models of religion. These roles (*coordination* and *signalling*) are formalised in Propositions 2 and 2*.

A direct implication of these propositions is that *ceteris paribus*, costlier religious requirements should ensure the achievement of higher levels of in-group trust and cooperation. Sosis (2000) and Sosis and Bressler (2003) provide evidence consistent with this implication using data on secular and religious communes the XIXth century U.S.A. Since religion is excepted to increase cooperation, religious communes are expected to survive longer than their non-religious counterparts. Sosis (2000) finds that religious communities were incidentally significantly more likely to outlast secular ones at every stage of their life (more than 4 times more likely). Sosis and Bressler (2003) find that religious communes demanded more than twice as many costly requirements[21] to their members compared to non-religious communes, and that these requirements were significantly and positively correlated to communes' lifespan. These findings "imply... that the greater longevity of religious communes with costlier requirements [is] due to greater intragroup cooperation and trust levels" (Norenzayan and Shariff, 2008: 61). This hypothesis is consistent with the model's implications.

Another implication of Propositions 2 and 2* is that members of religious groups should, *ceteris paribus*, be more cooperative, trusting and trusted than non-members, at least with coreligionists. Empirical evidence is consistent with this proposition.[22] Indeed, religious individuals are persistently "perceived to be more trustworthy and more cooperative" by others, and sociological surveys suggest that "individuals who report stronger [religious] beliefs... have stronger altruistic tendencies" (Norenzayan and Shariff, 2008: 59-60). Membership to a religious organisation has been found to foster in-group trust between coreligionists (see Sosis, 2004, for a review). Furthermore, participation in world religions has been found to be significantly correlated with increased cooperation in ultimatum games and dictator games, and is also correlated with punishment in third-party punishment games (Henrich et al., 2010). These findings altogether "support the notion that religion may have coevolved with complex societies to facilitate larger-scale interactions", notably through the fostering

[21] Defined either as (i) behaviours required by the commune and entailing energetic, time and/or financial cost; or (ii) as restricted behaviours (e.g. alcohol consumption) (Sosis and Bressler, 2003: 219).

[22] Note that the model is silent with respect to how members would interact with secular players. Indeed, it rests on the assumption that members restrict their interactions to the in-group. However, if we assume that members would act, or would be more likely to act, with secular players "as if" they were at the in-group equilibrium, then this evidence is directly consistent with the implications of the model.

of trust and the enforcement of cooperative norms (Henrich et al., 2010: 1481). The model is in direct line with these findings.

5 Application: Analysing the Size of Religion

5.1 Religion Participation and Affiliation

A useful feature of the model introduced in this paper is that it permits an analysis of the factors influencing both *religious affiliation*, defined as the proportion of players who partake in the religious groups (q_R^*), and *religious participation*, defined as the costliness of religious requirements and measured directly by $\hat{r^*}$, in a society.[23] Proposition 3 formalises some of the predictions stemming from the model in that regard.

Proposition 3.

Religious affiliation is decreasing in the material benefits to defection, π_d , and in the cost of punishment,[24] *$c(p)$, and is increasing in the material benefits to cooperation, π_c. Religious participation is decreasing in π_c and increasing in π_d and in $c(p)$.*

