Social Rationalizability*

P. Jean-Jacques Herings[1], Ana Mauleon[2] and Vincent J. Vannetelbosch[3]

[1] Department of Economics, University of Maastricht, The Netherlands
(e-mail: P.Herings@algec.unimaas.nl)
[2] Department of Economic Analysis, Universidad del Pais Vasco, Spain
(e-mail: jepmaeca@bs.ehu.es)
[3] THEMA, University of Cergy Pontoise, France
(e-mail: Vincent.Vannetelbosch@eco.u-cergy.fr)

Date: April 2000

Summary. Social environments constitute a framework in which it is possible to study how groups of agents interact in a society. The framework is general enough to analyse both non-cooperative and cooperative games. We identify a number of shortcomings of existing solution concepts that are used for social environments and propose a new concept called social rationalizability. The concept aims to identify the consequences of common knowledge of rationality and farsightedness within the framework of social environments. The set of socially rationalizable outcomes is shown to be non-empty for all social environments and it can be computed by an iterative reduction procedure. We introduce a definition of coalitional rationality for social environments and show that it is satisfied by social rationalizability.

Keywords and Phrases: Social environments, rationalizability, coalitional rationality.

JEL Classification Numbers: C72, C78

*Research for this article was mainly carried out while Ana Mauleon and Vincent Vannetelbosch were visiting CentER at Tilburg University. Financial support from the research projects TMR Network FMRX CT 960055 "Cooperation and Information", PI-1998-48 (government of the Basque Country) and G17-99 (UPV-EHU) is gratefully acknowledged.

1 Introduction

Many social, economic and political activities are conducted by groups or coalitions of individuals. For example, consumption takes place within households or families; production is carried out by firms which are large coalitions of owners of different factors of production; workers are organized in trade unions or professional associations; public goods are produced within a complex coalition structure of federal, state, and local jurisdictions; political life is conducted through political parties and interest groups; and individuals belong to networks of formal and informal social clubs.

The framework of social environments as introduced in Chwe [4] (see also Rosenthal [9]) specifies what each coalition can do if and when it forms. It is general enough to integrate the representation of a cooperative game, an extensive-form game with perfect information, and a normal-form game played in such a fashion that there are coalitional moves and countermoves. An example is the coalitional contingent threat situation due to Greenberg [5]. For social environments where coalitions can form through binding or non-binding agreements and actions are public, Chwe [4] and Xue [14] have proposed the solution concepts of the largest consistent set and the optimistic or conservative stable standards of behavior, respectively. The solution concepts predict which coalitions structures are possibly stable and could emerge.[1]

Both approaches have a number of nice features. Firstly, they do not rely on a very detailed description of the coalition formation process as noncooperative sequential games do, see e.g. Bloch [3].[2] No commitment assumption is imposed. Secondly, it incorporates the farsightedness of the coalitions. A coalition considers the possibility that, once it acts, another coalition might react, a third coalition might in turn react, and so on without limit. The main difference between Chwe [4] and Xue [14] is that Xue's approach strengthens the farsightedness notion. A farsighted individual considers only the final outcomes that might result when making choices. But, an individual with perfect foresight considers also how final outcomes can be reached. That is, possible deviations along the way to the final outcomes should be considered.[3]

[1] For a very specific social environment, namely the coalitional contingent threat situation, Mariotti [7] has defined an equilibrium concept: the *coalitional equilibrium*. Central to his concept is the notion of coalitional strategies and the similarity with subgame perfection (except that coalitions are formally treated as players).

[2] Sequential coalition formation games are quite sensitive to the exact coalition formation process and rely on the commitment assumption. Once some individuals have agreed to form a coalition they are committed to remain in that coalition. They can neither leave the coalition nor propose to change it later on.

[3] In Chwe [4] the specification of how individuals view and use their alternatives is formalized by the indirect dominance relation which captures some farsightedness of the individuals. In Xue [14] it is formalized by means of the theory of social situations developed by Greenberg [5]. A social situation allows to capture perfect foresight (which strengthens farsightedness) by extending the von Neumann and Morgenstern [13] notion of stability to accommodate different behavior on the part of the individuals in terms of their Knightian (pessimism or optimism) attitude towards uncertainty.

Both approaches suffer from a number of drawbacks as well, some of them pointed out by the authors themselves. For instance, as indicated in Chwe [4], the largest consistent set may fail to satisfy the requirement of individual rationality. An individual that is given the choice between two moves, where one yields with certainty a higher payoff than the other, might choose the move leading to the lower payoff according to the largest consistent set. This is perhaps somewhat less disturbing than it seems at first sight, since the largest consistent set aims to be a weak concept, a concept that rules out with confidence. It is therefore more surprising, as we show in this paper, that in certain social environments the largest consistent set may rule out too much. One drawback of both the optimistic and the conservative stable standards of behavior of Xue [14], is that both solution sets may be empty. This is worrisome as the idea of farsightedness suggests that since coalitions do take into account the far reaching consequences of their moves, they should be able to settle on some stable outcomes at least. We also present a number of examples where the stable standards of behavior lead to undesirable outcomes, for instance that both OSSB and even CSSB may rule out too little, or even worse, too much.

We aim for a solution concept that identifies the consequences of common knowledge of rationality and farsightedness within the framework of social environments, and that remedies the problems mentioned above. To achieve this goal, we propose to extend the rationalizability approach of Bernheim [2] and Pearce [8] to the framework of social environments. We use a cautious version of rationalizability that is also analyzed in Herings and Vannetelbosch [6]. Since social environments deal with the behavior of coalitions, whereas rationalizability is about the implications of rationality of individuals, we have to convert coalitional behavior into individual behavior. This is achieved by recognizing that individual participation in a coalition is basically characterized by two possibilities. An individual may either agree to a coalitional move, or object to it and block it. Unlike in non-cooperative game theory, in a social environment several coalitions may and could be willing to move at the same time. Conflicts of interest may arise, which can take the form of one coalition trying to preempt the move of another coalition, but also of coordination problems in and between coalitions. Individuals should therefore also have beliefs on how such conflicts of interest are solved.

