A Dynamic Model of Belief-Dependent Conformity to Social Norms

Sontuoso, Alessandro

University of Pennsylvania

27 December 2013

A Dynamic Model of Belief-Dependent Conformity to Social Norms

Alessandro Sontuoso[1]

University of Pennsylvania, 249 South 36th Street, Philadelphia, PA 19104

December 27, 2013

Human conduct is often guided by "conformist preferences", which thrive on behavioral expectations within a society, with conformity being the act of changing one's behavior to match the purported beliefs of others. Despite a growing research line considering preferences for a fair outcome allocation, economic theories do not explain the fundamental conditions for some social norm – whether of fairness or not – to be followed. Inspired by Bicchieri's account of norms (C.Bicchieri, The Grammar of Society. CambridgeUP [2006]), I develop a behavioral theory of norm conformity building on the Battigalli-Dufwenberg "psychological" framework (P.Battigalli and M.Dufwenberg, Dynamic Psychological Games, J.Econ.Theory, 144:1-35 [2009]).

KEYWORDS: conformist preferences, social norms, social dilemmas, psychological game theory, behavioral economics.

I. Introduction

Socio-economic behavior is generally modelled on rational choice theory's prescriptions: economic theory assumes that an agent has preferences satisfying some rationality requirements, yet most traditional economic applications simply view those requirements as implying that the self-interest of the agent is narrowly *self-centred* and unaffected by the others' outcome. On the other hand, the widely documented regularities of behavior inconsistent with the standard predictions of models with rational self-centred individuals have motivated alternative accounts. Everyday life examples of such "incidents" might be brought about by norms that informally prescribe how people ought to behave in the community or workplace, and which are enforced out of fear of social sanctions: Arrow's [1972] pioneering

[1] I am grateful to Cristina Bicchieri, Dirk Engelmann, David Rojo-Arjona, and Robert Sugden for their helpful comments. All errors are mine.

Philosophy, Politics and Economics, University of Pennsylvania, e-mail: sontuoso@sas.upenn.edu

investigation suggests that entrepreneurs, who could turn a profit on hiring labour cheaply from a racially discriminated group, were restrained from doing so owing to the establishment of social customs involving discriminatory tastes; or rather, as Akerlof [1980] claims, if the custom prohibits an employer from hiring labour at a reduced wage, employees will not cooperate in training new workers who undercut existing wages, because by doing so they would suffer a loss of reputation for participating in disobeying the norm. Other situations that are often explained by the enforcement of informal norms regulating social behavior include the voluntary supply of public goods (Sugden [1984]) and altruistic or reciprocity-based transactions such as gift-giving, *etc.* (Sacco *et al.* [2006]).

The above instances seem to be validated by a wealth of experimental evidence in mixed-motive (*i.e.*: social dilemma) games, which provide support against the traditional self-centred view of economic agents (Camerer [2003], Ch. 2, Fehr and Schmidt [2006], Ledyard [1995]). In this regard, the present investigation contributes to the existing explanatory literature by focusing on a conditional motivation that can make people comply with default rules of behavior in social dilemmas. In a nutshell, this essay suggests that many individuals have a tendency to follow the behavior, attitudes or judgements of others, with the others' observed or purported behavior being considered appropriate or normal (within a certain social group): here a *"behavioral rule"* capturing some appropriate behavior is formally defined as a correspondence that dictates a set of strategy profiles at each node of an extensive form game; under precisely stated conditions it is assumed that conformity is generated by the anticipation of some negative emotion, which would arise in the event of violations of the relevant rule. The conditions for a *"social norm"* to be followed by a certain population will be defined regardless of either the specifics or the intrinsic value of the behavioral rule; in other words, the conditions to be introduced in this paper shall apply to any rule of behavior that may be collectively adopted by a social group, thereby coming to

constitute a *social* norm (*e.g.*: norms of equality, reciprocity, revenge, efficiency, *etc.*).

It should be noted that relatively recent developments in behavioral game theory have substantially improved the analysis of strategic interaction by allowing for diverse assumptions about players' emotions and preferences. Some of the *social preference* theories, namely the so-called models of "reciprocal fairness", seem to be most effective in accounting for other-regarding behavior where intentions matter: think of Rabin [1993], Dufwenberg-Kirchsteiger [2004], Falk-Fischbacher [2006], Charness-Rabin [2002]. All such psychological game theory models assume that players have a preference for a somehow specified equitable payoff (Rabin [1993], Dufwenberg-Kirchsteiger [2004]) or they are intention-based inequity averse (Falk-Fischbacher [2006][2]) or they have a taste for both fairness and efficiency (captured by quasi-maximin preferences in Charness-Rabin [2002]); so, the aforementioned models may be interpreted as more or less implicitly assuming that players have internalized a variously defined, unique norm of fairness or reciprocity. Now, while each of those models can explain a substantial part of the experimental results on other-regarding behavior, by assuming a stable disposition towards some pre-defined notion of fairness any one model cannot generally explain the fact that different individuals are often motivated by *different forms of (possibly culture-dependent) other-regarding principles* (Henrich *et al.* [2001], Fischbacher and Gächter [2010]); also, the above models cannot generally account for an individual having a preference for a

---

[2] Falk and Fischbacher [2006] define "kindness" directly in relation to the payoff that the co-player gets: their model can therefore be viewed as an *intention-based* inequity aversion theory (as opposed to a *simple* inequity aversion theory *à la* Fehr and Schmidt [1999] or Bolton and Ockenfels [2000]).

certain outcome, *conditional on the fact that she expects others not to deviate from the precepts of the relevant rule of behavior* (*e.g.*: people often dislike vandalism or littering, although they are likely to indulge in misbehavior whenever evidence of vandalism or littering is present in the environment). Similarly, the aforementioned models are typically vulnerable to changes in the framing of games which, as it will be clear, affects the players' behavioral expectations: for example, it has been observed that subjects' altruistic behavior often varies with contextual factors involving the extent to which some subject i knows that her counterpart j is aware that i is responsible for some "inappropriate" behavior (in this respect, Dana *et al.* [2007] show that relaxing the players' common knowledge of a one-to-one mapping between actions and outcomes in Dictator Game experiments gives subjects the moral "wiggle room" to behave selfishly).

Now, surveys from various disciplines – including neuroscience and cognitive psychology – support the view that human conduct is often guided by *conformist preferences* (Klucharev *et al*. [2009], Montague and Lohrenz [2007]) which thrive on behavioral expectations within a society or group, with conformity being the act of changing one's behavior to match the purported beliefs of others (Cialdini and Goldstein [2004]). To that end, the present essay takes the investigation of other-regarding preferences in mixed-motive games one step further: despite a growing body of literature considering preferences for a fair outcome allocation among players, economic theories do not explain the fundamental conditions for some social norm (whether of fairness or not) to exist and to be in operation among players with conformist motivations. Therefore, inspired by Cristina Bicchieri's [2006] philosophical account of norms, here I develop an original behavioral theory of conditionally conformist preferences in social dilemmas, building on Battigalli and Dufwenberg's [2009] framework for the analysis of dynamic psychological games. To sum up, in what follows: I define a *behavioral rule* as a correspondence dictating the strategy profiles most "appropriate"

(according to some principle); I assume that conformist players, at each decision node, hold a conjecture about the active player's *rule-complying actions* available at that node; I then model the expected utility function of a conformist player as a linear combination of her material payoff and a psychological component representing a negative emotion arising from presumed norm violations. A *social norm* is said to exist and to be followed by a population whenever players maximize their expected utilities, given their correct beliefs and their conformist preferences being conditional on the following elements: *i.* they are aware of the existence of some rule of behavior; *ii.* they believe that the others will behave in keeping with some rule; *iii.* they believe that the others expect them to behave in keeping with some rule, and the cost of a potential violation is sufficiently high to make it unprofitable.[3]

The remainder of the essay is organized in this manner: II introduces some general notation on extensive form games and conditional systems of beliefs; III formally lays out the model; IV discusses an equilibrium solution; V provides some applications, and VI concludes.

II. Preliminaries

1. *Notation on extensive form games*

An extensive form game (with perfect recall) is given by the structure $\langle N, H, P, (\mathrm{I}_i)_{i \in N} \rangle$, where: $N = \{1, \ldots, n\}$ is the *set of players*, H is the finite *set of feasible histories*, P is the *player function*, I_i is the *information partition* of

---

[3] The conditions are derived from Bicchieri's [2006] pioneering account, although their formal implementation will introduce a number of advances on Bicchieri's framework, since here the players' utility function will directly reflect a mathematically-precise specification of the very conditions: a "psychological" utility function and the use of extensive form games with updating of beliefs will result in increased predictive power.

Player i. Each element of H is a history, which is a (finite) *sequence of actions* taken by the players: let $h(a^l)$ denote a sequence $(a^1, \dots, a^l)$, with a^l being the l-th action chosen along the game tree.[4] Further, let Z denote the *set of terminal histories*, with $H \backslash Z$ being the set of non-terminal histories; given that, let $A_i(h)$ denote the *set of feasible actions* for Player i at history h.

The *player function* P assigns to each element of $H \backslash Z$ an element of N, with $P(h)$ being the player choosing an action after the history h. Then, for each player $i \in N$, I_i denotes the *information partition* of Player i – and $I_i \in \mathrm{I}_i$ is an information set of Player i – where a partition I_i of $H_i = \{h \in H: P(h) = i\}$ has the property that $A(h) = A(h')$ if h and h' are in the same cell of the partition. The *material payoffs* of players' strategies are described by functions $m_i: Z \to \mathbb{R}$ for each player $i \in N$. Further, for each player $i \in N$ let S_i denote the *set of pure strategies* of Player i: hence, $s_i = \left(s_{i,h}\right)_{h \in H \backslash Z}$ is a strategy for Player i, that is, a plan specifying the action chosen at every history after which Player i moves (with $s_{i,h}$ being the action implemented by s_i if history h occurred). A strategy profile s is a tuple of strategies, with one strategy for each player of the game: let $S = \prod_{i \in N} S_i$ be the *set of strategy profiles*; similarly define $S_{-i} = \prod_{j \neq i} S_j$ for players j other than i. Finally denote the *set of Player i's pure strategies allowing history h* (*i.e.*: strategies leading to, and succeeding, h) as $S_i(h)$; strategy profiles allowing history h are defined as $S(h) = \prod_{i \in N} S_i(h)$, and $S_{-i}(h) = \prod_{j \neq i} S_j(h)$ for all players j other than i. With a slight abuse of notation, let $z(s)$ indicate a terminal history induced by some strategy profile $s \in S$.

---

[4] Notice that, in what follows, a node of the game tree is identified with the history leading up to it (*i.e.*: a path in the game tree) as in Osborne and Rubinstein [1994].

2. *Conditional systems of beliefs*

Battigalli and Dufwenberg [2009] provide a framework for the analysis of dynamic psychological games, where conditional higher-order systems of beliefs influence the players' motivation. As in their model, here behavioral strategies are used to describe Player j's beliefs about Player i's actions at each history after which i has to play: formally, a *behavioral strategy* of Player i is a collection of independent probability measures $\sigma_i = \left(\sigma_i(\cdot\,|h)\right)_{h\in H\setminus Z} \in \prod_{h\in H\setminus Z} \Delta\left(A_i(h)\right)$, where $\sigma_i(a|h)$ is the probability of action a at history h and $\Delta\left(A_i(h)\right)$ denotes the set of probability measures over the set of Player i's feasible actions at history h. Then, $\Pr_{\sigma_i}(\cdot\,|\hat{h}) \in \Delta\left(S_i(\hat{h})\right)$ is the probability measure over Player i's strategies, conditional on $\hat{h}$, derived from σ_i and therefore, for some pure strategy $s_i \in S_i(\hat{h})$, $\Pr_{\sigma_i}(s_i|\hat{h}) := \prod_{h\in H\setminus Z:\, h\succcurlyeq\hat{h}} \sigma_i(s_{i,h}|h)$ indicates the conditional probability of s_i, given that $\hat{h}$ has occurred (note that $h \succcurlyeq \hat{h}$ is a history subsequent or equal to $\hat{h}$, and $s_{i,h}$ is the action selected by s_i if history h took place).

Following Battigalli and Dufwenberg's [2007] notation, every player $i \in N$ holds a *system of first-order beliefs* $\alpha_i = \left(\alpha_i(\cdot\,|h)\right)_{h\in H_i}$ about the strategies of all the co-players (*e.g.*: in a game with perfect information, at each $h \in H_i$ Player i holds an updated belief $\alpha_i(\cdot\,|h) \in \Delta\left(S_{-i}(h)\right)$ such that she believes that all players have chosen all the actions leading to h with probability 1). At each $h \in H_i$ Player i further holds a *system of second-order beliefs* β_i about the first-order belief system of *each* of the opponents: for simplicity, for some $h \in H_i$, $\beta_i(h)$ indicates a collection of i's *point beliefs*

about every j's first-order belief (*i.e.*: $\beta_i(h)$ denotes i's point belief about $\alpha_{-i} = \left(\alpha_j(\cdot\,|h')\right)_{j\neq i,\, h'\in H_j}$). Given that, for each $j \neq i$, let $\beta_{i,j}^{S_i}(h) \in \Delta\big(S_i(h)\big)$ denote Player i's strategy-part of $\left(\alpha_j(\cdot\,|h')\right)_{h'\in H_j}$, which represents i's point belief about what some player $j \neq i$ believes about i's strategies.[5] Finally, it is assumed that players' beliefs at different information sets must satisfy Bayes' rule and common knowledge of Bayesian updating.