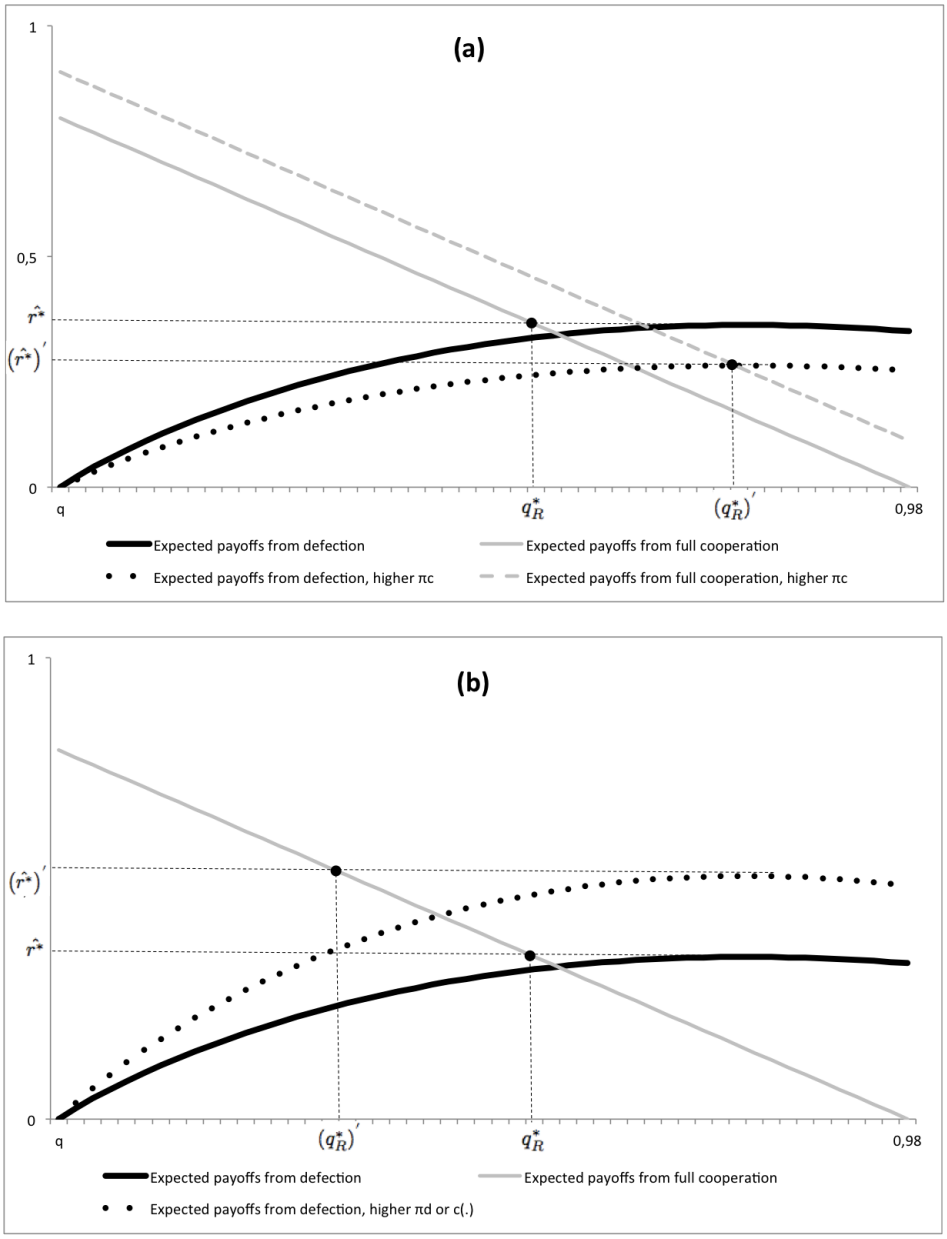
Proposition 3 proposes an integrated theory of the factors influencing the "size" of religion within a society. Figure 5 is useful to understand the mechanisms underlying Proposition 3. First, an increase in π_c pushes the "expected payoffs from full cooperation" curve up and the "expected payoffs from defection" curve down, as shown on Figure 5 (a). As the payoffs to cooperation increase, cooperators' negative emotions in the face of defection also increase, which translates into higher punishment (and thus smaller net payoffs to defection). The combined effects of the shifting of these two curves results in an increase in religious affiliation (from q_R^* to $(q_R^*)'$) and a decrease in religious participation (from $\hat{r^*}$ to $(\hat{r^*})'$). Second, an increase in π_d shifts upwards the "expected payoffs from defection" curve, as shown on Figure 5 (b). This entails an increase in religious participation (from $\hat{r^*}$ to $(\hat{r^*})'$), which in turn provokes a decrease in religious affiliation (from q_R^* to $(q_R^*)'$). An increase in the cost of punishment would have a similar effect as it would diminish cooperators' optimal punishment, which in turn would shift upwards the "expected payoffs from defection" curve. Aother implication of Proposition 3 is that religious affiliation and participation should be inversely correlated. Indeed, *ceteris paribus*, an increase in religious participation always entail a decrease in religious affiliation.[25]

[23] To simplify this analysis, I will restrict my attention without loss of intuition to the cases where, at equilibrium, some players partake in the religious group while some others do not, that is whenever $\pi_c > \hat{r^*}$ and $\pi_c - \tau_{max} < \hat{r^*}$.

[24] By "increasing cost of punishment", I mean that any level of punishment chosen by a player entails a greater total cost of punishment. For instance, a technological improvement that would the change punishment cost function from $c(p) = 5p$ to $c(p) = 2p$ would entail a "decreasing cost of punishment".

[25] This is consistent with the empirical evidence showing that more demanding religious groups are typically: see e.g. Iannaccone, 1992. This result also echo that of Levy and Razin (2012), who find that religious groups that are more demanding in their rituals are smaller are composed of individuals who are more "extreme" in their beliefs.

Figure 5: Religious Affiliation and Participation



5.2 Evidence

The model predicts first that an increase in the material benefits to defection, π_d, will increase religious participation. This implication reflects the phenomenon that in hard times, individuals tend to increasingly turn to their religious community and increase their devotion:

> In cases of extreme hardship, where common-pool resource problems abound and

> the threats of defection are high, the expectation is for the cost outlays to intensify and to become more frequent. When the chips are down the religious will produce more effort and expend more resources proving their faith. (Bulbulia, 2004: 672)

In this regard, Chen (2010) provides empirical evidence from the Indonesian financial crisis of 1996-97. He finds that religious participation, measured as participation in religious rituals and Islamic school attendance for children, increased amongst most affected households. He interprets these findings as evidence that economic distress stimulates religious participation, at least for certain households. Furthermore, the model also predicts that an increase in π_d will decrease religious affiliation. Hence, an increase in π_d should entail a decrease in religious participation for players who benefit less from cooperation (i.e. those who are less exposed to economic risks).[26] This is consistent with Chen's findings that households "that suffer less economic distress significantly decrease religious intensity" (2010: 303).