The equilibrium approach assumes that individuals have common expectations about their behaviors. That is, each individual holds a correct conjecture about the behavior of every other individual. But once we admit the possibility that an individual may have several behaviors that she could reasonably take, conjectures and behaviors actually played may be mismatched. This is what distinguishes the rationalizability approach from the equilibrium one. Indeed, in the rationalizability approach, the conjectures are not assumed to be correct, but are only constrained by considerations of rationality. Each individual believes that the behavior taken by every other individual is a best response to some conjecture on every other individual's

behavior, and, further, each individual assumes that every other individual reasons in this way and hence thinks that every other individual believes that every other individual's behavior is a best response to some conjecture, and so on. In other words, the individual rationality of the individuals is common knowledge.

We introduce two alternative definitions of the social rationalizability concept which we show to be equivalent definitions. The first one is strongly influenced by Battigalli's [1] extensive-form rationalizability. It is based on two assumptions: (1) the individuals are rational and endowed with a hierarchy of hypotheses, and (2) this is common knowledge at the original status-quo. Central to our new concept are the notions of *individual behavior* and of *implementability prior-belief*. An individual behavior describes, for each history, the coalitional moves the individual agrees to join and those she decides to block. Beliefs about which agreement is implemented among the set of agreements are derived from an implementability prior-belief over the entire set of feasible moves. Our second definition is motivated by Pearce's [8] original extensive-form rationalizability and is based on a reduction procedure.[4] We show the equivalence of our two definitions of social rationalizability.

Our main results are the following. The set of socially rationalizable outcomes is non-empty for the entire class of social environments. When we apply social rationalizability to the prisoner's dilemma, it follows that cooperation is sustained. Social environments deal with coalitional moves. It is therefore important that social rationalizability not only guarantees individual rationality, but also coalitional rationality. Among a set of Pareto ranked alternatives a coalition should be able to coordinate on the Pareto optimal one. Social rationalizability is shown to satisfy coalitional rationality.

The paper is organized as follows. In Section 2 we introduce some notations and primitives. We present the solution concepts of Chwe [4] and Xue [14], and we give the motivation for introducing a new concept. In Section 3 we propose two alternative definitions of social rationalizability and we show the equivalence of both of them. The examples are reconsidered and solved by our concept. In Section 4 we study the property of coalitional rationality and show it is satisfied by social rationalizability. Finally, Section 5 concludes.

[4] Other papers related to extensive-form rationalizability (EFR) are among others Bernheim [2], who introduced subgame-perfect rationalizability, Shimoji and Watson [10], who studied the equivalence between conditional dominance and EFR, and Vannetelbosch [11],[12], who defined rationalizability for multi-stage bargaining games.

2 Social Environments

2.1 Notations and Primitives

As in Chwe [4] and Xue [14], we define by $\Gamma = \langle I, Z, (u_i)_{i\in I}, \{\rightarrow_S\}_{S\subseteq I, S\neq\emptyset}\rangle$ a social environment, where $I = \{1, 2, ..., \#I\}$ is the set of individuals, Z is the finite set of outcomes, $\{\rightarrow_S\}_{S\subseteq I,S\neq\emptyset}$ are effectiveness relations defined on Z, and for every individual $i \in I$, $u_i : Z \rightarrow \mathbb{R}$ is her utility function. We denote by $\#I$ the cardinality of I. The relation $\rightarrow_S$ represents what coalition S can do: $x_0 \rightarrow_S x_1$ means that if x_0 is the status-quo, coalition S can make x_1 the new status-quo. It does not mean that coalition S can enforce x_1 no matter what anyone else does; after S moves to x_1 from x_0, another coalition $S^{'}$ might move to x_2, where $x_1 \rightarrow_{S'} x_2$. A priori no restrictions are imposed on the effectiveness relations $\{\rightarrow_S\}_{S\subseteq I,S\neq\emptyset}$. For example, the effectiveness relation can be empty, $x_0 \rightarrow_S x_0$ might be possible, and $x_0 \rightarrow_S x_1$ does not imply $x_1 \rightarrow_S x_0$. All actions or moves are public and the individuals care only about the end outcome, not how it is reached. Conventional game theoretic situations can be modeled as a social environment.

- For noncooperative games in normal-form there are at least two possibilities for representation, depending on whether coalitions may form or not. Let Z_i denote the nonempty set of pure strategies of individual i. In the first case, $Z = \prod_{i\in I} Z_i$ and $x \rightarrow_S y$ if $x_{I\setminus S} = y_{I\setminus S}$. In the second case, $Z = \prod_{i\in I} Z_i$ and $x \rightarrow_S y$ if $S = \{i\}$ and $x_{I\setminus\{i\}} = y_{I\setminus\{i\}}$.
- For a cooperative TU-game (v, I), where the payoff of the grand coalition has been normalized to $v(I) = 1$ and the payoffs of one-player coalitions to $v(\{i\}) = 0$, we set $Z = \{x \in \mathbb{R}^{\#I} \mid \sum_{i=1}^{\#I} x_i = 1 \text{ and } x_i \geq 0,\ \forall i \in I\}$ and $x \rightarrow_S y$ if $\sum_{i\in S} y_i \leq v(S)$. By restricting attention to integer payoffs, it is easy to incorporate the existence of a smallest money unit and to get a finite set of outcomes. For a cooperative NTU-game (v, I), we put $Z = \{x \in \mathbb{R}^{\#I} \mid x \in v(I) \text{ and } x_i \text{ individual rational}\}$ and $x \rightarrow_S y$ if $y_S \in v(S)$.