III. A model of social norms

1. *Behavioral rules*

I can now turn to shape an original theory of conformity to social norms. In this sub-section a *"behavioral rule"* is defined as a correspondence that dictates a set of strategy profiles at each decision node of the game tree. For a given history/node, the dictated set of strategy profiles is intended as indicating *behavior appropriate from that history onwards*.[6]

Definition 1. Given an extensive form game G, a *behavioral rule* is a set-valued function r that assigns to every non-terminal history $h \in H\backslash Z$ one or more elements from the set $S(h)$ of strategy profiles allowing history h; that is, a behavioral rule $r\colon H\backslash Z \to S$ is a correspondence dictating the strategy

---

[5] Recalling that a behavioral strategy σ_i is used to describe the other players' beliefs about Player i's behavior, the reader can anticipate that (as it will be imposed later on) in equilibrium $\alpha_i(s_{-i}|\hat{h}) \equiv \prod_{j\neq i} \Pr_{\sigma_j}(s_j|\hat{h})$. Besides, since in equilibrium α_i will be derived from the behavioral strategy profile $\sigma = (\sigma_i)_{i\in N}$, every player $j \neq i$ will hold the same beliefs about Player i's strategies, which implies that in equilibrium $\beta_i^{S_i}(h) \in \Delta\big(S_i(h)\big)$ represents Player i's beliefs about what every other player unanimously believes about i's strategies.

[6] This implies that if the set of strategy profiles dictated by the behavioral rule at the initial history is singleton and if, along the play, no player ever deviates from such prescripts, then that rule will dictate exactly the same strategy profile at all successor nodes.

profiles most "appropriate" – according to a certain principle – for each node of the given (mixed-motive) game.

Instances of such behavioral rules include instructions that prescribe behavior minimizing payoff-inequality among players, procedures that dictate behavior maximizing the players' joint welfare, rules instructing players to reciprocate the preceding action, *etc.*.[7] For example, consider a rule that prescribes behavior minimizing payoff-inequality among players: when one evaluates such a rule at the root of a game tree, the rule will dictate those strategy profiles that minimize the difference in payoffs among players – at a given terminal node – considering that every terminal node can be reached. Now, assume that one of the players deviates along the play (by choosing an action that was not part of the set of strategy profiles dictated at the root); then, when evaluating this behavioral rule at a node following such a deviation, the rule will dictate strategy profiles that minimize the difference in payoffs among players, *conditional on the terminal nodes that can still be reached*.[8]

Before I move on, it should be highlighted that a behavioral rule, as per definition 1, does not embody in itself an element of rationality. Further, it is assumed that all rules regulating social dilemmas are contained in a universal set of behavioral rules, while each player is only aware of the rules contained

---

[7] For a few specific (formal) definitions of behavioral rules, see section V below.

[8] Notice that the above definition of behavioral rule is different from the one suggested by López-Pérez [2008], where a "norm" is defined as a correspondence mapping h into $A(h)$ for all $h \in H$. In fact, here it is argued that defining a behavioral rule as a correspondence mapping non-terminal histories into *strategy profiles* allows to better capture the strategic complexity of many norm-driven situations: note that the present definition is useful when considering games with both conditionally and unconditionally conformist players (since defining a rule in such a way allows to take into account the behavior of an *unconditionally conformist* player who gets to move after someone's deviation). On a different note, disregarding the role of expectations in sustaining a social norm seems to be a conceptual drawback of López-Pérez's model, although that certainly makes his framework parsimonious.

in her personal subset (as determined by a collection of attitudes, values, goals, and practices characterizing her group, organization or institution); thus, denote by R the *set of behavioral rules*, and for each $i \in N$ let R_i be the *behavioral rule subset* of Player i, with $R_i \subseteq R$. To sum up, the interpretation is as follows: given a universal set of rules R, the culture of each player i marks out a subset R_i, stored in i's memory, which contains default rules of behavior in accordance with set usage, procedure, discipline or principle she is aware of. It is assumed that each player's rule subset may contain all or just part of the rules of the other players' subsets – depending on the extent to which players share the same culture – or may even be empty.

Now, given an extensive form game G and some behavioral rule $\dot{r}$, with $\dot{r} \in R$, let $\dot{r}(h^0)$ denote the set of strategy profiles that that rule dictates at the initial history, henceforth referred to as the *set of strategy profiles completely consistent with* $\dot{r}$.[9] Further, given a rule subset $R_i \subseteq R$ for each $i \in N$, denote by $R_i(h)$ the *set of rule-complying strategy profiles allowing history* h, which is defined as $R_i(h) = \{s \in S(h)\colon \exists r \in R_i \text{ s.t. } s = r(h)\}$: in other words, $R_i(h)$ is the set of strategy profiles consistent with any $r \in R_i$ that is evaluated at a certain history/node h. Given that, let $A_{i,h}\left(R_i(\hat{h})\right)$ denote the *set of Player* i*'s rule-complying actions at history* h, which depicts the set of actions prescribed (by any $r \in R_i$) to Player i at history $h \succcurlyeq \hat{h}$; so, if Player i – once at history h – takes an action being part of the rule-complying strategy profiles allowing $\hat{h}$, then $s_{i,h} \in A_{i,h}\left(R_i(\hat{h})\right)$. Finally, denote the *set of Player* i*'s rule-complying strategies allowing history* h as $S_i\big(R_i(h)\big)$, which

---

[9] The expression "completely consistent" alludes to the fact that at h^0 the behavioral rule dictates a set of strategy profiles indicating *behavior appropriate for the game as a whole*.

represents i's strategy-part of the set of rule-complying strategy profiles allowing h.

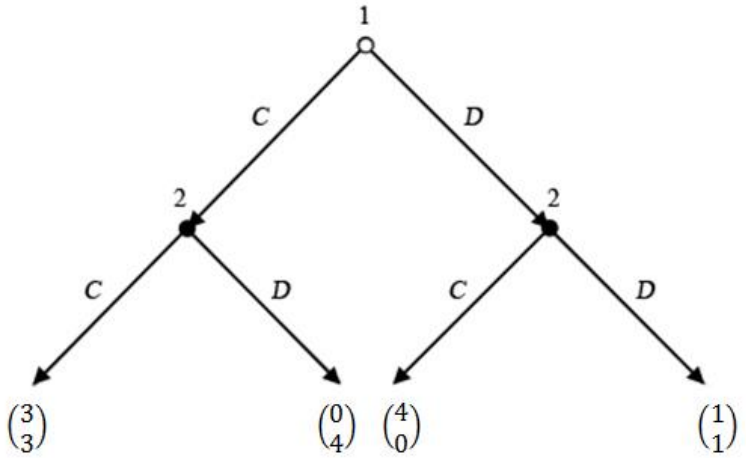

Figure 1 - Dynamic Prisoner's Dilemma *"DPS"*

Consider the above Dynamic Prisoner's Dilemma and let r be an "efficiency rule", defined as a behavioral rule dictating strategy profiles that, at each history/node, maximize the players' joint welfare. It is clear that, in this case, the set of strategy profiles completely consistent with r is singleton, that is, $r(h^0) = \{(C, CC)\}$; indeed, the strategy profile (C, CC) yields the payoff profile $(3,3)$, which maximizes the sum of the players' payoffs. Now assume that, for whatever reason, Player 1 deviates from the precepts of the efficiency rule by choosing D; thus, when evaluating the efficiency rule at $h(D)$, such a behavioral rule will still dictate a strategy profile that maximizes the players' joint welfare, but conditional on the terminal nodes that can be reached now (it follows that here $r\big(h(D)\big) = \{(D, CC)\}$, which yields the payoff profile $(4,0)$). Further, assuming that R_i contains only the efficiency rule, for $i = 1,2$, one

can denote the set of Player 1's rule-complying strategies allowing h^0 as $S_1\big(R_i(h^0)\big) = \{C\}$. Similarly, for Player 2, $S_2\big(R_i(h^0)\big) = \{CC\}$; hence the sets of Player 2's rule-complying actions at $h(C)$ and $h(D)$ can be expressed, respectively, as $A_{2,h(C)}\big(R_i(h^0)\big) = \{C\}$ and $A_{2,h(D)}\big(R_i(h^0)\big) = \{C\}$.[10]

2. *Norm-conjectures*

It is assumed that conformist players, conditional on each history/node of an extensive form game, hold a conjecture about the active player's rule-complying actions at that history.

Definition 2. Given an extensive form game G and for each $i \in N$ a rule subset $R_i \subseteq R$, a *norm-conjecture* of Player i is a collection of independent probability measures $\rho_i = \big(\rho_i(\cdot\,|h)\big)_{h\in H\setminus Z} \in \prod_{h\in H\setminus Z} \Delta\Big(A_{P(h)}(h)\Big)$, with $\rho_i(a|h)$ being the probability of action a at history h, such that:

$$\text{supp}\ \rho_i = \text{supp}\ \big(\rho_i(\cdot\,|h)\big)_{h\in H\setminus Z} \in \prod\nolimits_{h\in H\setminus Z} A_{P(h),h}\big(R_i(h^0)\big),$$

where supp ρ_i denotes the *support* of ρ_i, and $A_{P(h),h}\big(R_i(h^0)\big)$ is the set of rule-complying actions of the player active at h, as dictated at h^0 by any $r \in R_i$.[11]

In plain words, conditional on each $h \in H \setminus Z$ Player i holds a conjecture $\rho_i(\cdot\,|h)$ about the active player's (rule-complying) actions at h. It should be

---

[10] It should be noted that generally, in the event that $R_i \neq R_j$ (for some $i, j \in N$, with $j \neq i$), the set of Player i's rule-complying strategies – according to i's rule subset $R_i \subseteq R$ – may not be the same as the set of Player i's rule-complying strategies according to j's rule subset: in other terms, it might well be that $S_i\Big(R_i\big(\hat{h}\big)\Big) \neq S_i\Big(R_j\big(\hat{h}\big)\Big)$ for some history $\hat{h}$, which indicates that Player i and Player j disagree about which of Player i's strategies would constitute appropriate behavior.

[11] Recall that, for example, if some player $y \in N$ takes an action immediately after history h, then the value of the player function at h is y, *i.e.*: $P(h) = y$.

stressed that, possibly depending on the degree of cultural heterogeneity of the players' set N, it may not be obvious that one rule-complying strategy profile is more plausible than another rule-complying strategy profile, so conformist players have to form conjectures about "what would be normal to do" in the specific situation at hand. In fact, it should be noted that R_i may contain multiple elements (or, even if R_i is singleton, one behavioral rule might be ambiguous in that it could dictate multiple strategy profiles), hence a conformist player will have to form a conjecture $\rho_i = \left(\rho_i(\cdot\,|h)\right)_{h\in H\setminus Z}$ indicating what she reckons that a player would do upon reaching each $h \in H \setminus Z$. To sum up, R_i determines the set of rule-complying actions of the player active at h (*i.e.*: more precisely, the rules one is aware of determine a set $A_{P(h),h}\left(R_i(h^0)\right)$ of admissible/rule-complying actions), whereas ρ_i determines which of these admissible actions are plausibly taken in the current play of G. Before proceeding, it is convenient to make the following assumption.

Assumption 1. Given an extensive form game G and for each $i \in N$ a rule subset $R_i \subseteq R$, if $A_{y,\hat{h}}\left(R_i(h^0)\right) = \emptyset$ for some $\hat{h} \in H\setminus Z$ and $P\left(\hat{h}\right) = y$, then $\text{supp}\,\rho_i\left(\cdot\,|\hat{h}\right) \in A_y\left(\hat{h}\right)$.

Assumption 1 states that in the case in which at some history $\hat{h}$ the set of rule-complying actions is empty, ρ_i may assign positive probability to any action at that history (*i.e.*: to any action in $A_y\left(\hat{h}\right)$). In effect some behavioral rules (*e.g.*: the strict equality rule) may not be defined at all histories, so assumption 1 makes it possible for ρ_i to assign positive probability to actions at each and every information set. Here the interpretation is that, if a player $y \in N$ gets to move at a node at which none of the rules in R_i is defined, then y is believed to be free to take any available action (or, equivalently, every action is considered "rule-complying").

Now, I can move on to introduce the "relativist's conception of *moral choice*", as follows.

Definition 3. A strategy $s_i^* = \left(s_{i,h}\right)_{h \in H \setminus Z}$ is a *moral choice according to norm-conjecture* $\rho_i = \left(\rho_i(\cdot\,|h)\right)_{h \in H \setminus Z}$ if the following condition holds for all $s_i \in S_i$:

$$\Pr_{\rho_i}(s_i^*|h^0) \geq \Pr_{\rho_i}(s_i|h^0),$$

where $\Pr_{\rho_i}(s_i|h^0)$ is the conditional probability of a pure strategy of Player i at h^0 – derived from i's norm-conjecture ρ_i – and is calculated as $\Pr_{\rho_i}(s_i|h^0) = \prod_{h \in H_i} \rho_i\left(s_{i,h}\middle|h\right)$.