Proposition 3 also states that religious participation and affiliation depend on the cost of cooperative norms enforcement. When such cost is low, cooperation is relatively easy to enforce, and high costs of religious requirements are not necessary to ensure players' trust and cooperation (and vice-versa). Various elements can influence the cost of norm enforcement. The emergence of the modern state, with the deployment of its institutional apparatus (e.g. bureaucracy, efficient police forces and justice systems), certainly lowered the cost of enforcing cooperation. For example, if an individual violates one's property, it is clearly easier for one to refer to the police than to carry justice oneself. The model predicts that the emergence of such institutions should accompany a diminution in religious participation. This proposition is consistent with the commonly held view of the phenomenon of secularization as a dimension of modernization, observed over the past centuries and decades in many parts of the world.[27] Max Weber, one of the earliest theorists of the modern states, held such view, here summarized by Habermas (1990: 2):

> What Weber depicted was not only the secularisation of Western culture, but also and especially the development of modern societies... marked by... the two functionally intermeshing systems that had taken shape around the organisational cores of the capitalist enterprise and the bureaucratic state apparatus. To the

[26] Indeed, Chen stresses the "collective insurance" dimension of the religious group for its members. Clearly, wealthier or less exposed households can be viewed as having a higher τ_i as they certainly benefit less from such collective scheme (in other words: their benefits to cooperation are lower than for other households).

[27] The concept of modernization certainly cannot be pinned down to the emergence of formal institutions of law enforcement. Hence, it might be difficult, if not impossible, to clearly isolate the effect of these institutions on religious participation and affiliation. I only report here the common view that both phenomena (decreasing religious participation and establishment of formal institutions of law enforcement) are related, which is indeed consistent with the model. Also note that the model predicts as well an increase in religious affiliation in that context. This prediction is consistent with evidence that "American church membership rates have risen throughout most of the past two centuries" (Iannacconne, 1998: 1468) alongside modernization, at least until recently. More empirical research should shed light on this relationship.

> degree that everyday life was affected by this cultural and societal rationalisation, traditional forms of life... were dissolved.

Intergroup mobility is another factor affecting the cost of punishing free-riders. Indeed, "when residents have few constraints limiting their ability to transfer to another group, the threat of punishment and social ostracism are less effective free-rider deterrents" (Sosis and Alcorta, 2004: 268). The model predicts that an increasing intergroup mobility will in turn increase religious participation and decrease religious affiliation. This is consistent with Sosis and Alcorta's (2004: 268) observation that "costly in-group requirements [are] more prevalent in communities characterized by... high intergroup mobility." It is thus not surprising to observe that religious organisations have typically sought to restrict such mobility, either through open conflict with their surroundings (McBride and Richardson, 2012) or through stigmas limiting their members' capacity to interact with the out-group (Iannacconne, 1992).

Another implication of Proposition 3 is that an increase (decrease) in the material benefits to cooperation,π_c, will increase (decrease) religious affiliation and decrease (increase) religious participation. The development of social safety nets can be interpreted at least partially in light of this prediction. Indeed, "cooperation" within a community often aims at ensuring security for its members, notably "in the face of uncertainty regarding the ability to meet future needs for food, clothing, and shelter" (Olson, 2011: 136; c.f. Norris and Inglehart, 2004; Chen, 2010). When such needs are addressed by formal safety nets, the benefits of community-based cooperation diminish. That said, the development of social safety nets can also be interpreted as increasing players' "outside option":[28] indeed, secular players may see their payoffs increase *relatively* to religious players' payoffs. Globally, the combined effect of a decrease in π_c and of an increase in players' outside option would be ambiguous with respect to religious participation (as a decrease in π_c increases $\hat{r^*}$ while an increase in the outside option has the opposite effect), but would definitely entail a decrease in religious affiliation. This prediction is consistent with the evidence that "public safety net interventions can dilute incentives to maintain... informal coping" groups since "with incomes thus smoothed, households may no longer have sufficient incentive to band with others to form private risk sharing arrangements" (Cox and Fafchamps, 2008: 3714-56). It is also strongly consistent with the empirical evidence provided by Norris and Inglehart's (2004) cross-national and longitudinal study of religion affiliation and belief. They show that religious affiliation declines with increasing safety net coverage, measured by the involvement of the state in providing health, disability and pension insurance. They find, in contrast, that high levels of poverty, violence and economic inequality foster religious affiliation, *ceteris paribus* (c.f. Olson, 2011).[29]

[28] In the model, this outside option is assumed for simplicity to be 0; it is however easy to relax this assumption and analyse the effect of an increase in the "payoffs to secularity". Clearly, the curves on Figure 4 would remain unchanged; the proportion of players partaking in the group would thus stay the same. However, the optimal cost of religious requirements would decrease by an amount equal to the increase in the payoffs to non-membership. Indeed, recall that $\hat{r^*}$ must be equal to the payoffs defectors renounce to by remaining secular. If secular players' payoffs increase, then the net cost of renunciation diminishes accordingly.

[29] The concomitant rise of the welfare state and the massive religious disaffection observed in the Western world in the second half of the XXth century onwards, for example, can also be interpreted in that light.

Finally, the model also proves useful to analysing the impact of the state's interventions in religions markets. In particular, the state may seek to penalize or to encourage religious participation. The model predicts in this regard that a *subsidy* to religious participation will leave religious affiliation unaffected but will increase religious participation,[30] while a state *penalty* would have the opposite effect. Berman (2000) found that state subsidies to ultra-orthodox groups in Israel increased religious participation of young men. Similarly, Barro and McCleary (2006) found that "state regulation of religion lowers religious participation" (McCleary, 2011: 18). Hence, the available empirical evidence is consistent with the predictions of the model.