For social environments where coalitions can form through binding or non-binding agreements and actions are public, Chwe [4] and Xue [14] have proposed interesting concepts, the largest consistent set and the optimistic or conservative stable standards of behavior, respectively, to predict which coalition structures are possibly stable or could emerge.

2.2 The Largest Consistent Set

Based on the indirect dominance relation, Chwe [4] defined the largest consistent set (LCS). The indirect dominance relation captures the fact that farsighted coalitions consider the end outcome that their move(s) eventually may lead to. Moreover, a coalition may deviate from a status quo only if each of its members can be made strictly better off. So, an outcome y indirectly dominates x if y can replace x in a sequence of moves, such that at each move all

deviators are better off at the end outcome y compared to the status-quo they face. Formally, indirect dominance is defined as follows.

An outcome x is *indirectly dominated* by y, or $x \ll y$, if there exists a sequence $x_0, x_1, ..., x_m$, where $x_0 = x$ and $x_m = y$, and a sequence $S_0, S_1, ..., S_{m-1}$ such that $x_j \rightarrow_{S_j} x_{j+1}$ and $u_i(x_j) < u_i(y)\ \forall\ i \in S_j$, for $j = 0, 1, .., m-1$. Direct strict dominance is obtained by setting $m = 1$. An outcome x is *directly dominated* by y, or $x < y$, if there exists a coalition S such that $x \rightarrow_S y$ and $u_i(x) < u_i(y)\ \forall\ i \in S$. Obviously, if $x < y$, then $x \ll y$. The largest consistent set, $LCS(\Gamma)$, is defined as follows.

Definition 1 (Chwe, 1994) *A set $Y \subseteq Z$ is consistent if $x \in Y$ if and only if $\forall\ y, S$ such that $x \rightarrow_S y$, $\exists\ z \in Y$, where $y = z$ or $y \ll z$, such that we do not have $u_i(x) < u_i(z)$ for all $i \in S$. The largest consistent set $LCS(\Gamma)$ is the consistent set such that if $Y \subseteq Z$ is consistent then $Y \subseteq LCS(\Gamma)$.*

By considering indirect dominance, the largest consistent set captures the notion of farsightedness. An outcome is stable, that is an outcome is in the largest consistent set, if and only if deviations from it do not occur because the deviation itself or potential further deviations are not unanimously preferred to the original outcome by the coalition considering the deviation. Although there can be many consistent sets, Chwe [4] has shown that there uniquely exists a largest consistent set, $LCS(\Gamma)$, and that the largest consistent set is non-empty. One simple way to find $LCS(\Gamma)$ is to apply the following iterative procedure. Let $Y^0 \equiv Z$. Then, Y^k $(k = 1, 2, ...)$ is inductively obtained as follows: $x \in Z$ belongs to Y^k if and only if $\forall\ y, S$ such that $x \rightarrow_S y$, $\exists\ z \in Y^{k-1}$, where $y = z$ or $y \ll z$, such that we do not have $u_i(x) < u_i(z)$ for all $i \in S$. Then, $LCS(\Gamma)$ is $\bigcap_{k \geq 1} Y^k$.

2.3 Stable Standards of Behavior

We give the definitions of Optimistic Stable Standard of Behavior (OSSB) and Conservative Stable Standard of Behavior (CSSB) due to Xue [14]. Some notations and definitions have to be introduced. A path is a sequence $(x_0, x_1, ..., x_m)$ where for all $j = 0, 1, ..., m-1$, there exists a coalition $S_j \subseteq I$ such that $x_j \rightarrow_{S_j} x_{j+1}$ and $x_j, x_{j+1} \in Z$. Let Π be the set of paths in Z, and Π_x the set of paths in Z originating from x. Xue [14] defined a standard of behavior as a function $\sigma : Z \rightarrow 2^{\Pi}$ such that $\sigma(x) \subseteq \Pi_x$ for all $x \in Z$. A standard of behavior is said to be *internally stable* if $\forall x \in Z$, $\forall \alpha \in \sigma(x)$, $\nexists y \in \alpha$, $\nexists S \subseteq I$, $\nexists z \in Z$ such that $y \rightarrow_S z$ and S "prefers" $\sigma(z)$ to α. A standard of behavior is said to be *externally stable* if $\forall x \in Z$, $\forall \alpha \in \Pi_x \setminus \sigma(x)$, $\exists y \in \alpha$, $\exists S \subseteq I$, $\exists z \in Z$ such that $y \rightarrow_S z$ and S "prefers" $\sigma(z)$ to α. A standard of behavior is stable if it is both internally and externally stable.

As in Greenberg [5], Xue [14] distinguished an optimistic and a conservative approach to define "prefers." In the optimistic approach a coalition S prefers $\sigma(z)$ to α if $\exists \beta \in \sigma(z)$, $u_i(\alpha) < u_i(\beta)\ \forall i \in S$. In the conservative approach a coalition S prefers $\sigma(z)$ to α if $\forall \beta \in \sigma(z)$, $u_i(\alpha) < u_i(\beta)\ \forall i \in S$. An OSSB is a stable standard of behavior, where "prefers" is defined by the optimistic approach. A CSSB is a stable standard of behavior, where "prefers" is defined by the conservative approach. Formally,

Definition 2 (Xue, 1998) *Let σ be a standard of behavior. Then,*

(i) *σ is an OSSB if $\forall x \in Z$, $\alpha \in \Pi_x \backslash \sigma(x) \Longleftrightarrow \exists S \subseteq I$, $y \in \alpha$, and $z \in Z$ such that $y \rightarrow_S z$ and $\exists \beta \in \sigma(z) : u_i(\alpha) < u_i(\beta)\ \forall i \in S$.*