Notice that definition 3 describes moral choice as a strategy with the highest probability of being considered "currently-normal" or "appropriate", given a rule subset R_i: again, it should be stressed that such a choice is still independent of payoff-maximization considerations. That said, the aim of this paper is not to advocate moral relativism, but to utilize moral relativism as one of the features characterizing conformist individuals.

3. *Conditionally conformist preferences*

A norm-driven decision maker i is modelled as a player with *conditionally conformist* preferences, whose expected utility function is a linear combination of her material payoff and a component representing some anticipated negative emotion (*i.e.*: a function of the sum of losses that other conformist players j would suffer because of a rule violation). To that end, one needs to define some player j's expectation of her material payoff, given her strategy s_j and her initial belief $\alpha_j = (\cdot\,|h^0)$ about the strategies of the co-players; so, drawing on Battigalli and Dufwenberg's [2007] concept of *simple guilt*, such an expectation is given by $\mathrm{E}_{s_j,\alpha_j}\left[m_j|h^0\right] = \sum_{s_{-j}} \alpha_j\left(s_{-j}\middle|h^0\right) m_j\left(z\left(s_j, s_{-j}\right)\right)$. Here, if Player j is a conditionally conformist decision maker – and presumes

that her co-players are norm-driven too – she can form her belief α_j by assuming her co-players' behavior to be consistent with some rule r.

Now, the present theory assumes that players are naïve in the following way: if Player i presumes that her co-players are norm-driven, then she believes that they are aware of the same behavioral rules as hers.

Assumption 2. Given an extensive form game G and a rule subset $R_i \subseteq R$, (unless the players' awareness of the rule subsets is otherwise specified) Player i believes that the co-players' rule subsets are the same as hers; that is, Player i believes that $R_i = R_j, \forall\, j \in N$.

As a consequence of assumption 2, if Player j presumes that her co-players are norm-driven, then she believes that they hold the same norm-conjecture as hers. It follows that Player j will form her first-order belief α_j by assuming her co-players' behavior (at each history where they are active) to be consistent with her own norm-conjecture $\rho_j = \left(\rho_j(\cdot\,|h)\right)_{h \in H \setminus Z}$. Notice that, here, her initial belief $\alpha_j = (\cdot\,|h^0)$ will still correspond to a probability measure over the strategies of the opponents, except that now the support of α_j will contain only rule-complying strategies (according to j's rule subset R_j). Thus, the probability of a certain strategy profile of all players other than j is now given by $\alpha_j(s_{-j}|h^0) \equiv \Pr_{\rho_j}(s_{-j}|h^0) = \prod_{i \neq j} \prod_{h \in H_i:\, h \succcurlyeq h^0} \rho_j(s_{i,h}|h)$. Note that, for the sake of simplicity, the present theory assumes that *players cannot randomize*, yet randomized choices may enter the analysis as an expression of the players' beliefs about the opponents' (rule-complying) strategies. Given that, a *norm-driven decision maker's preferences* are defined as follows.

Definition 4. A norm-driven decision maker has conformist preferences characterized by a utility function u_i^C of the form

$$u_i^C(z, s_{-i}, \alpha_j) = m_i(z) - k_i d_i^C d_i^E \left(1 + \sum_{j \neq i} max\left\{0, \mathrm{E}_{\rho_i, s_j, \alpha_j}[m_j | h^0] - m_j(z)\right\} \right),$$

with $s_{-i} \in S_{-i}(z)$, $k_i \in [0, \infty)$ and where:

- k_i is Player i's *sensitivity to the presumed norm*;
- d_i^C is a dummy variable equal to one if i is aware of one or more behavioral rules applicable to the given game (*i.e.*: $d_i^C = 1$ whenever $R_i \neq \emptyset$), equal to zero otherwise;
- d_i^E is a dummy variable equal to one if i believes that every $j \neq i$ is aware and will also adhere to some $r \in R$, equal to zero otherwise.

It is now clear that the anticipated negative emotion is a function of any positive difference between the initially expected payoff to j and the payoff j would get in the event of a rule violation. Note that i does not know what α_j is: in effect β_i provides i's estimation of α_j, which i will compute by presuming that j holds the same norm-conjecture as hers; that is, Player i's estimation of $\alpha_j(s_{-j}|h^0)$ will be given by $\beta_i(s_{-j}|h^0) \equiv \Pr_{\rho_i}(s_{-j}|h^0) = \prod_{y \neq j} \prod_{h \in H_y:\, h \succcurlyeq h^0} \rho_i(s_{y,h}|h)$.

To sum up, if $d_i^C = 1$, $d_i^E = 1$, and $k_i > 0$ Player i will exhibit conformist preferences. It should be stressed that the sensitivity parameter k_i sets the size of a hypothetical feeling of uneasiness of member i of a group in which, because of a rule violation, some other member's welfare gets reduced: the underlying assumption is that individuals may feel resentment at injustice (Sugden [2000], Elster [1989], Ch. 6) – or rather here resentment at behavior different from an established pattern – and the anticipation of such resentment

would bring about a negative emotion on the part of a potential deviator; it is assumed that (while d_i^C and d_i^E are endogenously determined) k_i is exogenously given.[12, 13]

4. *Social norms*

Given the above apparatus, I shall introduce a set of conditions for a social norm to exist or, more precisely, conditions for a behavioral rule r to constitute a "social norm for i". Before proceeding it should be highlighted that the present theory differs from conventional social preference models in that the form of the current utility function incorporates a taste for conditional preferences, as is the case of an individual having a preference for conformity to some principle – whatever the relevant behavioral rule prescribes to her –

---

[12] Obviously if $k_i = 0$ or $d_i^C = 0$ or $d_i^E = 0$ the utility function reduces to one of standard non-conformist motivation. The reader can anticipate that the psychological component of such an expected utility function is always null in equilibrium. In fact, if i correctly expects that at least one player $j \neq i$ will not adhere to some $r \in R$, then d_i^E takes on value 0 (hence the psychological disutility is null); moreover, if i correctly expects that every player $j \neq i$ will adhere to some $r \in R$ and i herself adheres to that rule, then d_i^E takes on value 1 but no member's welfare gets reduced (hence the psychological disutility is null).

[13] It should be noted that the utility function of definition 4 differs from the one proposed by Bicchieri (Bicchieri [2006], Ch. 1) since, according to Bicchieri's function, Player i would suffer a loss in utility also in the case in which she conforms to the norm but Player j does not and, by doing so, i gets a material payoff lower than the one implied by the norm. The two specifications further differ in that Bicchieri's utility function does not involve a psychological component such that the opponents' beliefs explicitly affect a player's preferences (a fact that – when adapting Bicchieri's utility function to dynamic games – would rule out the possibility of updating beliefs about the opponents' norm-driven behavior). Here is a brief description of Bicchieri's [2006] utility function: considering a normal form game, a norm N_i is defined as a (set-valued) function from one's expectation about the opponents' (rule-complying) strategies to one's own strategies, that is, $N_i: L_{-i} \rightarrow S_i$, with $L_{-i} \subseteq S_{-i}$; a strategy profile is said to instantiate a norm for Player j if $s_{-j} \in L_{-j}$, and to violate a norm if $s_j \neq N_j(s_{-j})$. Player i's utility function is a linear combination of i's material payoff $\pi_i(s)$ and a component that depends on norm compliance: $U_i(s) = \pi_i(s) - k_i \max_{s_{-j} \in L_{-j}} \max_{m \neq j} \{\pi_m(s_{-j}, N_j(s_{-j})) - \pi_m(s), 0\}$, where $k_i \geq 0$ shows i's sensitivity to the norm and j refers to the norm violator. The norm-based component represents the maximum loss resulting from all norm violations: the first maximum operator aims at taking care of the possibility that there might be multiple rule-complying strategy profiles; the second maximum operator ranges over all the players other than the norm violator j.

only on condition that the others do not deviate from the precepts of that rule.[14]

Definition 5. Let $r \in R$ be a behavioral rule applicable to a certain class $\mathbb{C}$ of mixed-motive games, where each game is a structure $G = \langle N, H, P, \left(u_i^C\right)_{i \in N} \rangle$. r is a *social norm* for Player i of G, if the following conditions hold for i.

1. (contingency) $r \in R_i \Longrightarrow d_i^C = 1$.
2. (conditional preference) $\exists s^* \in r(h^0)\ \text{s.t.}\ u_i^C\left(z(s_i^*, s_{-i}^*)\right) \geq u_i^C\left(z(s_i, s_{-i}^*)\right)$ for $\forall\, s_i \in S_i$, where $r(h^0)$ is the set of strategy profiles completely consistent with r. That is, i prefers to adhere to r in a play of G *if*:

   2.1. (empirical expectations)

$$\left(\begin{array}{c} \alpha_i(s_{-i}|h^0) = \prod_{j \neq i} \Pr_{\rho_i}\left(s_j \middle| h^0\right)\ for\ \forall\, s_{-i} \in S_{-i}, \\ with\ \text{supp}\left(\rho_i(\cdot\,|h)\right)_{h \in H \setminus Z} \in \prod_{h \in H \setminus Z} A_{P(h),h}\left(r(h^0)\right) \end{array}\right) \Longrightarrow d_i^E = 1;$$

$$and$$

   2.2. (normative expectations)

   i. $\beta_i^{S_i}(h^0) = \left(\Pr_{\rho_i}(s_i|h^0)\right)_{s_i \in S_i}$, *and*

   ii. $k_i > 0\ and\ sufficiently\ large.$

A few comments are in order.[15] Condition *1* states that i is aware of some behavioral rule r applicable to game G. Condition *2.1* states that i believes that

---

[14] Recall that $-k_i d_i^C d_i^E \left(1 + \sum_{j \neq i} max\left\{0, \text{E}_{\rho_i, s_j, \alpha_j}\left[m_j|h^0\right] - m_j(z)\right\}\right)$ is to be intended as an anticipated negative emotion on the part of member i of a group in which peer j's welfare gets reduced: notice that the endogenously defined dummy d_i^E makes i's psychological loss immaterial unless i is the only deviator; that is, i is inclined to feel an aversion to rule-breaking through an anticipated negative emotion but, if someone else is already expected not to be adhering to the presumed norm (*i.e.*: if $d_i^E = 0$), then i does not care about j's welfare any longer. This is in contrast with conventional social preference models, where players may deviate (as long as it is convenient to them) while still experiencing a psychological loss.

every $j \neq i$ adheres to $r \in R_i \cap R_j$;[16] that is, i's first-order belief is derived from i's norm-conjecture ρ_i, with the support of ρ_i containing rule-complying actions as dictated at h^0 by r (save for cases in which at some $\hat{h}$ the set of rule-complying actions is empty, in which case assumption 1 holds). Then, the interpretation of condition *2.2* is that i believes that every $j \neq i$ believes that she ought to behave according to ρ_i. More precisely, condition *2.2* holds whenever its two components hold at once: *(i)* the first expression (*i.e.*: $\beta_i^{S_i}(h^0) = \left(\Pr_{\rho_i}(s_i|h^0)\right)_{s_i \in S_i}$) states that i believes that every $j \neq i$ expects her to behave according to i's norm-conjecture ρ_i, that is, i's second-order belief is derived from ρ_i; *(ii)* the second component (*i.e.*: $k_i > 0$ and $sufficiently\ large$) states that i's cost of some rule violation is psychologically hurting (whenever $k_i > 0$, and even more so when $\mathrm{E}_{\rho_i, s_j, \alpha_j}\left[m_j|h^0\right] - m_j(z) > 0$ for some $j \in N$, with $j \neq i$) and is high enough to make i's deviation from s_i^* unprofitable.[17]

Now, the above conditions for a social norm to exist are to be intended as those necessary for a behavioral rule r to be held in place: if fulfilled for every $i \in N$ a strategy profile dictated by r is an equilibrium, provided that all beliefs are correct and that players maximize expected utilities. Hence,

---

[15] The above set of conditions introduces a mathematically-precise definition of social norm, which formulizes Bicchieri's [2006] philosophical conditions. In this respect note that Bicchieri's construct differs from the present conditions, among the other issues, in that here if different players – incorrectly – expect different norms to be followed, then different behavioral rules constitute a social norm for each of the players. Obviously such a situation is impossible in equilibrium, where beliefs are correct.

[16] Notice that $A_{P(h),h}\left(r(h^0)\right)$ is the set of rule-complying actions of the active player at history h, as dictated at h^0 by r.

[17] Formally the sufficiently-large-k_i requirement implies that $k_i \geq \max\left\{\hat{k}_i^{s_i}, \ldots, \hat{k}_i^{s_i}\right\}$, where each $\hat{k}_i^{s_i}$ is a sensitivity parameter such that $u_i^C\left(z(s_i, s_{-i}^*)\right) = u_i^C\left(z(s_i^*, s_{-i}^*)\right)$ for some $s_i \in S_i$, with $s^* \in r(h^0)$.

definition 5 results in a social norm (existing and) being "followed by population N" if the conditions in remark 1 simultaneously hold.