6 Conclusion

This paper presents a novel view of the role of religion in fostering cooperation. It proposes a game theoretic model of religion as an institution arising endogenously to coordinate players' trust, which in turn assures the enforcement of cooperative norms. The model also shows that when players are heterogeneous, notably with respect to the strength of their religious beliefs, religion may also serve as a signalling device to exclude those who would never cooperate. Finally, the model enables clear and tractable predictions about the levels of religious affiliation and religious within a society.

Different interesting extensions to the model should be considered for further research. For example, work could be done to investigate the dynamic effects of different rules for updating players' beliefs. This approach would shed light on the evolution of the "size" of religion within a society and of the mechanisms underlying such evolution at a micro level. Second, the model in its present form does not allow for more than one religious group. Relaxing this assumption could give insight into inter-group behaviour.

7 Appendices

Proof to Proposition 1

Let $\Pi_d(q)$ ($\Pi_c(q)$) denote the expected final utility from defection (from cooperation) for a level of q when $q_i = E_i(q_j) = q,\ \forall i,\ j \in [0,1]$. An equilibrium obtains when $\Pi_d(q) = \Pi_c(q)$, or in cases of a corner solution when $q = 1$ and $\Pi_d(q) < \Pi_c(q)$.

[30] To see why, consider Figure 4. Under the assumptions that a state subsidy would leave π_c, π_d and the cost of punishment unchanged, the curves on Figure 4 would stay the same, leaving group affiliation unaffected. However, to remain an efficient free-riding deterrent and coordination mechanism, the cost of religious requirements would now have to take into account the subsidy that members receive from the state. The optimal cost of religious requirements would thus increase from r^* to $r^* + s$, where "s" would measure the size of the state subsidy. The reasoning for state penalties is analogous.

Lemma 1: $\Pi_d(q)$ ***is concave in*** q, $\forall q \in (0,1)$. *Proof: Let* $\pi_d - p^*(q)$ *be rewritten as* B. *Then,* $\Pi_d(q) = qB$. *We know from equation (10) that* $p^*(q)$ *is strictly increasing and concave in* q. *Hence,* $\frac{\partial B}{\partial q} = -(p^*)'(q) < 0$, *and* $\frac{\partial^2 B}{\partial q^2} = -(p^*)''(q) > 0$. *Hence,* B *is strictly decreasing and convex in* q. *Since* q *is naturally increasing and linear in* q, *then necessarily* qB *is concave.* ■

Consider now the two following cases.

Case 1: $\Pi_d(q) \geq 0 \ \forall q \in [0,1]$.
Case 2: $\Pi_d(q) \geq 0 \ \forall q \in [0,\bar{q}]$, and $< 0 \ \forall q \in (\bar{q},1]$

Since $\Pi_d(q)$ is concave $\forall q \in (0,1]$, Cases 1 and 2 are exhaustive. I will now show separately, for each of these two cases, that Proposition 1 must hold. Note first that the proof to the existence of the low equilibrium is trivial for both cases as $\Pi_c(0) = \Pi_d(0) = 0$ in both cases.

Case 1: $\Pi_d(q) \geq 0 \ \forall q \in [0,1]$.

Lemma 2: If Case 1 holds, then $\Pi_c(q)$***is convex in*** q, $\forall q \in (0,1]$. *Proof: Note that* $\Pi_c(q)$ *can be written at lenght as:*

$$\Pi_c(q) = q\pi_c + (1-q)\left[-\gamma q \pi_c (\pi_d - p^*(q)) - c[p^*(q)]\right] \tag{14}$$

The first derivative of expression (14) can be written as:

$$\frac{\partial \Pi_c(q)}{\partial q} = \pi_c + \gamma \pi_c q (\pi_d - p^*(q)) + c[p^*(q)]$$

$$(1-q)\left[-\gamma \pi_c (\pi_d - p^*(q)) + \gamma q \pi_c (p^*)'(q) - c'(p) \cdot (p^*)'(q)\right] \tag{15}$$

Using equation (10), equation (15) can be simplified as:

$$\frac{\partial \Pi_c(q)}{\partial q} = \pi_c + \gamma \pi_c q (\pi_d - p^*(q)) + c[p^*(q)] + (1-q)\left[-\gamma \pi_c (\pi_d - p^*(q))\right] \tag{16}$$

Using equation (16), the second derivative of expression (14) can be written as:

$$\frac{\partial^2 \Pi_c(q)}{\partial q^2} = 2\gamma \pi_c [\pi_d - p^*(q)] + (1-q)\gamma \pi_c \cdot (p^*)'(q) \tag{17}$$

which is always strictly positive when $\Pi_d(q) \geq 0$, *implying that* $\Pi_c(q)$ *is strictly convex on this domain.* ■

Using Lemmas 1 and 2, to prove Proposition 1 when Case 1 holds is straightforward.