(ii) *σ is a CSSB if $\forall x \in Z$, $\alpha \in \Pi_x \backslash \sigma(x) \Longleftrightarrow \exists S \subseteq I$, $y \in \alpha$, and $z \in Z$ such that $y \rightarrow_S z$ and $\forall \beta \in \sigma(z) \neq \emptyset : u_i(\alpha) < u_i(\beta)\ \forall i \in S$.*

2.4 Motivation and Examples

As has already been mentioned by Chwe [4] himself, the LCS is blurring or avoiding important issues, and hence, suffers substantial drawbacks. One drawback is that the LCS does not incorporate any idea of best response. Thereby, it is not very surprising that the LCS does not always rule out all unreasonable moves. Figure 1 shows a social environment with one individual

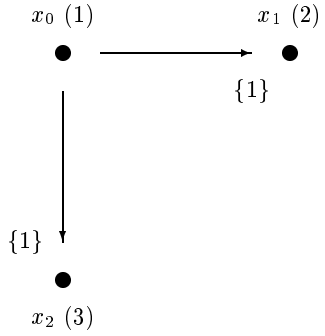

Figure 1: Individual rationality.

that is currently at the status quo x_0 where she gets 1 unit of utility. She has the possibility to move to outcome x_1 and obtain 2 units of utility, or to go to outcome x_2 and receive 3 units of utility. In the social environment of Figure 1, $LCS(\Gamma) = \{x_1, x_2\}$. This is unreasonable as a simple optimization dictates individual 1 to move to x_2, in order to get a utility equal to 3 instead of 2. So, the LCS does not satisfy individual rationality.[5]

[5] Two other problems have also been mentioned by Chwe [4]. First, the LCS does not incorporate the decision of subcoalitions to veto coalitional moves. Second, a coalition considers what further moves other coalitions will

It is more surprising that we have found social environments where LCS rules out too much. This problem is more serious as LCS is developed to be a weak concept that rules out with confidence. In the social environment of Figure 2, there are three individuals that have the

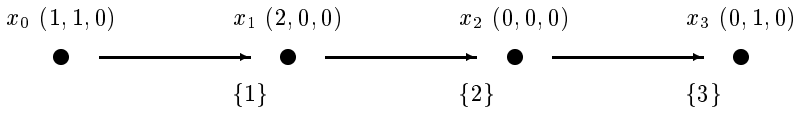

Figure 2: LCS may rule out too much

opportunity to move in a sequential manner. The status quo is x_0. The utility tuples achievable at the four outcomes are indicated in parentheses, with the utility of individual i in position i. The direct dominance relation is given by $x_0 < x_1$ and the indirect one by $x_0 \ll x_1$. It follows that $LCS(\Gamma) = \{x_1, x_2, x_3\}$, so outcome x_0 is ruled out. However, individual 1 only wants to move from outcome x_0 to outcome x_1 if she is sure that individual 2 will not move from x_1 to x_2. Individual 2 does have incentives to move from x_1 to x_2 as the move to x_2 enables individual 3 to move to x_3. It is only when individual 2 is sure that 3 does not move that he is indifferent between moving and not moving. Even under such extreme beliefs individual 2 would not loose from moving to x_2. It is therefore certainly reasonable for individual 1 not to move from outcome x_0 to x_1. A concept that aims to rule out with confidence should not rule out outcome x_0.

The OSSB seems to perform better than LCS for the social environment of Figure 2. It holds that the unique OSSB is defined by $\sigma(x_0) = \{(x_0)\}$, $\sigma(x_1) = \{(x_1, x_2, x_3)\}$, $\sigma(x_2) = \{(x_2), (x_2, x_3)\}$ and $\sigma(x_3) = \{(x_3)\}$. The uniqueness of OSSB follows from Claim 3.11 in Xue [14]. So individual 1 will not make the move from x_0 to x_1, because she fears the move of individual 2 from x_1 to x_2. Less convincing is that $(x_1, x_2) \notin \sigma(x_1)$. Individual 2 hopes for the best, so he is convinced that individual 3 moves from x_2 to x_3. This is not consistent with the fact that $\sigma(x_2)$ contains both (x_2) and (x_2, x_3).

The CSSB is a truly weak concept. It doesn't rule out anything in the social environment of Figure 2. But even though a CSSB is typically a very weak concept, it may also rule out too much. In the social environment of Figure 3 there is a unique CSSB, given by $\sigma(x_0) = \emptyset$, $\sigma(x_1) = \{(x_1)\}$ and $\sigma(x_2) = \{(x_2)\}$. The uniqueness of CSSB follows from Claim 3.11 in Xue

make once it moves, but does not consider what other coalitions will do if it does not move. Hence, the LCS does not allow for the possibility of coalitions moving to preempt the moves of other coalitions. Social rationalizability (as well as Xue's [14] concepts) overcomes these problems.

[14]. Although a unique CSSB exists, it is empty-valued for some status quos. A standard of

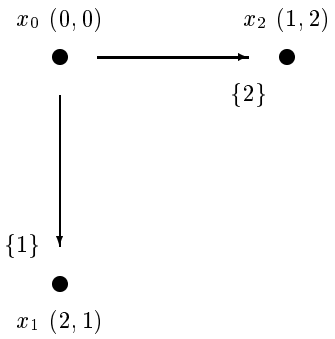

Figure 3: CSSB and OSSB may rule out too much.

behavior that prescribes $\sigma(x_0) = \{(x_0, x_1), (x_0, x_2)\}$, violates internal stability when one also assigns the obvious $\sigma(x_1) = \{(x_1)\}$ and $\sigma(x_2) = \{(x_2)\}$, since $(x_0, x_2) \in \sigma(x_0)$, $x_0 \rightarrow_{\{1\}} x_1$, and $\sigma(x_1)$ is preferred to (x_0, x_2).