Remark 1. A social norm r^* (exists and) *is followed* by population N if: every player $i \in N$ has conformist preferences represented by a utility function u_i^C, with $d_i^C = 1$, $d_i^E = 1$, and $k_i > 0$; every player i maximizes her expectation of u_i^C; every i holds correct beliefs about every j's ($j \in N$, with $j \neq i$) first-order belief and behavior; every player i's behavior is consistent with one of the end-nodes yielded by $r^* \in R_i \cap R_j$ (according to norm-conjectures $\rho_j = \rho_i$ for $\forall\, j \in N$); k_i is sufficiently large for every $i \in N$.

Note that the expression "a social norm r^* is followed by population N" (or "every player $i \in N$ conforms to r^*") implies that every player in the population plays her part of one of the strategy profiles contained in $r^*(h^0)$.

IV. Equilibrium concept

In this section an equilibrium concept for mixed-motive games with belief-dependent conformist preferences is discussed: by imposing the requirement that all beliefs (and norm-conjectures) are correct in equilibrium, I derive a "Social Sequential Equilibrium" as a special case of the sequential equilibrium notion of Kreps and Wilson [1982]. Kreps and Wilson's definition of equilibrium consists of sequentially rational, consistent assessments where: *(i)* An *assessment* is a profile of behavioral strategies and conditional first-order beliefs (along with higher-order beliefs in Battigalli and Dufwenberg's [2009] specification). *(ii)* An assessment is *consistent* if the profile $\alpha = (\alpha_i)_{i \in N}$ of first-order beliefs about the opponents' strategies is derived from the

behavioral strategy profile $\sigma = (\sigma_i)_{i \in N}$, that is, for $\forall\, i \in N$, $\forall\, s_{-i} \in S_{-i}$, $\forall\, h \in H_i$, it must be that $\alpha_i(s_{-i}|h) = \prod_{j \neq i} \Pr_{\sigma_j}(s_j|h)$;[18] given that, Battigalli and Dufwenberg's [2009] specification of sequential equilibria for psychological games extends the consistency requirement by demanding that higher-order beliefs at each information set are correct for $\forall\, i \in N$, $\forall\, h \in H_i$, that is, $\beta_i(h) = \alpha_{-i}$. *(iii)* Finally, an assessment is *sequentially rational* if, for every player i and every information set $h \in H_i$, the strategy of i is a best response to the other players' strategies given i's beliefs at h.

In the present framework I further extend the consistency requirement by imposing that Player i's (correct) beliefs about every j's first-order beliefs are derived from norm-conjectures ρ_i, with $\rho_i = \rho_j$ (for $\forall\, j \in N$, with $j \neq i$). It follows the definition of a *"socially consistent assessment"*.

Definition 6. A *socially consistent* assessment is a profile $(\sigma, \rho, \alpha, \beta) = (\sigma_i, \rho_i, \alpha_i, \beta_i)_{i \in N}$ that specifies behavioral strategies, norm-conjectures, first- and second-order beliefs, such that for $\forall\, i \in N$, $\forall\, s_{-i} \in S_{-i}$, $\forall\, h \in H_i$:

(i) $\alpha_i(s_{-i}|h) = \prod_{j \neq i} \Pr_{\sigma_j}(s_j|h)$;

(ii) $\beta_i(h) = \alpha_{-i}$;

(iii) $\beta_i^{S_i}(h^0) = \left(\Pr_{\rho_i}(s_i|h^0)\right)_{s_i \in S_i}$ *and*

$$\left(\rho_j(\cdot\,|h)\right)_{j \neq i,\, h \in H \setminus Z} = \left(\rho_i(\cdot\,|h)\right)_{h \in H \setminus Z}.$$

Notice that condition *(iii)* in definition 6 is the distinguishing feature of a socially consistent assessment in that it implies that (not only are beliefs

---

[18] Notice that, since α_i is derived from $\sigma = (\sigma_i)_{i \in N}$, the beliefs of every player $j \neq i$ about Player i's strategies must be the same.

derived from a behavioral strategy profile but also) a behavioral strategy profile $\sigma = (\sigma_i)_{i \in N}$ contains probability measures which equal those contained in every player's norm-conjecture ρ_i.[19]

The core equilibrium concept for mixed-motive games with belief-dependent conformist preferences can now be presented.

Definition 7. Given an extensive form game $G = \langle N, H, P, \left(u_i^C\right)_{i \in N} \rangle$ and a rule subset $R_i \subseteq R$ for each $i \in N$, a *Social Sequential Equilibrium* ("SSE") of G is a socially consistent assessment, such that for $\forall\, i \in N,\ \forall\, h \in H_i,\ \forall\, s_i^* \in S_i(h)$:

$$\Pr_{\sigma_i}(s_i^* | h) > 0 \Leftrightarrow \left(s_i^* \in S_i\left(R_i(h^0)\right) \textit{ and } s_i^* \in \arg\max_{s_i \in S_i(h)} \mathrm{E}_{s_i, \alpha_i, \beta_i, \rho_i}\left[u_i^C | h\right] \right).$$

In plain words, a socially consistent assessment is a social sequential equilibrium *iff* each probability measure $\Pr_{\sigma_i}(\cdot\,|h)$ assigns positive conditional probability only to conditional expected-payoff maximizing *rule-complying* strategies; that is, every player i holds the same conjecture about the actions consistent with some rule in R_i and maximizes the expectation of the utility function (given her correct belief systems). Note that here it is assumed common knowledge of the utility functions u_i^C, implying that the sensitivity parameters k_i are commonly known (and are, in effect, *sufficiently large*) as well as the fact that each player i knows that every $j \neq i$ adheres to some

---

[19] Following Kreps and Wilson [1982], condition *(i)* can be written under the assumption that there is a strictly positive sequence $\sigma^t \to \sigma$ such that each σ^t is *completely mixed* and each belief $\alpha_i^t(s_{-i}|h)$ is derived from σ^t using Bayes' rule. This allows not to restrict a player's belief system to information sets reached with positive probability only: in other words, the probability of events conditional on zero-probability events must approximate probabilities that are derived from behavioral strategies assigning positive probability to every action at every information set.

$r \in R_i \cap R_j$ (*i.e.*: resulting in $d_i^E = 1$), given that each player's rule subset is non-empty (*i.e.*: resulting in $d_i^C = 1$).[20]

So, it should be highlighted that if condition *1* of definition 5 (*i.e.*: $r \in R_i \Longrightarrow d_i^C = 1$) holds for every player $i, j \in N$ and every $i, j \in N$ holds correct norm-conjectures $\rho_j = \rho_i$ (as well as first- and second-order beliefs), then either of the following *equilibrium scenarios* is possible:

(i) conditions *2.1-2.2* of definition 5 hold for every player $i, j \in N \Longrightarrow$ social norm r^* exists (for $\forall\, i, j \in N$) and is followed by population $N \Longrightarrow$ a *social sequential equilibrium* of G occurs;

(ii) conditions *2.1-2.2* of definition 5 do not hold $\Longrightarrow$ social norm r does not exist for any player $i, j \in N$ (and it is not followed by population N) $\Longrightarrow$ a social sequential equilibrium of G does not occur (yet a *subgame perfect equilibrium* occurs if G is a game with observable actions; a *standard sequential equilibrium à la* Kreps and Wilson occurs otherwise).

Note that in scenario *(ii)* the utility function reduces to one of classical, non-conformist motivation, which justifies the standard notions of equilibrium adopted. It should be stressed that – for a given extensive form game G, and a

---

[20] Note that (if $d_i^C = 1$ and $d_i^E = 1$) one could define a consistent assessment *à la* Battigalli and Dufwenberg [2009] by dropping condition *(iii)* of definition 6 above; given that, their equilibrium notion can be obtained by dropping the requirement that each probability measure $\Pr_{\sigma_i}(\cdot\,|h)$ assign positive conditional probability only to rule-complying strategies in definition 7 above. Also note that every game with *simple guilt* has a sequential equilibrium *à la* Battigalli and Dufwenberg [2007] irrespective of the magnitude of the sensitivity parameter $(k_i)_{i \in N}$ (*i.e.*: parameter θ_{ij} in their notation); conversely, here, a player with utility function u_i^C has conditionally conformist preferences such that if, for some player $j \neq i$, j's cost of a rule violation is *not* high enough to make j's deviation from s_j^* unprofitable (*i.e.*: k_j is not sufficiently large), then condition *2.1* of definition 5 will not hold for every other player i (*i.e.*: resulting in $d_i^E = 0$) and so the utility function u_i^C will reduce to one of classical ("non-psychological") motivation, thereby implying a standard notion of equilibrium.

rule subset $R_i \subseteq R$ for each $i \in N$ – the existence of a social sequential equilibrium is ultimately conditional on players' sensitivity parameters $(k_i)_{i \in N}$: this sublimely captures the fragility of social norms in actual society.

That said, having assumed sufficiently large $(k_i)_{i \in N}$ parameters, an existence proof that relies on Selten's *trembling hand* argument can be conveniently adapted from Battigalli and Dufwenberg [2009].[21] Finally, the following result is a direct consequence of definition 7.

Remark 2. Given an extensive form game G, and a rule subset $R_i \subseteq R$ for each $i \in N$, if a certain social norm $r^* \in R_i$ (exists and) is followed by population N, then some social sequential equilibrium of G occurs.

Notice that the converse is not necessarily true as a certain socially consistent, sequentially rational assessment (*i.e.*: a social sequential equilibrium) might be induced by multiple behavioral rules in R, some of which may not even belong to R_i for some $i \in N$.[22]

---

[21] Battigalli and Dufwenberg point out that, in some cases, other solution concepts might depict the dynamics of certain types of belief-dependent motivations more satisfactorily than some variant of Kreps and Wilson's [1982] sequential equilibrium could do (*e.g.*: "weakly consistent perfect Bayesian equilibrium", self-confirming equilibrium, *etc.*). Space constraints do not permit further discussion here, but the reader may refer to Battigalli and Dufwenberg [2009] for a *psychological forward induction* argument.

[22] For example, consider a 2-player game and let each player's rule subset be defined as $R_i = \{r^E, r^M\}$. Then, assume that, at the root of the game, behavioral rule r^E dictates the strategy profiles $r^E(h^0) = \{(b,c),(a,c)\}$ whereas rule r^M dictates the strategy profiles $r^M(h^0) = \{(a,c),(a,d)\}$ (with each pair of lower-case letters denoting a strategy profile). Further, assume that both players have preferences represented by utility functions $(u_i^C)_{i \in N}$ and that, while holding correct beliefs, they play the strategy profile (a,c). Now, while (a,c) is a social sequential equilibrium of the game, this does not necessarily imply that, say, r^E (rather than r^M) constitutes a social norm and is being followed by the players of the game. Interestingly, this well captures the case of a traveller who, once in a foreign country, observes some locals interacting (without taking part in the actual game herself): while the outcome of the interaction may turn out to be compatible with some of the behavioral rules stored in the observer's mind, she may not be able to tell which one has been held in place, especially if the foreign country is particularly culturally-different from

V. Applications

In this section I turn to analyse a few dynamic interactions accounting for conditionally conformist preferences. Before that, I shall formally define a few behavioral rules reflecting principles which are usually assumed to regulate behavior in social dilemmas.[23]

- *Equality principle*:
  $r^E(H) = \{s \in S(h)\colon h \in H \setminus Z;\ z(s) \text{ s.t. } m_1(z) = \ldots = m_n(z)\}$.
- *Inequity-reducing principle*:
  $r^F(H) = \left\{s \in S(h)\colon h \in H \setminus Z;\ z(s) \text{ s.t. } z \in \arg\min_{z \in Z}\left(\frac{1}{n}\sum_{i \in N}[m_i(z) - \overline{m}(z)]^2\right)\right\}$.[24]
- *Classical-utilitarian welfare maximization principle*:
  $r^M(H) = \left\{\begin{matrix} s \in S(h)\colon \\ h \in H \setminus Z;\ z(s) \text{ s.t. } z \in \arg\max_{z \in Z}(\sum_{i \in N} m_i(z)) \end{matrix}\right\}$.
- *Rawlsian* (*minimax*) *welfare maximization principle*:
  $r^W(H) = \left\{s \in S(h)\colon h \in H \setminus Z;\ z(s) \text{ s.t. } z \in \arg\max_{z \in Z} W\left(\boldsymbol{m}(z)\right)\right\}$,
  where $W\left(\boldsymbol{m}(\hat{z})\right)$ denotes a *Rawlsian social welfare function* and is defined as $W\left(m_1(\hat{z}), \ldots, m_n(\hat{z})\right) = \min_{i \in N}\{m_1(\hat{z}), \ldots, m_n(\hat{z})\}$.

It should be stressed that the above rules do not aim at representing the whole range of norms that may emerge in strategic interactions but is only meant to provide a simple illustration of the conditions under which conformity sets in.

---

hers; on a smaller case, a similar problem occurs the first time we happen to interact with members of a group, organization or institution whose social norms we do not yet know.

[23] Note that, below, $r(H)$ denotes the union of the sets of strategy profiles dictated by the behavioral rule at each history $h \in H \setminus Z$.