Proof of $\pi_c < \pi_d - p^*_{max} \Rightarrow \exists$ *only one low equilibrium*:
By contradiction, suppose that there exists at least one other equilibrium, denoted $\check{q}$, with $\Pi_c(\check{q}) = \Pi_d(\check{q})$. Since $\Pi_c(0) = \Pi_d(0) = 0$ and $\Pi'_d(0) > \Pi'_c(0) > 0$, it must be the case, from Lemmas 1 and 2, that $\Pi_c(q) < \Pi_d(q) \ \forall q \in (0,\check{q})$ and $\Pi_c(q) > \Pi_d(q) \ \forall q \in (\check{q},1)$, which contradicts that $\pi_c < \pi_d - p^*_{max}$. ※

Proof of $\pi_c < \pi_d - p^*_{max} \Leftarrow \exists$ *only one low equilibrium:*
By contradiction, suppose that $\pi_c > \pi_d - p^*_{max}$. Then, clearly $q = 1$ forms an equilibrium, which is a contradiction. ※

Proof of $\pi_c > \pi_d - p^*_{max} \Rightarrow \exists$ *3 equilibria:*
The low and the high equilibria, in that case, are trivial. Also, since $\Pi_c(0) = \Pi_d(0) = 0$ and $\Pi'_d(0) > \Pi'_c(0) > 0$ but $\Pi_c(1) > \Pi_d(1) > 0$, then given Lemmas 1 and 2 and the fixed point theorem, $\exists \hat{q} :\ \Pi_c(\hat{q}) = \Pi_d(\hat{q})$, with necessarily that $\Pi'_c(\hat{q}) > \Pi'_d(\hat{q})$. Suppose that there is another equilibrium, denoted $\bar{q}$, for which $\Pi_c(\bar{q}) = \Pi_d(\bar{q})$. Suppose wlog that $\bar{q} > \hat{q}$. Then, it must be the case that for some $\ddot{q} \in (\hat{q}, \bar{q})$, $\Pi'_c(q) > \Pi'_d(q)\, \forall q \in [\hat{q}, \ddot{q}]$ while $\Pi'_c(q) < \Pi'_d(q)\, \forall q \in [\ddot{q}, \bar{q}]$, which contradicts either Lemma 1 or Lemma 2, or both. ※

Proof of $\pi_c > \pi_d - p^*_{max} \Leftarrow \exists$ *3 equilibria:*
By contradiction, suppose that there are 3 equilibria but $\pi_c \leq \pi_d - p^*_{max}$. Then, clearly, either Lemma 1 does not hold, or Lemma 2 does not hold, or neither hold, which is a contradiction. ※

This completes the proof of Proposition 1 for Case 1.

Case *2:* $\Pi_d(q) \geq 0\ \forall q \in [0, \bar{q}]$**, and** $< 0\ \forall q \in (\bar{q}, 1]$

With Case 2, we know that $\Pi_c(1) > 0 > \Pi_d(1)$. Hence, $q = 1$ is necessarily always an equilibrium. We also know that $\Pi_c(0) = \Pi_d(0) = 0$ and $\Pi'_d(0) > \Pi'_c(0) > 0$. Also, we know that $\Pi_c(\bar{q}) = \pi_c > \Pi_d(\bar{q}) = 0$. Therefore, by the fixed point theorem and as Lemmas 1 and 2 hold $\forall q \in (0, \bar{q}]$, there exists a unique $\hat{q} \in (0, \bar{q}) :\ \Pi_c(\hat{q}) = \Pi_d(\hat{q})$.

Proof of $\pi_c > \pi_d - p^*_{max} \Rightarrow \exists$ *3 equilibria:*
Suppose by contradiction that there are not 3 equilibria. Since we've already determined the necessary existence of at least 3 equilibria (namely $q = 0$, $q = 1$ and $q = \hat{q}$), then there must be more than 3 equilibria for this statement to hold. Suppose there is at least a fourth equilibrium, denoted $\check{q}$. As Lemmas and 2 hold $\forall q \in (0, \bar{q}]$, it must be true that $\check{q} \in (\bar{q}, 1]$, with $\Pi_c(\check{q}) = \Pi_d(\check{q})$. However, we know that $\Pi_c(q) > 0\ \ \forall q \in (\bar{q}, 1]$, while $\Pi_d(q) < 0\ \forall q \in (\bar{q}, 1]$, which is a contradiction. ※

Proof of $\pi_c > \pi_d - p^*_{max} \Leftarrow \exists$ *3 equilibria:*
Suppose by contradiction that there are 3 equilibria but $\pi_c \leq \pi_d - p^*_{max}$. By the definition of Case 2, this condition can be rewritten $\pi_c \leq \pi_d - p^*_{max} < 0$, which is a contradiction. ※

This completes the proof to Proposition 1. ■

Proof to Proposition 1*

Note that equation (14) for the qth player can be rewritten as follows:

$$\Pi_c(q) = q\pi_c + (1-q)\left[-\gamma q \pi_c \left(\pi_d - p^*(q)\right) - c\left[p^*(q)\right]\right] - \tau_q \tag{18}$$

Where τ_q is the qth player's type. Knowing that $\tau_q = \Phi^{-1}(q)$ by definition, the first derivative of expression (18) can be written as follows after simplification using equation (*10*):

$$\frac{\partial \Pi_c(q)}{\partial q} = \pi_c + \gamma \pi_c q \left(\pi_d - p^*(q)\right) + c\left[p^*(q)\right]$$

$$+ (1-q)\left[-\gamma \pi_c \left(\pi_d - p^*(q)\right)\right] - \frac{\partial \Phi^{-1}(q)}{\partial q} \tag{19}$$

The second derivative of expression (18) can thus be written as:

$$\frac{\partial^2 \Pi_c(q)}{\partial q^2} = 2\gamma \pi_c \left[\pi_d - p^*(q)\right] + (1-q)\gamma \pi_c \cdot \left(p^*\right)'(q) - \frac{\partial^2 \Phi^{-1}(q)}{\partial q^2} \tag{20}$$

If Φ is weakly convex, then Φ^{-1} is weakly concave, entailing that equation (20) is always positive whenever $\pi_d - p^*(q) > 0$. Hence, the proof to Proposition 1 (Case 1) applies here. Note that the Proof to Proposition 1 (Case 2) does not apply, however: if $\pi_d - p^*(q) < 0$ for some q, then $\Pi_c(q)$ is not necessarily convex

$\forall q: \ \pi_d - p^*(q) < 0$, entailing the possibility of an intermediate equilibrium with $\Pi_c(q) = \Pi_d(q) < 0$. Such an equilibrium is not very interesting, however, as it implies that more cooperation is less desirable to some players than no cooperation at all.