The unique OSSB coincides with the CSSB for the social environment of Figure 3, and may therefore also be empty-valued and rule out too much, a feature that is less surprising for OSSB. The example becomes even more striking when we add a move $x_0 \rightarrow_{\{1,2\}} x_3$ with payoffs -1 for both individuals. Then the unique CSSB and the unique OSSB are given by $\sigma(x_0) = \emptyset$, $\sigma(x_1) = \{(x_1)\}$, $\sigma(x_2) = \{(x_2)\}$ and $\sigma(x_3) = \{(x_3)\}$. The solution concepts CSSB and OSSB do not distinguish the moves to x_1 and x_2 on the one hand, and the move to x_3 on the other. Another possibility is to add a move $x_3 \rightarrow_{\{1\}} x_0$ and to put the utility of both individuals to -1 at x_3. The standard of behavior $\sigma(x_3) = \{(x_3)\}$, $\sigma(x_0) = \emptyset$, $\sigma(x_1) = \{(x_1)\}$, and $\sigma(x_2) = \{(x_2)\}$ is both an OSSB and a CSSB. The worst outcome is stable.

CSSB and OSSB may also rule out too little. In the social environment of Figure 4, the only sensible standard of behavior is $\sigma(x_0) = \{(x_0)\}$. Nevertheless, the standard of behavior

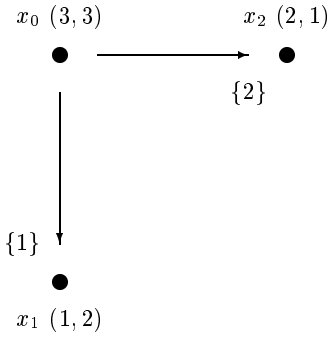

Figure 4: OSSB and CSSB may rule out too little.

$\sigma(x_0) = \{(x_0), (x_0, x_1), (x_0, x_2)\}$, $\sigma(x_1) = \{(x_1)\}$ and $\sigma(x_2) = \{(x_2)\}$ is both the unique CSSB and the unique OSSB. It may look like this phenomenon is caused by the absence of the no-move. But even if we add moves $x_0 \to_{\{1\}} x_0$, $x_0 \to_{\{2\}} x_0$, $x_0 \to_{\{1,2\}} x_0$, then the standard of behavior defined by $\sigma(x_0) = \{(x_0), (x_0, x_1), (x_0, x_2), (x_0, x_0), (x_0, x_0, x_1), (x_0, x_0, x_2), (x_0, x_0, x_0), \cdots\}$, $\sigma(x_1) = \{(x_1)\}$ and $\sigma(x_2) = \{(x_2)\}$ is a CSSB. OSSB seems to do better now, as the unique OSSB is given by $\sigma(x_0) = \{(x_0), (x_0, x_0), (x_0, x_0, x_0), \cdots\}$, $\sigma(x_1) = \{(x_1)\}$, and $\sigma(x_2) = \{(x_2)\}$.

In order to remedy these drawbacks, we propose a notion of rationalizability for social environments, which identifies the coalitions that are likely to form and the outcomes that might occur when (1) the individuals are rational and endowed with a hierarchy of hypotheses, and (2) this is common knowledge at the original status-quo.

3 Rationalizable Social Behaviors

3.1 Individual and Social Behaviors

In what follows, we denote the move of coalition S from x to y, $x \to_S y$, by (xy, S). The no-move at status-quo x is denoted by $(xx, \emptyset)$. One has to distinguish between $(xx, \emptyset)$ and $(xx, \{i\})$. Indeed, $(xx, \{i\})$ means that individual i can move from x to x. The set of all possible moves and no-move is given by $M = \{(xy, S) | x, y \in Z, x \to_S y\} \cup \{(xx, \emptyset) | x \in Z\}$. An original status-quo is given, and it is denoted x_0. We consider histories starting at x_0. We denote by $h = (x_0, m_1, m_2, ..., m_{k-1})$ a history of length k, where $x_0 \in Z$ is the status-quo, $m_j = (m_j^- m_j^+, m_j^c) \in M$, $m_1^- = x_0$, $m_j^+ = m_{j+1}^-$ and m_j^c denotes the coalition of individuals that moves from m_j^- to m_j^+, $j = 1, ..., k-1$. The length of a history h is denoted $l(h)$ with $l(h) = 1$ for $h = (x_0)$. To make the length of a history h explicit, we sometimes denote it by h^k, where k is the length of the history. Let $h^- = x_0$ be the *original* status-quo of h and $h^+ = m_{l(h)-1}^+$ be the end outcome of h. Given h^k and $j \leq k$ $(j, k \in \mathbb{N})$, we call h^j a sub-history of h^k if h^j consists of the first j elements of h^k, and we write $h^j \leq h^k$. A history is different from a path as used in the theory of stable standards of behavior. A path only gives a sequence of outcomes, whereas for a history it also matters which coalition made the move from one outcome to another.

The set of feasible moves after history h is denoted by $M(h) = \{m \in M | h^+ = m^-\} \setminus \{(h^+ h^+, \emptyset)\}$ for all h. It does not include the no-move. Let $M_i(h) = \{(xy, S) \in M(h) | i \in S\}$ be the set of feasible moves after history h involving individual i. The set of individuals that has a move after history h is denoted $I(h) = \{i \in I | M_i(h) \neq \emptyset\}$.

We denote by H the set of all histories with finite length and by $H(J)$ the set of histories with at most J moves. That is, $H(J) = \{h \in H | l(h) \leq J + 1\}$. Temporarily we fix J and consider only histories in $H(J)$. Let $H_i(J) = \{h \in H(J) | M_i(h) \neq \emptyset\}$. It is the set of histories that contain at most J moves and after which individual i is involved in a move. Individual i's

opponents are denoted by $-i$. As general notation, we denote by $\Delta(X)$ the set of all probability measures on X. For finite X, we denote by $\Delta^0(X)$ the set of all probability measures giving positive probability to each member of X.