[24] Note that $\overline{m}(z)$ denotes the mean value of the players' material payoffs, for a given terminal node z.

(Also, note that a behavioral rule could be constructed by combining two or more of the above principles.)

1. *Dictator Games*

Consider the following variant of the *dictator game*: as in the original version (Forsythe *et al.* [1994]) each subject is given an endowment to allocate; but, here, assume that whatever money the dictator gives to her co-player will be multiplied by a factor v.

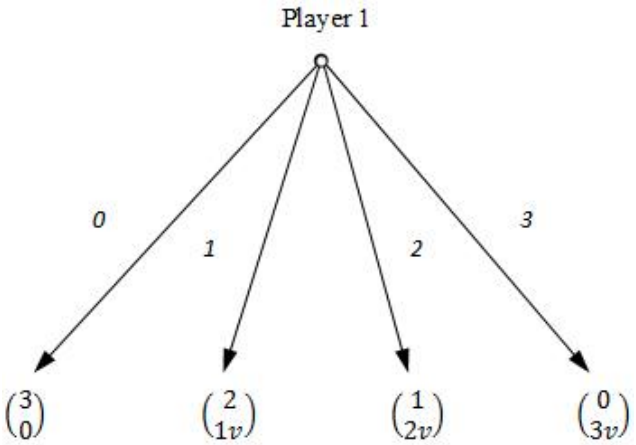

Figure 2 - Dictator Game with factorized donations *"FDG"*

For example, consider the game tree of Figure 2: Player 1 is endowed with \$3 and can choose to give any (integer) amount between 0 and 3 to Player 2; also, let $v \in \{0.5, 2\}$. It is straightforward to see that, when $v = 0.5$, the strategies dictated by the above-defined behavioral rules are as follows: $r^E(h^0) = \{2\}$, $r^F(h^0) = \{2\}$, $r^M(h^0) = \{0\}$ and $r^W(h^0) = \{2\}$; instead, when $v = 2$, $r^E(h^0) = \{1\}$, $r^F(h^0) = \{1\}$, $r^M(h^0) = \{3\}$ and $r^W(h^0) = \{1\}$.

Now, let the dictator's endowment and donation be denoted by M and x, respectively, with the donation being any integer $x \in [0, M]$. Below are a few results showing how the dictator's optimal donation (*i.e.*: action) varies with behavioral rules and factor v.

Proposition 1. Given the behavioral rules r^E, r^F, the only SSE of FDG is $x = \frac{M}{1+v}$, whenever $k_1 \geq \frac{M}{1+v(1+M)}$.

Proof: Firstly note that, for a given endowment M and for some factor v, the set of strategies dictated by behavioral rule $r^F(h^0)$ (or $r^E(h^0)$) is singleton, that is, $r^F(h^0) = \left\{x = \frac{M}{1+v}\right\}$. Hence, the norm-conjecture induced by r^F, for $\forall\, i \in N$, is such that $\rho_i\left(x = \frac{M}{1+v}\right) = 1$. Given that, Player 1 can form her belief β_1 by assuming her co-player's first-order beliefs to be consistent with her norm-conjecture: thus Player 1's expectation of Player 2's (expected) material payoff at h^0 equals $\mathrm{E}_{\rho_i,\beta_1}[m_2|h^0] = \rho_i\left(x = \frac{M}{1+v}\right) \cdot v\frac{M}{1+v}$; this implies that Player 1's expectation of Player 2's potential disappointment at $(x = 0|h^0)$ would equal $\mathrm{E}_{\rho_i,\beta_1}[m_2|h^0] - m_2\big(z(0)\big) = v\frac{M}{1+v} - 0 = v\frac{M}{1+v}$; in turn, this implies that Player 1's psychological utility at $(x = 0|h^0)$ would be given by $u_1^C(z(0), \rho_i, \beta_1) = M - \hat{k}_1^0\left[1 + \left(v\frac{M}{1+v}\right)\right]$. On the other hand, Player 1's utility (=payoff) at $\left(x = \frac{M}{1+v}|h^0\right)$ is given by $u_1^C\left(z\left(\frac{M}{1+v}\right), \rho_i, \beta_1\right) \equiv m_1\left(z\left(\frac{M}{1+v}\right)\right) = M - \frac{M}{1+v}$; it follows that Player 1's conformist preferences against $x = 0$ can be expressed as $u_1^C(z(0), \rho_i, \beta_1) \leq u_1^C\left(z\left(\frac{M}{1+v}\right), \rho_i, \beta_1\right) \Longrightarrow M - \hat{k}_1^0\left[1 + \left(v\frac{M}{1+v}\right)\right] \leq M - \frac{M}{1+v}$, which implies that $k_1 \geq \frac{M}{1+v(1+M)}$.

Corollary 1. Given the behavioral rule r^W, the only SSE of FDG is $x = \frac{M}{1+v}$, whenever $k_1 \geq \frac{M}{1+v(1+M)}$. **Proof**: See Appendix.

Proposition 2. Given the behavioral rule r^M, the following SSE of FDG may occur:

a) for $v < 1$, the only SSE is $x = 0$, $\forall\, k_1$ *or*

b) for $v > 1$, the only SSE is $x = M$ whenever $k_1 \geq \frac{M}{1+vM}$.

Proof: See Appendix.

Andreoni and Miller [2002] designed a similar experimental game, where each dictator was given a menu of choices with different endowments: specifically,

endowments were either 40, 60, 75, 80 or 100, while v varied between $1/3$ and 4. How does this relate to the above analysis? Quite clearly one can hypothesize that, in the experiment, a player with preferences for conformity to some equity principle would increase her allocation to her partner if she knew she had a low v, compared to the case of a high v (where she would decrease her donation); instead, a player with preferences for conformity to some efficiency principle would keep (almost) everything if v was less than 1, and give away (almost) everything if v was greater than 1. Now, Andreoni and Miller's experiment was not designed to test for belief-dependent conformist preferences, but simply to check whether people have consistent preferences across different rounds (featuring different M or v). Interestingly their experimental results show that, for different subjects, different forms of other-regarding principles were practiced, and that a majority of subjects behaved consistently across rounds. Indeed, about 40% of subjects exhibited selfish preferences, around 25% conformed to an equity principle, 11% maximized overall social welfare, and the remaining 24% acted idiosyncratically from round to round.[25]

Conventional social preference models find it hard to explain how subjects' actions may vary in accord with their knowledge about the partners' knowledge of a one-to-one mapping between – fair *vs.* unfair – actions and outcomes. In this respect, one of Dana *et al.*'s [2007] experimental treatments allowed dictators the possibility of losing agency if they did not choose an action within a relatively long time interval (in a standard dictator game). More precisely, dictators were instructed that they would have a 10-second

[25] In the absence of data about subjects' conjectures, one might well assume that those subjects exhibiting consistent preferences believed it was appropriate to behave according to some principle.

interval during which to enter their choices and that, if they had not chosen an action at a randomly selected point in the interval, the computer program would cut them off and choose between an equal and an unequal payoff allocation (the latter being advantageous to the dictator), with same probability. Given that only the dictator would be notified if a cut-off occurred, the respective receiver could not tell whether her payoff was determined by the dictator's action or the computer program: so, in the eyes of the receiver, this feature made it plausible for unequal outcomes to have resulted from a random device even in the case in which they were actually due to the dictator's action; interestingly, Dana *et al.*'s results show that – among dictators that were not cut off – a majority picked the selfish action, that is, a proportion higher than the proportion of selfish choices in the baseline treatment (where there was no possibility of being cut off). Their results further show that the dictators' response time was often longer than in the baseline treatment, a fact that resulted in almost 1 in 4 dictators being cut off; therefore, it seems that many subjects were willing to delay making a choice, perhaps trying to avoid the responsibility of making an unfair choice. Now, the present theory is capable of making sense of these results, as follows: in the baseline treatment the subjects' rule subset could be defined as $R_i = \{r^F\}$ while in the above-mentioned treatment it could be defined as $R_i = \{r^F, r^X\}$, with r^X being a "random device rule" (which prescribes that strategies be chosen through a random device). Given that *any* final outcome is compatible with the random device rule – and given that *only* the dictator was notified if a cut-off occurred – it turns out that *dictators exploited this asymmetry*: *(i)* by choosing the selfish action more often than they did in the

baseline treatment (thinking that the receiver would believe that the dictator had been cut off, that is, had followed the random device rule); *(ii)* or else by delaying their choice in order to be cut off (hence effectively following the random device rule).[26]

2. *Ultimatum Games*

The *ultimatum game* (Güth *et al.* [1982]) provides a simple 2-player model of bargaining, where Player 1 (the "proposer") suggests how to divide a given sum M, and Player 2 (the "responder") can either accept or reject this proposal: if the responder rejects, neither player receives anything; if she accepts, the money is split according to the proposal (*i.e.*: $M - x$ to Player 1, and x to Player 2). Assume the proposer's set of actions contains integers in the interval $[0, M]$ (where M is an even number) and the responder's set of actions at each decision node is $\{Y, N\}$, with Y denoting acceptance and N denoting rejection. In this case, $r^F(h^0)$ (or $r^E(h^0)$) contains strategy profiles whereby the proposer offers *any* amount, and the responder *either* accepts *or* declines an offer $x = M/2$,[27] and declines all offers other than $x = M/2$.

Proposition 3. Given the behavioral rules r^E, r^F, there is only one SSE of the ultimatum game: for $\rho_i(s_1 = M/2|h^0) = 1$, $\rho_i\left(s_{2,M/2} = Y|M/2\right) = 1$ and

---

[26] It should be stressed that this interpretation of Dana *et al.*'s [2007] results is founded on the assumption that in many games people consider choosing through a random device *as appropriate as* picking an equal allocation in the first place (see Bicchieri and Chavez [2013] for strong experimental evidence in support of such an assumption). Formally, defining the players' rule subset as $R_i = \{r^F, r^X\}$ implies that norm-conjectures ρ_i are such that any one of the dictator's actions could be assigned probability 1, which in turn implies that the dictator's utility from choosing an unequal payoff allocation would not involve a psychological loss (as, in any case, there would be no deviation from ρ_i).

[27] In fact, whether the responder accepts or declines a 50-50 offer, players will get an equal amount of money (for a given terminal node), *i.e.*: in the case of acceptance each will get half the sum M, whereas in the case of rejection each will get \$0.

$\rho_i\left(s_{2,x\neq M/2} = Y | x \neq M/2\right) = 0$; the only SSE is given by $\left(\frac{M}{2}, Y\ for\ \frac{M}{2}\ and\ N\ for\ x \neq \frac{M}{2}\right)$ whenever $k_2 \geq M$. **Proof**: See Appendix.[28]

Unlike models that combine intentionality with distributional concerns, like Falk-Fischbacher [2006] or Dufwenberg-Kirchsteiger [2004], the present theory is *less indeterminate* (given some inequity-reducing principle such as r^F), since it only involves one psychological equilibrium besides the standard subgame perfect equilibrium.[29] In this respect, the reader might object that the present theory is *more indeterminate* than the others if one considers alternative behavioral rules (*e.g.*: either r^F, r^M, r^W or r^X). However, in the laboratory this kind of indeterminacy in predictions is easily overcome if the game of interest is preceded by a pre-play stage featuring an elicitation method so as to obtain ratings of the extent to which different actions (in different hypothetical games) are believed to be collectively perceived as appropriate; then, from the actions that are considered appropriate, the experimenter can

---

[28] It should be noted that the strategy profile $\left(\frac{M}{2}, Y\ for\ x \geq \frac{M}{2}\ and\ N\ for\ x < \frac{M}{2}\right)$ would *not* qualify as an SSE – given r^F – because (as per definition 7 above) in equilibrium every player should assign positive probability to rule-complying actions, if any, also at decision nodes off the equilibrium path. In fact, it should be recalled that r^F dictates a strategy profile that minimizes differences in material outcomes, which implies that a conditionally conformist responder should be willing to give up an offer $x > M/2$ if the value she attaches to principles (*i.e.*: k_2) is greater than the material payoffs at stake: one may think of it as unwillingness to accept an undue inducement. This goes beyond the purposes of the present theory, but of course one can easily define a less stringent rule such that it dictates the strategy profile $\left(\frac{M}{2}, Y\ for\ x \geq \frac{M}{2}\ and\ N\ for\ x < \frac{M}{2}\right)$, in order to make more realistic predictions. In this regard, experimental data (Camerer [2003], Ch. 2) are partly consistent with the unique SSE yielded by the over-simplistic r^F, since modal offers are usually 40 to 50 percent and such offers are rarely rejected; also, there are hardly any offers in the categories 0 to 10 percent and over 50 percent (but it is hard to believe that someone's k_2 could be so large to reject an offer over 50 percent in the lab, as dictated by r^F).