Finally, note that when Φ is not weakly convex, than the proof to Proposition 1 does not hold. Following the proof to Proposition 1, the only things we can say in such a case is that (i) a low equilibrium always exists; (ii) a high equilibrium exists whenever $\pi_c - \tau_{max} \geq \pi_d - p^*_{max}$; and (iii) if (ii) applies, then by the fixed point theorem, at least one intermediate equilibrium must also exist. ■

Proof to Proposition 2

I'll consider the following two cases and prove Proposition 2 for these two cases separately.

Case 1: $r^* = \Pi_d(\hat{q})$

Conditional on $m_i = M$, then $\Pi_d(q) > 0 \ iff \ q > \hat{q}$, for all players. By definition, $\forall q > \hat{q}$, we know that $\Pi_c(q) = \Pi_d(q)$. At equilibrium, clearly $(m_i^*, \ s_i^*) \neq (M, \ D)$. To see why, suppose by contradiction that a player chooses $m_i = M$ and $s_i = D$. Then, if $q \leq \hat{q}$, $\Pi_d(q) - r^* \leq 0$, and $m_i^* \neq M$ as $u_i(., \ m_i = M | s_i = D) < u_i(., \ m_i = N | s_i = D)$. If $q > \hat{q}$, then $s_i^* \neq D$ as $u_i(., \ s_i = D | m_i = M) < u_i(., \ s_i = C | m_i = M)$. Hence, we know that $(s_i^* | m_i = M) = C$ for all players. Then, the maximisation problem $\max_{m_i \in \mathcal{M}_i} u_i(., \ s_i^*, \ s_{-i}^*)$ yields the unique solution $m_i = M$ iff $\pi_c > r^*$, for all players.

Case 2: $r^* = \Pi_d(q^*)$

Conditional on $m_i = M$, then $\Pi_d(q) < 0$. Then, iff $\pi_c > r^*$, for reasons analogous to those presented with Case 1, it must be the case at equilibrium that $(m_i = M) \Rightarrow (s_i = C)$, for all players. The rest of the proof follows immidiately from the preceding case.

This completes the proof to Proposition 2. ■

Proof to Proposition 2*

The Proof to Proposition 2* stems logically from the proof to Proposition 2. ■

Proof to Proposition 3

Note that with the case $\pi_c > \hat{r^*} > \pi_c - \tau_{max}$, we have necessarily from Definition 4 that $\hat{r^*} = \Pi_d(q^*)$. To see why, suppose by contradiction that $\hat{r^*} = \Pi_d(\hat{q})$, with $\hat{q} < q^*$ and $\Pi_{c,q}(q) > \Pi_d(q) \ \forall q \in (\hat{q}, 1]$. Then, it follows necessarily that $\Pi_{c,q}(1) > \Pi_d(1)$, which implies that $\hat{r^*} < \pi_c - \tau_{max}$, which is a contradiction.

We also know that the type of the member of the religious group with the highest type can be written as $\tau_R = \pi_c - \hat{r^*}$. This member has a utility of zero. Furthermore, the level of religious affiliation q_R can be written as $q_R = \Phi(\tau_R)$ as all players with types lower than τ_R partake in the religious groups. Hence, religious affiliation can be rewritten $q_R = \Phi\left(\pi_c - \hat{r^*}\right)$, with $\Phi'(.) > 0$.

To do comparative statics with these expressions is straightforward. First, we note that $\frac{\partial \hat{r^*}}{\partial \pi_c} = -\frac{\partial p^*(q)}{\partial \pi_c}$, which is always negative (Remark 1). It follows immidiately that $\frac{\partial q_R}{\partial \pi_c} = \Phi'\left(\pi_c - \hat{r^*}\right) \cdot \left(1 + \frac{\partial p^*(q)}{\partial \pi_c}\right) > 0$. Second, $\frac{\partial \hat{r^*}}{\partial \pi_d} = 1 > 0$, which entails naturally that $\frac{\partial q_R}{\partial \pi_c} = -\Phi'\left(\pi_c - \hat{r^*}\right) < 0$. An increase in the cost of punishment has an effect analytically similar to an increase in π_d. ■

References

Abramitzky, R., 2011. On the (Lack of) Stability of Communes: An Economic Perspective, in R. McCleary (ed.), *The Oxford Handbook of the Economics of Religion*, Oxford: Oxford University Press.

Allen, D. S., 2003. Angry Bees, Wasps and Jurors: The Symbolic Politics of $\grave{o}\rho\gamma\eta$ in Athens, in S. Braund and G.W. Most (eds), *Ancient Anger Perspectives From Homer to Galen*. Cambridge: Cambridge University Press.

Axelrod, R., and W.D. Hamilton, 1981. The Evolution of Cooperation, *Science*, 211(4489), 1390–6.