A social behavior selects after any history a unique move or a no-move. We denote it by $b = (b(h))_{h \in H(J)}$ where $b(h) \in M(h) \cup \{(h^+h^+, \emptyset)\}$. Let B be the set of all social behaviors. Our aim is to find those social behaviors that are rationalizable. From the rationalizable social behaviors, we derive the set of outcomes that are stable. We aim for a concept that is weak, so rules out with confidence. To do this, we examine individual behaviors first.

We model an individual behavior as, for each history, the set of coalitional moves the individual agrees to join and those she decides to block. Observe that the framework of social environments does not exclude that an individual might agree to join more than one coalitional move (if possible). Formally, a behavior of individual i is $b_i = (b_i(\cdot \mid h))_{h \in H_i(J)}$ where $b_i(\cdot \mid h) : M_i(h) \to \{0, 1\}$. If $b_i((xy, S) \mid h) = 1$ then $i \in S$ agrees to join in the potential move of coalition S from x to y. If $b_i((xy, S) \mid h) = 0$ then $i \in S$ blocks the move of coalition S from x to y. The set of all possible behaviors of individual i is denoted by B_i.

It may happen that the individuals agree on more than one move. We denote by $\mathcal{M}(h) = \left\{ \overline{M} \middle| \, \emptyset \neq \overline{M} \subset M(h) \right\}$ the collection of sets of feasible moves after h. For every history $h \in H(J)$, the agreement function is a mapping $f(\cdot \mid h) : \prod_{i \in N} B_i \to \mathcal{M}(h) \cup \{(h^+h^+, \emptyset)\}$ which associates to the profiles of individual behaviors the set of moves after history h on which there is agreement, so

(i) $f((b_i)_{i \in N} \mid h) = \overline{M} \in \mathcal{M}(h)$ if $\forall (xy, S) \in \overline{M}$, $\forall i \in S$, we have $b_i((xy, S) \mid h) = 1$ and $\forall (xy, S) \in M(h) \setminus \overline{M}$, $\exists i \in S$ such that $b_i((xy, S) \mid h) = 0$;

(ii) $f((b_i)_{i \in N} \mid h) = (h^+h^+, \emptyset)$ if $\forall (xy, S) \in M(h)$, $\exists i \in S$ such that $b_i((xy, S) \mid h) = 0$.

Individual behaviors depend on histories only. In particular, individual behaviors are not allowed to depend on the set of moves on which there has been agreement in the past. One interpretation consistent with such individual behaviors is that after each history the individuals behaviors are transmitted to a mediator, which determines a move in the set of moves on which there is agreement, or selects the no-move when no agreement is possible. The mediator reports this move in the agreement set or the no-move to the individuals, but not the agreement set itself.

A profile of individual behaviors induces a social behavior or a number of social behaviors. A social behavior is induced by a profile of individual behaviors if for each history the move

prescribed by the social behavior is a move on which there is agreement by all individuals involved in the move, or the no-move when no agreement is possible.

3.2 Beliefs, Conjectures and Payoffs

A problem arises when there are several moves on which agreement is possible. One alternative is to assume that all individuals have uniform implementability prior-beliefs on the set $M(h)$. The likelihood of a particular move in the set of moves on which there is agreement, is then determined by Bayesian updating. This results in uniform ex post beliefs on the agreement set. We allow the individuals to have general implementability prior-beliefs on the set $M(h)$. Moreover, it is assumed that the implementability prior-beliefs of the individuals are cautious. Let $q_i = (q_i(\cdot \mid h))_{h \in H(J)}$ be the implementability prior-belief of individual i, where $q_i(\cdot \mid h) : M(h) \to \Delta^0(M(h))$. Hence, given a set of agreements $\overline{M} \subseteq M(h)$, the probability individual i assigns to the implementation of the move $\overline{m} \in \overline{M}$ is given by $\overline{q}_i(\overline{m} \mid h, \overline{M}) = (q_i(\overline{m} \mid h)) \cdot [\sum_{m \in \overline{M}} q_i(m \mid h)]^{-1}$ if $\overline{m} \in \overline{M}$ and $\overline{q}_i(\overline{m} \mid h, \overline{M}) = 0$ otherwise. Let Q_i be the set of all functions q_i.

The basis for rationalizability is that individuals form conjectures about each others' behavior and then optimize subject to these conjectures. We restrict the individuals to hold uncorrelated conjectures[6] about the behaviors of their opponents. After each history $h \in H_i(J)$ at which individual i is involved in a move, she holds such conjectures. A conjecture of individual i is a mapping $c_i : H_i(J) \to \prod_{j \neq i} \Delta(B_j)$. We denote by $c_i(h')(b_{-i})$ the probability individual i conjectures at history h' that her opponents behavior is b_{-i}. We denote $c_i^j(h')(b_j) \in \Delta(B_j)$ the probability individual i conjectures at history h' that player j's behavior is b_j. Notice that a conjecture may change as the course of the social situation unfolds, and that there is only a need for an individual to form conjectures when an individual is potentially involved in a move.

A conjecture c_i reaches $h \in H_i(J)$ if there is an individual behavior b_i, and there are individual behaviors of her opponents b_{-i} in the support of c_i such that (b_i, b_{-i}) reaches h. A profile (b_i, b_{-i}) reaches $h = (x_0, m_1, ..., m_k)$ if $b_i(m_j \mid h^j) = 1 \ \forall i \in m_j^c$, $j = 1, ..., k$. A behavior b_i reaches h if there is b_{-i} such that (b_i, b_{-i}) reaches h. A set $A_{-i} \subseteq B_{-i}$ reaches h if there is (b_i, b_{-i}) with $b_{-i} \in A_{-i}$ reaching h.