[29] In those models responders have thresholds of offers they always accept or reject (based on the subjects' fairness sensitivity), which in effect yield multiple equilibria.

deduce some general rules and can therefore apply such rules to make predictions in other mixed-motive games. (In this connection, Krupka and Weber [2013] make use of an incentivized pre-play elicitation method for identifying social norms, which uses simple coordination games where people guess what is believed to be more or less appropriate in each context.) For instance in the case of the ultimatum game if, in some pre-play stage, subjects expressed preferences compatible with an efficiency rule such as the classical-utilitarian welfare maximization principle r^M, then the present theory would predict a unique SSE where – for any positive k_2 – the proposer offers \$0 and the responder accepts all offers.[30]

3. *Trust Games*

Consider the following *trust game*. At the initial node h^0, Player 1 (the "trustor") chooses either "a" or "b": when opting for "b" the game terminates and material outcomes are allocated as shown in the vector of payoffs at the end-node $z(b)$; if Player 1 opts for "a" the choice passes to Player 2 (the "trustee"), who in turn can decide on "c" or "d", the consequences of which are shown in the vector of payoffs at the end-nodes $z(c)$ and $z(d)$, respectively. Let the parameters x and v be such that $x > 0$ and $v > 1$.

---

[30] Similarly, some algebra shows that if subjects expressed preferences for conformity to a Rawlsian (minimax) welfare maximization principle r^W, then the present theory would predict a unique SSE whereby the proposer offers $M/2$ whenever $k_1 \geq \frac{M}{2+M}$, and the responder accepts all offers.

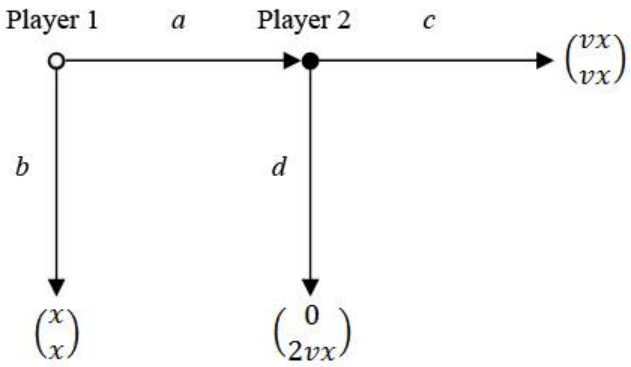

Figure 3 - Trust Game *"TG"*

The following results refer to potential, alternative specifications of the rule subsets.

Proposition 4. Given the behavioral rules r^E, r^F, the only SSE of TG is (a, c), whenever $k_2 \geq \frac{vx}{1+vx}$. **Proof**: See Appendix.

Corollary 2. Given the behavioral rule r^W, the only SSE of TG is (a, c), whenever $k_2 \geq \frac{vx}{1+vx}$. **Proof**: See Appendix.

Proposition 5. Given the behavioral rule r^M, the following SSE of TG may occur:

a) for $\rho_i(c|h(a)) = 1$; (a, c), whenever $k_2 \geq \frac{vx}{1+vx}$ *or*

b) for $\rho_i(c|h(a)) = 0$; (a, d), whenever $k_1 \geq \frac{x}{1+2vx-x}$.

Proof: See Appendix.

Notice that scenario *(b)* in proposition 5 (*i.e.*: SSE (a, d)) provides an instance of a socially undesirable solution: in fact, Player 1 conforms to r^M because her cost of a norm violation is high enough to make her deviation from $s_1^* = a$ unprofitable (if $k_1 \geq \frac{x}{1+2vx-x}$).[31]

The above exercise suggests that the range of equilibria observed across experimental trust games might vary with conjectures about norms, with such beliefs being induced by a variety of context- and culture-dependent principles. For example, Xiao and Bicchieri [2010] analyse a trust game similar to the one above (but where the trustee's action set contains multiple options), and compare experimental results with a treatment variant in which differences in the payoff distribution at $z(b)$ make it possible for the precepts of the *equality* principle and a *reciprocity* principle to conflict. It should be noted that, unlike the equality rule r^E – which is defined with reference to material payoffs at terminal nodes only – the reciprocity principle takes into account also actions: that is, the reciprocity rule r^C can be defined so as to dictate a strategy profile where the trustor passes an amount x (choosing a, in the above notation) on to the trustee, and the trustee chooses any action (this time among her multiple options) such that she returns at least x to the trustor.

---

[31] Even though an equilibrium consisting of the strategy profile (a, d) may at first seem an unrealistic solution, this actually captures many situations characterized by the internalization of a socially undesirable norm: an example is given by a set of circumstances where a woman (marries and) brings a dowry to a man of dubious reputation; in effect, she (Player 1) may lucidly expect the man (Player 2) to use and invest the dowry, keeping all the proceeds for himself and, still, she may prefer to marry him if the local culture pushes women to get a husband. Thus if the cost of deviating is high, the influence of culture and its norms is such that the woman is indifferent between remaining unmarried (*i.e.*: $(s_1, s_2) = (b, \cdot)$) and getting married-but-losing-everything (*i.e.*: $(s_1, s_2) = (a, d)$). Another example of a social norm inducing an extremely undesirable outcome is female genital mutilation. In a number of countries in Africa and the Middle East this practice is supported by both men and women: most interestingly, in the majority of cases it is particularly supported by women as they consider such a practice a source of authority and honour (Bicchieri [2013]).

Xiao and Bicchieri's [2010] experimental results show that the trustees' normative expectations are consistent with a reciprocity principle only when it is in their interest, and are otherwise consistent with an equality principle; so, the conclusion is that different behavioral rules are made salient in different experimental treatments, and solutions vary accordingly.[32]

Furthermore, there is ample consensus that different cultures give prominence to different behavioral rules and, hence, different conjectures about norms. In this connection, Johnson and Mislin [2011] have collected data from 162 replications of the original Berg *et al.* [1995] trust game (with a total of 23,000 participants) so as to identify the effect of experimental protocols and geographic variation on trust and trustworthiness. Among the other things, their findings show that trustworthiness is significantly affected by the factor by which the experimenter multiplies the amount sent (*i.e.*: v, in the above notation). Moreover, Johnson and Mislin find robust evidence that subjects send less in trust games conducted in Africa than those in North America, which might indicate that people in Africa are sensitive to behavioral rules different from those followed in North America.

4. *Public Goods Games*

In a *public goods game* each of n players is endowed with a sum of money M, and can voluntarily invest part (or all) of it in a public good that has a total per-unit value of $v > 1$. Let Player i's contribution be any (integer) amount

---

[32] In terms of the present theory, their results can be interpreted as follows: first, define the players' rule subset as $R_i = \{r^E, r^C\}$, then let the active player select an action (among the admissible actions as determined by R_i) that best serves her interests and accordingly fix everyone's norm-conjectures ρ_i. More concretely, if R_i is defined as above for all players, and a player chooses an action which is consistent with only one of those rules, say r^E, then this means that all players *will have* formed their norm-conjectures in such a way as to assign positive probability only to actions prescribed by r^E.

$x_i \in [0, M]$, and her payoff function be given by $m_i\big(z(x_i, \dots, x_n)\big) = M - x_i + \frac{v\sum_j^n x_j}{n}$. Let the game be modelled as a sequential game with imperfect information, where players do not know what actions were chosen by previous players.

Proposition 6. Given the behavioral rule r^M, the only SSE of the public goods game involves a strategy profile where every player's contribution equals $x_i = M$ whenever $k_i \geq \frac{M(n-v)}{n+vM(n-1)}$ for $\forall\, i \in N$. **Proof**: See Appendix.[33]

Corollary 3. Given the behavioral rules r^E, r^F, there is a set of SSE of the public goods game, in which each equilibrium involves a strategy profile where every player's contribution equals $x_i = x$ whenever $k_i \geq \frac{x(n-v)}{n+vx(n-1)}$ for $\forall\, i \in N$. (The proof is analogous to that of proposition 6, and is therefore omitted.)

It should be noted that proposition 6 implies that, when the value v is greater than the number of players n, it is always optimal to follow the social norm: obviously, in that case the return on investment is a motivation sufficiently strong to make one follow the norm. On the other hand, when $(n - v)$ is positive, the value a player attaches to principles (*i.e.*: k_i) must provide a sufficiently strong motivation for one to be willing to conform.

Once again, the above exercise suggests that the range of equilibria observed across experimental public goods games might vary with conjectures

---

[33] Note that the Rawlsian (minimax) welfare maximization principle r^W yields the same SSE.

about norms,[34] with such beliefs being induced by a variety of context- and culture-dependent principles. The intuition is confirmed by a large cross-cultural experimental study undertaken in fifteen small-scale societies (Henrich *et al.* [2001]): the investigation addressed the question of whether the individuals' social environments shape behavior, by recruiting experimental subjects from small-scale societies that present a wide variety of economic and cultural features. Henrich *et al.*'s [2001] results show that *group*-level differences in the structure and organization of everyday economic activity explain a substantial part of the experimental variation observed across societies; in other words, the higher the degree of market integration and the higher the payoffs to cooperation of everyday life, the greater the level of cooperation in experimental games.

VI. Closing Remarks

This essay has presented an original theory of conformist preferences in social dilemmas, building on Battigalli and Dufwenberg's [2009] framework for the analysis of dynamic psychological games. The present theory departs from the

---

[34] In this regard, note that Fischbacher and Gächter [2010] report a large degree of preference heterogeneity in public goods games: in a first experiment, using a variant of Selten's strategy method, they assessed subjects' willingness to contribute; their data show that about 55% of subjects linearly condition their contributions on the contribution levels of others, 23% of subjects exhibit selfish preferences, about 12% of subjects increase their own contributions with the contribution levels of others up to a point (and then decrease their own contributions with increasing levels of others' contributions), and the remaining 10% of subjects exhibit idiosyncratic preferences. In a second experiment, Fischbacher and Gächter elicited subjects' beliefs about the others' contributions, and found that subjects' contributions depend directly on such beliefs; also, they found that subjects are on average "imperfect conditional cooperators" in that they match others' contributions only partly (*e.g.*: by contributing a little less than they expect others to contribute), which may explain why contributions decline in repeated public goods games. To conclude, Fischbacher and Gächter [2010] have designed their investigation so that they could use the first experiment to make a point prediction for each participant about her contribution in the second experiment (given her beliefs); in terms of the present theory, this roughly corresponds to *using the first experiment to derive a behavioral rule for each participant, and then using such a rule to make predictions in the second experiment*.

existing game-theoretic literature on social preferences, since it conceives of social norms as equilibrium selection devices while providing a set of conditions for a social norm to exist. Although the motivational factors considered here are related to the much-investigated concepts of fairness and reciprocity, it should be noted that the focus of this study has been on a "mere" conformity motivation in social dilemmas, implying that the peers' (presumed) behavior and expectations – be it fair or not – serve the individual as a means to guiding her own actions.

It should be stressed that the focus of this study has been on why rules are followed, rather than on the specifics of what the rules are. This implies that the present theory can account for the reasons that have led to the *perpetuation* of a given norm, but not for the reasons that have led to the evolution of an individual's rule subset (which is exogenously determined) and consequent norm-conjectures. Notice that this is due to the fact that the model partly relies on past behavior to explain future compliance: in effect, it is the individual's culture that marks out each player's rule subset so that it contains rules of behavior in accordance with set usage; as a consequence, this theory implies a tendency for individuals to conform to the presumed "currently-normal" behavior.

Now, the above considerations might seem to limit the potential for policy application of this theory in that it relies on an exogenous (culture-dependent) specification of the rule subsets, which implies that the system will *not* evolve away from its current position *unless* some variation in conjectures about norms occurs. But it is precisely because of the fact that social norms depend on such conjectures that this theory suggests that – if beliefs are "manipulated" – it may be possible to induce pro-social behavior at low cost! Indeed, a finely-tuned process of belief transmission can effectively favour the occurrence of the desired policy outcome: for instance, social psychology research conducted at several U.S. universities shows that students hold exaggerated beliefs about the alcohol consumption habits of their peers

(Berkowitz and Perkins [1986]). Such studies have concluded that students consume greater quantities of alcohol in order to fit in with their perceptions of acceptable social behavior, that is, in order to comply with their presumed drinking norm in operation on campus. Research further shows that students that participate in a peer-oriented discussion (focusing on correcting biased perceptions) report drinking significantly less: in particular, a study from the *National Institute on Alcohol Abuse and Alcoholism*, an agency of the United States Department of Health and Human Services, shows that several educational institutions that consistently organize peer-oriented discussions have experienced reductions of up to twenty percent in high-risk drinking over a relatively short period of time (NIAAA [2002]).

VII. References

Akerlof, George A. 1980. "A Theory of Social Custom, of Which Unemployment May be One Consequence" *The Quarterly Journal of Economics*, 94(4): 749-775.

Andreoni, James and John Miller. 2002. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism" *Econometrica*, 70(2): 737-753.

Arrow, Kenneth J. 1972. "Models of Job Discrimination" in *Racial Discrimination in Economic Life*, ed. Anthony H. Pascal. Lexington, Mass.: Heath.

Battigalli, Pierpaolo and Martin Dufwenberg. 2007. "Guilt in Games" *American Economic Review: Papers and Proceedings*, 97(2): 170-176.

------. 2009. "Dynamic Psychological Games" *Journal of Economic Theory*, 144(1): 1-35.

Berg, Joyce, John Dickhaut and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History" *Games and Economic Behavior*, 10(1): 122-142.