Balliet, D. and P.A.M. Van Lange, 2013. Trust, Punishment, and Cooperation Across 18 Societies: A Meta-Analysis, *Perspectives on Psychological Science*, 8, 363-79.

Barro, R. and R. McLeary, 2005. Which Countries Have Sate Religions?, *Quarterly Journal of Economics*, 120(4), 1331-70.

Battigalli, P. and M. Dufwenberg, 2007. Guilt in Games, *American Economic Review Papers and Proceedings*, 97, 170-76.

——, 2009. Dynamic Psychological Games, *Journal of Economic Theory*, 144, 1-35.

Berman, E., 2000. Sect, Subsidy and Sacrifice: An Economist's View of Ultra-Orthodox Jews, *Quarterly Journal of Economics*, 115, 905-23.

Bosman, R. and F. Van Winden, 2002. Emotional Hazard in a Power-to-take Experiment, *Economic Journal*, 112(476), 147-169.

Bosman, R., Sutter, M. and F. Van Winden, 2005. The Impact of Real Effort and Emotions in the Power-to-take Game, *Journal of Economic Psychology*, 26(3), 407-29.

Bowles, S. and H. Gintis, 2002. Social Capital and Community Governance, *Economic Journal*, 112(483), 419-36.

Boyd, R., Gintis, H., Bowles, S. and P.J. Richerson, 2003. The evolution of altruistic punishment, *Proceedings of the National Academy of Sciences,* 100(6), 3531-35.

Bulbulia, J., 2004. The Cognitive and Evolutionary Psychology of Religion, *Biology and Philosophy*, 19, 655-86.

——, 2012. Spreading Order: Religion, Cooperative Niche Construction, and Risky Coordination Problems, *Biology and Philosophy*, 27(1), 1-27.

Bulbulia, J. et al. (eds), 2008. The Evolution of Religion: Studies, Theories, and Critiques. Santa Margarita, CA: Collins Foundation Press.

Charness, G. and M. Dufwenberg, 2006. Promises and Partnership, *Econometrica*, 74(6), 1579-601.

Chen, D.L., 2010. Club Goods and Group Identity: Evidence from Islamic Resurgence during the Indonesian Financial Crisis, *Journal of Political Economy*, 118(2), 300-54.

Cox, D. and M. Fafchamps, 2008. Extended Family and Kinship Networks: Economic Insights and Evolutionary Directions, in T.P. Shultz and J. Strauss (eds), *Handbook of Development Economics vol 4*, Amsterdam: Elsevier.

De Quervain, D.J.F. et al., 2004. The Neural Basis of Altruistic Punishment. *Science*, 305, 1254-258.

Dubreuil, B., (2010a). Human Evolution and the Origins of Hierarchies: The State of Nature, Cambridge: Cambridge University Press.

——, (2010b). Paleolithic Public Good Games: Why Human Culture and Cooperation Did Not Evolve in One Step, *Biology and Philosophy*, 25, 53-73.

Dufwenberg, M. and U. Gneezy, 2000. Measuring Beliefs in an Experimental Lost Wallet Game, *Games and Economic Behavior,* 30(2), 163-82.

Dufwenberg, M. and G. Kirchsteiger, 2004. A Theory of Sequential Reciprocity, *Games and Economic Behavior*, 47(2), 268-98.

Durlauf, S.N. and M. Fafchamps, 2005. Social Capital, in P. Aghion and S. Durlauf (eds.), *Handbook of Economic Growth*, Amsterdam: Elsevier.

Falk, A., Fehr, E. and U. Fischbacher, 2005. Driving Forces Behind Informal Sanctions, *Econometrica*, 73(6), 2017-30.

——, 2008. Testing Theories of Fairness-Intentions Matter, *Games and Economic Behavior,* 62, 287-303.

Falk, A. and U. Fischbacher, 2005. Modeling Fairness and Reciprocity, in H. Gintis, S. Bowles, R. Boyd and E. Fehr (eds), *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, Cambridge: MIT Press.

——, 2006. A Theory of Reciprocity, Games and Economic Behavior, 54, 293-315.

Fehr, E. and A. Falk, 2002. Psychological Foundations of Incentives, *European Economic Review*, 46, 687-724.

Fehr, E. and S. Gächter, 2000. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980–94.

——, 2002. Altruistic Punishment in Humans, *Nature*, 415, 137-40.

Fessler, D.M.T., 2010. Madmen: An Evolutionary Persepective on Anger and Men's Violent Responses to Transgression, in M. Potegal, G. Stemmler and C. Spielberg (eds), *International Handbook of Anger*, New-York: Springer.

Fischbacher, U. and S. Gächter, 2010. Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments, *American Economic Review*, 100(1), 541-56.

Fischbacher, U., S. Gächter and E. Fehr, 2001. Are People Conditionally Cooperative? Evidence from a Public Goods Experiment, *Economics Letters*, 71(3), 397-404.

Gächter, S., Renner, E. and M. Sefton, 2008. The Long-Run Benefits of Punishment, *Science*, 322(5907), 1510.

Gächter, S., Herrmann, B. and C. Thöni, 2010. Culture and Cooperation, *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 365(1553), 2651-61.

Habermas, J., 1990. The Philosophical Discourse of Modernity. Cambridge: Polity Press.

Henrich, J. 2004. Cultural Group Selection, Coevolutionary Processes and Large-scale Cooperation. *Journal of Economic Behavior and Organization*, 53, 3-35.