Given $b_i \in B_i$, $q_i \in Q_i$, $c_i : H_i(J) \to \prod_{j \neq i} \Delta(B_j)$, and $h' \in H_i(J)$, the probability individual i at h' believes that history $h = (x_0, m_1, m_2, ..., m_k) \geq h'$ will be followed by the move m_{k+1} is denoted by $d_i(h')(m_{k+1} \mid h)$ with $d_i(h')(\cdot \mid h) \in \Delta(M(h) \cup \{(h^+ h^+, \emptyset)\})$. Whenever (b_i, c_i)

[6] The analysis where individuals hold correlated conjectures about the behaviors of their opponents is very similar.

reaches h, this realization probability is determined as follows:

$$d_i(h^{'})(m_{k+1} \mid h) = \frac{\sum_{b_{-i}} \left[p(b_{-i} \mid c_i(h^{'})) \cdot p(h, m_{k+1} \mid b_i, b_{-i}, q_i)\right]}{\sum_{(xy,S)\in M(h)\cup\{(h^{+}h^{+},\emptyset)\}} \sum_{b_{-i}} \left[p(b_{-i} \mid c_i(h^{'})) \cdot p(h, (xy, S) \mid b_i, b_{-i}, q_i)\right]},$$

for $m_{k+1} \in M(h)$, where $p(b_{-i} \mid c_i(h^{'})) = c_i(h^{'})(b_{-i})$ and $p(h, (xy, S) \mid b_i, b_{-i}, q_i)$ is the probability that $h = (x_0, m_1, m_2, ..., m_k)$ realizes and is followed by $m_{k+1} = (xy, S)$, given b_i, b_{-i} and q_i. If $(xy, S) \in M(h)$ then $p(h, (xy, S) \mid b_i, b_{-i}, q_i) = \prod_{j=1}^{k+1} \overline{q}_i(m_j \mid h^j, f(b_i, b_{-i} \mid h^j))$. If $(xy, S) = (h^{+}h^{+}, \emptyset)$ then $p(h, (xy, S) \mid b_i, b_{-i}, q_i) = \prod_{j=1}^{k} \overline{q}_i(m_j \mid h^j, f(b_i, b_{-i} \mid h^j))$ if $f(b_i, b_{-i} \mid h) = (h^{+}h^{+}, \emptyset)$, and $p(h, (xy, S) \mid b_i, b_{-i}, q_i) = 0$ if $f(b_i, b_{-i} \mid h) \neq (h^{+}h^{+}, \emptyset)$, which reflects that when there is agreement on some moves the no-move is never implemented.

Given (b_i, c_i, q_i), where (b_i, c_i) reaches h', the expected utility of individual i conditional on reaching history $h^{'}$ is

$$U_i(h^{'})(b_i, c_i, q_i) = \sum_{x\in Z} \left[\sum_{(h', m_{l(h')}, ..., m_k)\in h^{-1}(\{x\})} \prod_{j=l(h')}^{k} d_i(h^{'})(m_j \mid (h', m_{l(h')}, \ldots, m_{j-1})) \right] u_i(x),$$

where $h^{-1}(\{x\}) = \{h \in H(J) \mid l(h) = J \text{ and } h^{+} = x \text{ or } h = (x_0, m_1, ..., m_{k-1}, (xx, \emptyset)) \text{ with } k < J\}$ is the set of histories of length at most J ending at $x \in Z$.

3.3 Social Rationalizability

We next propose two alternative definitions of social rationalizability which we show to be equivalent. The first one is strongly influenced by Battigalli's [1] extensive-form rationalizability and is based on the notion of a hierarchy of nested hypotheses. The second one is motivated by Pearce's [8] original extensive-form rationalizability and is based on a reduction procedure.[7]

The concept of social rationalizability based on the approach of Battigalli is based on two assumptions: (1) the individuals are rational and endowed with a hierarchy of hypotheses, and (2) this is common knowledge at the original status-quo. A rational individual i maximizes her expected payoff at each history h reached by the play, subject to her *consistent* updating system of conjectures, c_i.

[7] Pearce's [8] extensive-form rationalizability (EFR), like most extensive-form theories, does not adequately deal with counterfactuals and strategic manipulations of conjectures. Battigalli [1] overcomes such drawbacks by providing an alternative characterization of EFR which is not a reduction procedure. Only individuals' updating systems of conjectures are restricted. Such restrictions are modeled as a hierarchy of nested hypotheses, ruling out strategic manipulation. This hierarchy corresponds to the sequence of strategy sets given by Pearce's [8] iterative deletion procedure.

Definition 3 *A consistent updating system for individual i is a mapping $c_i : H_i(J) \to \prod_{j \neq i} \Delta(B_j)$, such that for all $g, h \in H_i(J)$:*
(i) $c_i(h)$ reaches h,
(ii) if $g < h$ and $c_i(g)$ reaches h, then $c_i(g) = c_i(h)$.

The consistency of the updating system requires that the conjecture at history h is consistent with h being reached and that no conjecture is changed unless falsified. That is, individuals update according to Bayes rule whenever possible. An individual behavior b_i is individually rational if it is a best response to some cautious consistent updating system c_i and to some implementability prior-belief q_i. In Definition 4, R_i^1 is the set of individual behaviors of i that are individually rational. Higher degrees of rationality are constructed recursively.