Berkowitz, Alan D. and H. Wesley Perkins. 1986. "Problem Drinking Among College Students: A Review of Recent Research" *Journal of American College Health*, 35: 21-28.

Bicchieri, Cristina. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

------. 2013. *Norms in the Wild*. Cambridge: Cambridge University Press. Forthcoming.

Bicchieri, Cristina and Alex K. Chavez. 2013. "Norm Manipulation, Norm Evasion: Experimental Evidence" *Economics and Philosophy*, 29(2): 175-198.

Bolton, Gary E. and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition" *The American Economic Review*, 90(1): 166-193.

Camerer, Colin F. 2003. *Behavioral Game Theory. Experiments in Strategic Interaction*. Princeton, N.J.: Princeton University Press.

Charness, Gary and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests" *The Quarterly Journal of Economics*, 117(3): 817-869.

Cialdini, Robert B. and Noah J. Goldstein. 2004. "Social Influence: Compliance and Conformity" *Annual Review of Psychology*, 55(1): 591-621.

Dana, Jason, Roberto A. Weber and Jason Xi Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness" *Economic Theory*, 33(1): 67-80.

Dufwenberg, Martin and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity" *Games and Economic Behavior*, 47(2): 268-298.
Elster, Jon. 1989. *The Cement of Society*. Cambridge: Cambridge University Press.
Falk, Armin and Urs Fischbacher. 2006. "A Theory of Reciprocity" *Games and Economic Behavior*, 54(2): 293-315.
Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation" *The Quarterly Journal of Economics*, 114(3): 817-868.
------. 2006. "The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories" in *Handbook of the Economics of Giving, Altruism and Reciprocity: Vol. 1*, ed. Serge-Christophe Kolm and Jean Mercier Ythier. Amsterdam: North-Holland/Elsevier.
Fischbacher, Urs and Simon Gächter. 2010. "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments" *The American Economic Review*, 100(1): 541-556.
Forsythe, Robert, Joel L. Horowitz, N. E. Savin and Martin Sefton. 1994. "Fairness in Simple Bargaining Experiments" *Games and Economic Behavior*, 6(3): 347-369.
Güth, Werner, Rolf Schmittberger and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining" *Journal of Economic Behavior & Organization*, 3(4): 367–388.
Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis and Richard McElreath. 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies" *The American Economic Review*, 91(2): 73-78.
Johnson, Noel D. and Alexandra A. Mislin. 2011. "Trust Games: A Meta-Analysis" *Journal of Economic Psychology*, 32(5): 865-889.
Klucharev, Vasily, Kaisa Hytőnen, Mark Rijpkema, Ale Smidts and Guillén Fernández. 2009. "Reinforcement Learning Signal Predicts Social Conformity" *Neuron*, 61(1): 140-151.
Kreps, David M. and Robert Wilson. 1982. "Sequential Equilibria" *Econometrica*, 50(4): 863-894.
Krupka, Erin L. and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11(3): 495-524.
Ledyard, John. 1995. "Public Goods Experiments" in *The Handbook of Experimental Economics*, ed. John H. Kagel and Alvin E. Roth. Princeton, N.J.: Princeton University Press.
López-Pérez, Raúl. 2008. "Aversion to Norm-Breaking: A Model" *Games and Economic Behavior*, 64(1): 237-267.
Montague, P. Read and Terry Lohrenz. 2007. "To Detect and Correct: Norm Violations and Their Enforcement" *Neuron*, 56(1): 14-18.
NIAAA National Advisory Council on Alcohol Abuse and Alcoholism. 2002. "How to Reduce High-Risk College Drinking: Use Proven Strategies, Fill Research Gaps" *NIAAA College Materials*. National Institutes of Health: U. S. Department of Health and Human Services. Available: http://www.collegedrinkingprevention.gov/media/FINALPanel2.pdf (Accessed: 2013, November 20).
Osborne, Martin J. and Ariel Rubinstein. 1994. *A Course in Game Theory*. Cambridge, Mass.: MIT Press.
Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics" *The American Economic Review*, 83(5): 1281-1302.
Sacco, Pier Luigi, Paolo Vanin and Stefano Zamagni. 2006. "The Economics of Human Relationships" in *Handbook of the Economics of Giving, Altruism and Reciprocity: Vol. 1*, ed. Serge-Christophe Kolm and Jean Mercier Ythier. Amsterdam: North-Holland/Elsevier.
Sugden, Robert. 1984. "The Supply of Public Goods through Voluntary Contributions" *The Economic Journal*, 94(376): 772-787.
------. 2000. "The Motivating Power of Expectations" in *Rationality, Rules and Structure*, ed. Julian Nida-Rümelin and Wolfgang Spohn. Amsterdam: Kluwer.
Xiao, Erte and Cristina Bicchieri. 2010. "When Equality Trumps Reciprocity" *Journal of Economic Psychology*, 31(3): 456-470.

VIII. For Online Publication: Appendix

Proof of Corollary 1. Given an end-node $\hat{z} \in Z$, a Rawlsian social welfare function is defined as $W\big(m_1(\hat{z}), m_2(\hat{z})\big) = \min_{i \in N} \{m_1(\hat{z}), m_2(\hat{z})\}$; such a function has to be evaluated at each of the $M+1$ end-nodes of the game tree. Then, it is straightforward to see that here the set of maximizers of W is singleton: so, for a given endowment M and for some factor v, the set of strategies dictated by behavioral rule $r^W(h^0)$ is singleton, that is, $r^W(h^0) = \left\{x = \frac{M}{1+v}\right\}$. The rest of the proof is analogous to that of proposition 1.

Proof of Proposition 2. ***(a)*** The proof of the first equilibrium is straightforward, and is therefore omitted. ***(b)*** As for the second (somewhat extreme) scenario, note that for a given endowment M and for some factor v, the set of strategies dictated by behavioral rule $r^M(h^0)$ is singleton, that is, $r^M(h^0) = \{x = M\}$. Hence, the norm-conjecture induced by r^M, for $\forall\, i \in N$, is such that $\rho_i(x = M) = 1$. Given that, Player 1 can form her belief β_1 by assuming her co-player's first-order beliefs to be consistent with her norm-conjecture: thus Player 1's expectation of Player 2's (expected) material payoff at h^0 equals $\mathrm{E}_{\rho_i,\beta_1}[m_2|h^0] = \rho_i(x = M) \cdot vM = vM$; this implies that Player 1's expectation of Player 2's potential disappointment at $(x = 0|h^0)$ would equal $\mathrm{E}_{\rho_i,\beta_1}[m_2|h^0] - m_2\big(z(0)\big) = vM - 0 = vM$; in turn, this implies that Player 1's psychological utility at $(x = 0|h^0)$ would be given by $u_1^C(z(0), \rho_i, \beta_1) = M - \hat{k}_1^0[1 + (vM)]$. On the other hand, Player 1's utility (=payoff) at $(x = M|h^0)$ is given by $u_1^C(z(M), \rho_i, \beta_1) \equiv m_1\big(z(M)\big) = 0$; it follows that Player 1's conformist preferences against $x = 0$ can be expressed as $u_1^C(z(0), \rho_i, \beta_1) \leq u_1^C(z(M), \rho_i, \beta_1) \Longrightarrow M - \hat{k}_1^0[1 + (vM)] \leq 0$, which implies that $k_1 \geq \frac{M}{1+vM}$.

Proof of Proposition 3. Recall that $r^F(h^0)$ (or $r^E(h^0)$) contains strategy profiles whereby the proposer offers any amount, and the responder either accepts or declines an offer $x = M/2$, and declines all offers other than $x = M/2$. This implies that the norm-conjecture induced by r^F, for $\forall\, i \in N$, is such that: ρ_i may take on value 1 for any one of Player 1's actions; and ρ_i takes on value 0 for all actions Y following an offer $x \neq M/2$, whereas ρ_i may take on either value 0 or value 1 for action Y following $M/2$. Given that, Player 1 can form her belief α_1 by assuming her co-player's behavior to be consistent with her norm-conjecture: Player 1's initial belief $\alpha_1 = (\cdot\,|h^0)$ corresponds to a probability measure over the strategies of the opponent, with the support of α_1 containing only Player 2's rule-complying strategies; hence, Player 1 can calculate her expected utility from each of her actions, as follows. First, denoting the probability that Player 2 accepts an $M/2$ offer as $\hat{\rho}$ (*i.e.*: $\hat{\rho} := \rho_i\big(s_{2,M/2} = Y|M/2\big)$), Player 1's expectation of Player 2's (expected) material payoff at h^0 equals $\mathrm{E}_{\rho_i,\alpha_1,\beta_1}[m_2|h^0] = \rho_i(M/2|h^0) \cdot \left(\hat{\rho} \cdot \frac{M}{2} + (1-\hat{\rho}) \cdot 0\right) = \rho_i(M/2|h^0) \cdot \hat{\rho} M/2$. Further, if Player 1 chooses any action $x \neq M/2$, then Player 2 will update her beliefs based on the fact that $\rho_i(x \neq M/2|h^0) = 1$ for a certain

$x \in A_1$; this implies that Player 1's expectation of Player 2's potential disappointment at $(x \neq M/2|h^0)$ would be null as there is no deviation, and the above expression for $\mathrm{E}_{\rho_i,\alpha_1,\beta_1}[m_2|h^0]$ is simply updated so that $\rho_i(M/2|h^0) = 0$; in turn, this implies that Player 1's expected utility at $h(x \neq M/2)$ would not involve a psychological loss. Now, before considering Player 1's remaining action (*i.e.*: offering $M/2$), let's look at the strategic interaction from Player 2's perspective: in order to calculate the optimal action at each history after which Player 2 has to move, she will compare her utility from conforming against the utility from deviating from the presumed norm (*e.g.*: accepting an offer $x \neq M/2$); so, for all histories following actions $x \neq M/2$, Player 2's utility from deviating would equal $u_2^C(z, \rho_i, \beta_2) = x - \hat{k}_2^x[1+0] = x - \hat{k}_2^x$;[35] instead, Player 2's utility at $\left(N|x \neq \frac{M}{2}\right)$ would simply correspond to her material payoff (*i.e.*: $u_2^C(z(x,N), \rho_i, \beta_2) \equiv m_2\big(z(N)\big) = 0$). In brief, Player 2's conformist preferences can be expressed as: $u_2^C(z(x,Y), \rho_i, \beta_2) \leq u_2^C(z(x,N), \rho_i, \beta_2) \Longrightarrow x - \hat{k}_2^x \leq 0 \Longrightarrow \hat{k}_2^x \geq x$, which means that – given some offer $x \neq M/2$ – Player 2 will prefer to conform (thereby rejecting offer x) if her sensitivity parameter is weakly greater than x; then, the sufficiently-large-k_i requirement (for social norms r^E, r^F to exist for Player 2) imposes that $k_2 \geq \max\left\{\left(\hat{k}_2^x\right)_{x=0,\dots,M}\right\}$, that is, $k_2 \geq M$. On the other hand, if Player 1 offered $M/2$, Player 2 would fortify her belief that $\hat{\rho} := \rho_i\big(s_{2,M/2} = Y|M/2\big) = 1$ (in fact, it is mutually beneficial), which implies that Player 2's preferences, when $\hat{\rho} = 1$, are such that $u_2^C\left(z\left(\frac{M}{2}, N\right), \rho_i, \beta_2\right) < u_2^C\left(z\left(\frac{M}{2}, Y\right), \rho_i, \beta_2\right)$. Player 1 will figure this out and indeed make an $M/2$ offer, which Player 2 will accept. To conclude, recalling that a socially consistent assessment is an SSE (definition 7 above) if each probability measure $\mathrm{Pr}_{\sigma_i}(\cdot\,|h)$ assigns positive conditional probability only to conditional expected-payoff maximizing rule-complying strategies, it follows that the only SSE is the one given by proposition 3: in plain words, the proposer will make a 50-50 offer, and the responder will reject any offer other than that (whenever $k_2 \geq M$). In contrast, if $k_2 < M$ behavioral rule r^F is not a social norm and is not followed by population N (by remark 1 above).