——, 2006. Cooperation, Punishment, and the Evolution of Human Institutions, *Science*, 312(5770), 60-1.

Henrich, J. and R. Boyd, 2001. Why People Punish Defectors: Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas, *Journal of Theoretical Biology*, 208(1), 79-89.

Henrich, J. et al., 2006. Costly Punishment Across Human Societies, *Science*, 312(5781), 1767-70.

——, 2010. Markets, Religion, Community Size, and the Evolution of Fairness and Punishment, *Science*, 327, 1480-84.

Hopfensitz, A. and E. Reuben, 2009. The Importance of Emotions for the Effectiveness of Social Punishment, *Economic Journal*, 119, 1534-559.

Huettel, S. and R. E. Kranton, 2012. Identity Economics and the Brain: Uncovering the Mechanisms of Social Conflict, *Philosophical Transactions of the Royal Society*, 367, 680-91.

Iannacconne, L.R., 1992. Sacrifice and Stigma: Reducing Free-Riding in Cults, Communes, and Other Collectives. *Journal of Political Economy*, 100, 271-91.

——, 1994. Why Strict Churches Are Strong, *American Journal of Sociology*, 99, 1180-211.

——, 1998. Introduction to the Economics of Religion, *Journal of Economic Literature*, 36(3), 1465-95.

Jensen, K., 2010. Punishment and Spite, the Dark Side of Cooperation, *Philosophical Transactions of the Royal Society*, 365, 2635-50.

Kandori, M., 1992. Social Norms and Community Enforcement, *Review of Economic Studies*, 59(1), 63-80.

Leeson, P.T., 2008. Social Distance and Self-Enforcing Exchange, *Journal of Legal Studies*, 37, 161-88.

Leeson, P.T. and C.J. Coyne, 2012. Conflict-Inhibiting Norms, in M.R. Garfinkel and S. Skaperdas (eds.), *The Oxford Handbook of The Economics of Peace and Conflict*, Oxford: Oxford University Press.

Levy, G. and R. Razin, 2012a. Religious Beliefs, Religious Participation, and Cooperation, *American Economic Journal: Microeconomics*, 4(3), 121-51.

——, (forthcoming). Calvin's Reformation in Geneva: Self and Social Signalling, *Journal of Public Economic Theory.*

McBride, M.T. and G.Richardson, (2012). Religion, Conflict, and Cooperation, in M.R. Garfinkel and S. Skaperdas (eds.), *The Oxford Handbook of The Economics of Peace and Conflict*, Oxford: Oxford University Press.

McLeary, R., 2011. The Economics of Religion as a Field of Inquiry, in R. McCleary (ed.), *The Oxford Handbook of the Economics of Religion*, Oxford: Oxford University Press.

Norenzayan, A. and A. Shariff, 2008. The Origin and Evolution of Religious Prosociality, *Science*, 322, 58-62.

Norris, P. and R. Inglehart, 2004. Sacred and Secular: Religion and Politics Worldwide. Cambridge: Cambridge University Press.

Olson, D.V.A., 2011. Toward Better Measures of Supply and Demand for Testing Theories of Religious Participation, in R. McCleary (ed.), *The Oxford Handbook of the Economics of Religion*, Oxford: Oxford University Press.

Ostrom, E., 2000. Social Capital: A Fad or Fundamental Concept?, in P. Dasgupta and P. Serageldin (eds), *Social Capital: A Multifaceted Perspective*, Washington: World Bank.

Potegal, M. and R. Novaco, 2010. A Brief History of Anger, in M. Potegal, G. Stemmler and C. Spielberg (eds), *International Handbook of Anger*, New-York: Springer.

Rabin, M., 1993. Incorporating Fairness into Game Theory and Economics, *American Economic Review*, 83(5), 1281-301.

Reuben, E. and F. Van Winden, 2008. Social Ties and Coordination on Negative Reciprocity: The Role of Affect, *Journal of Public Economics*, 82(2), 34-53.

Roberts, S.C. et al., 2013. Who Punishes? Personality Traits Predict Individual Variation in Punitive Sentiment, *Evolutionary Psychology*, 11(1), 186-200.

Rustagi, D., Engel, S. and M. Kosfeld, 2010. Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management, *Science* 330(6006), 961–965.

Smith, A., 2009. Belief-Dependent Anger in Games. Working Paper, Department of Economics, University of Arizona.

Sobel, J., 2005. Interdependent Preferences and Reciprocity, *Journal of Economic Literature*, 43, 392-436.

Sosis, R., 2000. Religion and Intragroup Cooperation: Preliminary Results of a Comparative Analysis of Utopian Communities, *Cross-Cultural Research*, 34, 70-87.

——, 2005. Does Religion Promote Trust?: The Role of Signaling, Reputation and Punishment, *Interdisciplinary Journal of Research on Religion*, 1(7), 1-30.

Sosis, R. and C.S. Alcorta, 2003. Signaling, Solidarity, and the Sacred: The Evolution of Religious Behavior, *Evolutionary Anthropology*, 12, 264-74.

——, 2004. Ritual, Emotions, and Sacred Symbols: The Evolution of Religion as an Adaptive Complex, *Human Nature*, 16(4), 323-59.

Sosis, R. and E. Bressler, 2003. Cooperation and Commune Longevity: A Test of the Costly Signaling Theory of Religion, *Cross-Cultural Research*, 37, 211-39.