Proof of Proposition 4. Firstly, note that $r^F(h^0)$ (or $r^E(h^0)$) contains the following dictated strategy profiles: $r^F(h^0) = \{(b,c),(a,c)\}$. So the norm-conjecture induced by r^F, for $\forall\, i \in N$, can be represented by the following matrix:

---

[35] Notice that Player 2's expectation of Player 1's (expected) material payoff at $(x \neq M/2|h^0)$ equals $\mathrm{E}_{\rho_i,\beta_2}\left[m_1|x \neq \frac{M}{2}\right] = 0$.

$$\rho_i = \begin{bmatrix} \rho_i(a|h^0) & \rho_i(b|h^0) \\ \rho_i\big(c|h(a)\big) & \rho_i\big(d|h(a)\big) \end{bmatrix} = \begin{bmatrix} \hat{\rho} & 1-\hat{\rho} \\ 1 & 0 \end{bmatrix},$$

where $\hat{\rho} \in \{0,1\}$; that is, if $\hat{\rho} = 0$ then strategy profile (b, c) is implemented, whereas if $\hat{\rho} = 1$ then (a, c) is implemented. Given that, Player 1 can form her belief α_1 by assuming her co-player's behavior to be consistent with her norm-conjecture: this way Player 1 can calculate her expected payoff as well as the opponent's expected payoff and potential disappointment from 1's not conforming to the presumed norm. In brief, Player 1's expectation of Player 2's (expected) material payoff at h^0 equals $\mathrm{E}_{\rho_i,\alpha_1,\beta_1}[m_2|h^0] = vx\hat{\rho} + x(1-\hat{\rho})$. Further, if Player 1 chose action b, then Player 2 would update her beliefs based on the fact that $\hat{\rho} = 0$: this implies that Player 1's expectation of Player 2's potential disappointment at $(b|h^0)$ would be null as $\hat{\rho} = 0$ in the expression $\mathrm{E}_{\rho_i,\alpha_1,\beta_1}[m_2|h^0] - m_2\big(z(b)\big) = vx\hat{\rho} + x(1-\hat{\rho}) - x = vx\hat{\rho} - x\hat{\rho}$; in turn, this implies that Player 1's utility at $(b|h^0)$ would simply correspond to her material payoff because there would be no deviation from ρ_i, hence no psychological loss (*i.e.*: $u_1^C(z(b), \rho_i, \alpha_1, \beta_1) \equiv m_1\big(z(b)\big) = x$). On the other hand, Player 1's expected utility (=expected payoff) at $h(a)$ is given by $\mathrm{E}_{\rho_i,\alpha_1}[m_1|h(a)] = 1 \cdot vx = vx$; it follows that Player 1's conformist preferences can be expressed as $u_1^C(z(b), \rho_i, \alpha_1, \beta_1) \leq \mathrm{E}_{\rho_i,\alpha_1}[m_1|h(a)] \Longrightarrow x < vx$, which is always satisfied because $v > 1$. Similarly, Player 2's expectation of Player 1's (expected) material payoff at h^0 equals $\mathrm{E}_{\rho_i,\alpha_2,\beta_2}[m_1|h^0] = vx\hat{\rho} + x(1-\hat{\rho})$. But if Player 1 chooses action a, then Player 2 will update her beliefs based on the fact that $\hat{\rho} = 1$: this implies that Player 2's expectation of Player 1's potential disappointment at $\big(d|h(a)\big)$ would equal $\mathrm{E}_{\rho_i,\beta_2}[m_1|h(a)] - m_1\big(z(d)\big) = vx \cdot 1 - 0 = vx$; in turn, this implies that Player 2's utility at $\big(d|h(a)\big)$ would equal $u_2^C(z(d), \rho_i, \beta_2) = m_2\big(z(d)\big) - k_2\Big[1 + \Big(\mathrm{E}_{\rho_i,\beta_2}[m_1|h(a)] - m_1\big(z(d)\big)\Big)\Big] = 2vx - k_2(1+vx)$, whereas Player 2's utility (=payoff) at $\big(c|h(a)\big)$ is simply given by $u_2^C(z(c), \rho_i, \beta_2) \equiv m_2\big(z(c)\big) = vx$. Finally, Player 2's conformist preferences against d can be expressed as $u_2^C(z(d), \rho_i, \beta_2) \leq u_2^C(z(c), \rho_i, \beta_2) \Longrightarrow 2vx - k_2(1+vx) \leq vx \Longrightarrow k_2 \geq \frac{vx}{1+vx}$. It follows that the only norm-conjecture induced by r^F that yields an SSE is given by the matrix:

$$\rho_i = \begin{bmatrix} \rho_i(a|h^0) & \rho_i(b|h^0) \\ \rho_i\big(c|h(a)\big) & \rho_i\big(d|h(a)\big) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix},$$

in which case (a, c) is an SSE for $k_2 \geq \frac{vx}{1+vx}$; instead, if $k_2 < \frac{vx}{1+vx}$, behavioral rule r^F is not a social norm and is not followed by population N (by remark 1 above).

Proof of Corollary 2. Given an end-node $\hat{z} \in Z$, a Rawlsian social welfare function is defined as $W\big(m_1(\hat{z}), m_2(\hat{z})\big) = \min_{i \in N}\{m_1(\hat{z}), m_2(\hat{z})\}$; such a function has to be evaluated at each of the three end-nodes of the game tree, *i.e.*: $W\Big(m_1\big(z(b)\big), m_2\big(z(b)\big)\Big) = \min_{i \in N}\{x, x\} = x$; $W\Big(m_1\big(z(d)\big), m_2\big(z(d)\big)\Big) = \min_{i \in N}\{0, 2vx\} = 0$; $W\Big(m_1\big(z(c)\big), m_2\big(z(c)\big)\Big) = \min_{i \in N}\{vx, vx\} = vx$. It follows that here the set of maximizers of W is singleton: so, at the initial node the behavioral rule

r^W dictates only the strategy profile $r^W(h^0) = \{(a, c)\}$; then, the norm-conjecture induced by r^W, for $\forall\, i \in N$, can be represented by the following matrix:

$$\rho_i = \begin{bmatrix} \rho_i(a|h^0) & \rho_i(b|h^0) \\ \rho_i\big(c|h(a)\big) & \rho_i\big(d|h(a)\big) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

The rest of the proof is trivial.

Proof of Proposition 5. Note that $r^M(h^0)$ contains the following dictated strategy profiles: $r^M(h^0) = \{(a, c), (a, d)\}$. Hence, the norm-conjecture induced by r^M, for $\forall\, i \in N$, can be represented by the following matrix:

$$\rho_i = \begin{bmatrix} \rho_i(a|h^0) & \rho_i(b|h^0) \\ \rho_i\big(c|h(a)\big) & \rho_i\big(d|h(a)\big) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \hat{\rho} & 1-\hat{\rho} \end{bmatrix},$$

where $\hat{\rho} \in \{0,1\}$; that is, if $\hat{\rho} = 0$ then strategy profile (a, d) is implemented, whereas if $\hat{\rho} = 1$ then (a, c) is implemented. Thus Player 1's expectation of Player 2's (expected) material payoff at h^0 equals $\mathrm{E}_{\rho_i,\alpha_1,\beta_1}[m_2|h^0] = vx\hat{\rho} + 2vx(1-\hat{\rho})$. Further, if Player 1 chose action b, then her expectation of Player 2's potential disappointment at $(b|h^0)$ would equal $\mathrm{E}_{\rho_i,\alpha_1,\beta_1}[m_2|h^0] - m_2\big(z(b)\big) = vx\hat{\rho} + 2vx(1-\hat{\rho}) - x = 2vx - x - vx\hat{\rho}$; in turn, this implies that Player 1's utility at $(b|h^0)$ would equal $u_1^C(z(b), \rho_i, \alpha_1, \beta_1) = m_1\big(z(b)\big) - k_1\Big[1 + \Big(\mathrm{E}_{\rho_i,\alpha_1,\beta_1}[m_2|h^0] - m_2\big(z(b)\big)\Big)\Big] = x - k_1(1 + 2vx - x - vx\hat{\rho})$. On the other hand, Player 1's expected utility (=expected payoff) at $h(a)$ is given by $\mathrm{E}_{\rho_i,\alpha_1}[m_1|h(a)] = vx\hat{\rho} + 0 \cdot (1-\hat{\rho}) = vx\hat{\rho}$: it follows that Player 1's conformist preferences against b can be expressed as $u_1^C(z(b), \rho_i, \alpha_1, \beta_1) \leq \mathrm{E}_{\rho_i,\alpha_1}[m_1|h(a)] \Longrightarrow x - k_1(1 + 2vx - x - vx\hat{\rho}) \leq vx\hat{\rho}$, which implies $k_1 \geq \frac{x - vx\hat{\rho}}{1 + 2vx - x - vx\hat{\rho}} \Longrightarrow \begin{cases} k_1 \geq \frac{x-vx}{1+vx-x} & if\ \hat{\rho} = 1 \\ k_1 \geq \frac{x}{1+2vx-x} & if\ \hat{\rho} = 0 \end{cases}$, where the first case is always satisfied because $v > 1$. Given that, if Player 1 does choose action a, then two scenarios are possible. ***(a)*** If Player 2 believes that Player 1 chose a because she believed that $\hat{\rho} = 1$, then Player 2's expectation of Player 1's potential disappointment at $\big(d|h(a)\big)$ would equal $\mathrm{E}_{\rho_i,\beta_2}[m_1|h(a)] - m_1\big(z(d)\big) = vx \cdot 1 - 0 = vx$; in turn, this implies that Player 2's utility at $\big(d|h(a)\big)$ would equal $u_2^C(z(d), \rho_i, \beta_2) = m_2\big(z(d)\big) - k_2\Big[1 + \Big(\mathrm{E}_{\rho_i,\beta_2}[m_1|h(a)] - m_1\big(z(d)\big)\Big)\Big] = 2vx - k_2(1 + vx)$, whereas Player 2's utility (=payoff) at $\big(c|h(a)\big)$ is simply given by $u_2^C(z(c), \rho_i, \beta_2) \equiv m_2\big(z(c)\big) = vx$. So, in this case Player 2's conformist preferences against d can be expressed as $u_2^C(z(d), \rho_i, \beta_2) \leq u_2^C(z(c), \rho_i, \beta_2) \Longrightarrow 2vx - k_2(1 + vx) \leq vx \Longrightarrow k_2 \geq \frac{vx}{1+vx}$: it follows that for $\hat{\rho} = 1$, (a, c) is an SSE for $k_2 \geq \frac{vx}{1+vx}$. ***(b)*** The second scenario entails the following reasoning. If Player 2 believes – *in a self-serving way* – that Player 1 chose a (even though she believed that $\hat{\rho} = 0$) because k_1 is large enough, then Player 2's expectation of Player 1's potential disappointment at $\big(d|h(a)\big)$ would be null as $\hat{\rho} = 0$ in the expression $\mathrm{E}_{\rho_i,\beta_2}[m_1|h(a)] - m_1\big(z(d)\big) = vx\hat{\rho} - 0$; in turn, this implies that Player 2's utility

at $\big(d|h(a)\big)$ would simply correspond to her material payoff because there would be no deviation from ρ_i, hence no psychological loss (*i.e.*: $u_2^C(z(d),\rho_i,\beta_2) \equiv m_2\big(z(d)\big) = 2vx$). So, in this case Player 2's conformist preferences can be expressed as $u_2^C(z(c),\rho_i,\beta_2) \leq u_2^C(z(d),\rho_i,\beta_2) \Longrightarrow vx < 2vx$, which is always satisfied: it follows that for $\hat{\rho} = 0$, (a,d) is an SSE for $k_1 \geq \frac{x}{1+2vx-x}$.

Proof of Proposition 6. Note that $r^M(h^0)$ contains a unique strategy profile (x_i^*, x_{-i}^*) where every player's contribution equals $x_i^* = M$. Hence, the norm-conjecture induced by r^M, for $\forall\, i \in N$, is such that $\rho_i\big(x_j = M\big) = 1$ for $\forall\, j \in N$. Thus Player i's expectation of Player j's (expected) material payoff at h^0 equals $\mathrm{E}_{\rho_i,\alpha_i,\beta_i}\big[m_j|h^0\big] = M - x_g^* + \frac{v\sum_g^n x_g^*}{n} = vM$. Further, if Player i chose action $x_i = 0$, then her expectation of Player j's potential disappointment would equal $\mathrm{E}_{\rho_i,\alpha_i,\beta_i}\big[m_j|h^0\big] - m_j\big(z(x_i = 0, x_{-i}^*)\big) = vM - \frac{vM(n-1)}{n} = \frac{vM}{n}$; in turn, this implies that Player i's utility at $(x_i = 0|x_{-i}^*)$ would equal $u_i^C(z(0,x_{-i}^*),\rho_i,\alpha_i,\beta_i) = m_i\big(z(0,x_{-i}^*)\big) - k_i\Big[1 + (n-1)\Big(\mathrm{E}_{\rho_i,\alpha_i,\beta_i}\big[m_j|h^0\big] - m_j\big(z(0,x_{-i}^*)\big)\Big)\Big] = M + \frac{vM(n-1)}{n} - k_i\Big[1 + (n-1)\Big(\frac{vM}{n}\Big)\Big] = M + \frac{vM(n-1)}{n} - k_i\Big(\frac{n+vM(n-1)}{n}\Big)$. On the other hand, Player i's utility (=payoff) at $(x_i^*|x_{-i}^*)$ is given by $u_i^C(z(x_i^*,x_{-i}^*),\rho_i,\alpha_i,\beta_i) = m_i\big(z(x_i^*,x_{-i}^*)\big) = vM$: it follows that Player i's conformist preferences against $x_i = 0$ can be expressed as $u_i^C(z(0,x_{-i}^*),\rho_i,\alpha_i,\beta_i) \leq u_i^C(z(x_i^*,x_{-i}^*),\rho_i,\alpha_i,\beta_i) \Longrightarrow$
$M + \frac{vM(n-1)}{n} - k_i\Big(\frac{n+vM(n-1)}{n}\Big) \leq vM$, which implies that $k_i \geq \frac{M(n-v)}{n+vM(n-1)}$.