Definition 4 *Let $R^0 = \prod_{i \in I} B_i$. For $n \geq 1$, $R^n = \prod_{i \in I} R_i^n$ is inductively defined as follows: for all $i \in I$, $b_i \in R_i^n$ if there exists $q_i \in Q_i$ and a consistent updating system c_i such that*
(i) for all $h' \in H_i(J)$, $c_i(h') \in \prod_{j \neq i} \Delta^0(R_j^{k^*})$ where k^* is the maximal element in $\{0, 1, ..., n-1\}$ such that $R_{-i}^{k^*}$ reaches h',
(ii) for all $h' \in H_i(J)$, if b_i reaches h', then b_i is a best response to $(c_i(h'), q_i)$ at h', that is, for all $\widehat{b}_i \in B_i$, $U_i(h')(b_i, c_i, q_i) \geq U_i(h')(b_i/\widehat{b}_i^{h'}, c_i, q_i)$, where $b_i/\widehat{b}_i^{h'}$ is the behavior which results from b_i when behavior at h' and its followers $g > h'$ is specified by $\widehat{b}_i$.
The set $R^\infty(J) = \lim_{n \to \infty} R^n$ is the set of rationalizable individual behaviors where histories contain at most J moves.

Definition 4 can be interpreted as follows. The sequence R_j^1, R_j^2, R_j^3, ... $(j \neq i)$ represents for individual i a hierarchy of increasingly strong hypotheses about the behavior of individual j. When individual i adopts a behavior $b_i \in R_i^\infty(J)$, she always holds the strongest hypothesis which is consistent with the history reached (part (i) in Definition 4) and optimizes accordingly. Two important distinctions to extensive form rationalizability are that optimization takes place against both c_i and q_i, and that conjectures are cautious.

The concept of social rationalizability based on the ideas in Pearce [8] is a reduction procedure and is defined as follows.

Definition 5 *Let $P^0 = \prod_{i \in I} B_i$. For $n \geq 1$, $P^n = \prod_{i \in I} P_i^n$ is inductively defined as follows: for all $i \in I$, $b_i \in P_i^n$ if*
(i) $b_i \in P_i^{n-1}$,
(ii) there exists $q_i \in Q_i$ and a consistent updating system c_i such that for all $h' \in H_i(J)$ that are reached by b_i and P_{-i}^{n-1} it holds
(a) $c_i(h') \in \prod_{j \neq i} \Delta^0(P_j^{n-1})$,
(b) for all $\widehat{b}_i \in P_i^{n-1}$, $U_i(h')(b_i, c_i, q_i) \geq U_i(h')(b_i/\widehat{b}_i^{h'}, c_i, q_i)$.

The set $P^{\infty}(J) = \lim_{n\to\infty} P^n$ *is the set of rationalizable individual behaviors where histories contain at most* J *moves.*

Theorem 1 claims that the two definitions of social rationalizability are equivalent. Throughout the rest of the paper we focus on social rationalizability à la Pearce.

Theorem 1 *For all* $n \geq 0, R^n = P^n$.

Proof. Obviously, $R^0 = P^0$. We give a proof by induction, so suppose $R^{n-1} = P^{n-1}$. Consider some $b_i \in R_i^n$. Since $R^n \subseteq R^{n-1} = P^{n-1}$, it holds that $b_i \in P_i^{n-1}$, and Condition (i) in Definition 5 is satisfied. Suppose $h' \in H_i(J)$ is reached by b_i and P_{-i}^{n-1}. By the definition of R_i^n, there exists $q_i \in Q_i$ and a consistent updating system c_i such that $c_i(h') \in \prod_{j\neq i} \Delta^0(R_j^{n-1}) = \prod_{j\neq i} \Delta^0(P_j^{n-1})$ and b_i is a best response to $(c_i(h'), q_i)$ at h', that is, for all $\widehat{b}^i \in B_i \supseteq P_i^{n-1}$, $U_i(h')(b_i, c_i, q_i) \geq U_i(h')(b_i/\widehat{b}_i^{h'}, c_i, q_i)$. It follows that Conditions (iia) and (iib) in Definition 5 are satisfied, so $b_i \in P_i^n$.

Consider some $b_i \in P_i^n$. Since $P^n \subseteq P^{n-1} = R^{n-1}$, it holds that $b_i \in R_i^{n-1}$. Since $b_i \in R_i^{n-1}$, there exists $q_i \in Q_i$ and a consistent updating system c_i such that if b_i reaches $h \in H_i(J)$ then b_i is a best response to $c_i(h) \in \prod_{j\neq i} \Delta^0(R_j^{k^*})$ and q_i, where $k^* \leq n-2$. Since $b_i \in P_i^n$, there exists $\widehat{q}_i \in Q_i$ and a consistent updating system $\widehat{c}_i$ such that if b_i and $P_{-i}^{n-1} = R_{-i}^{n-1}$ reach $h \in H_i(J)$, then $\widehat{c}_i(h) \in \prod_{j\neq i} \Delta^0(P_j^{n-1}) = \prod_{j\neq i} \Delta^0(R_j^{n-1})$, and for all $\widehat{b}_i \in P_i^{n-1} = R_i^{n-1}$, $U_i(h)(b_i, \widehat{c}_i, \widehat{q}_i) \geq U_i(h)(b_i/\widehat{b}_i, \widehat{c}_i, \widehat{q}_i)$. The use of a cautious $\widehat{q}_i$ and a cautious consistent updating system $\widehat{c}_i$ implies that $\widehat{c}_i(h) = \widehat{c}_i(h') \in \prod_{j\neq i} \Delta^0(R_j^{n-1})$ for all $h, h' \in H_i(J)$ reached by R_{-i}^{n-1}.

We define $\widetilde{c}_i$ by

$$\begin{aligned} \widetilde{c}_i(h) &= \widehat{c}_i(h) \text{ if } h \in H_i(J) \text{ is reached by } R_i^{n-1} \\ \widetilde{c}_i(h) &= c_i(h) \text{ if } h \in H_i(J) \text{ is not reached by } R_i^{n-1}, \end{aligned}$$

and $\widetilde{q}_i$ by

$$\begin{aligned} \widetilde{q}_i(\cdot \mid h) &= \widehat{q}_i(\cdot \mid h) \text{ if } h \in H(J) \text{ is reached by } R_i^{n-1}, \\ \widetilde{q}_i(\cdot \mid h) &= q_i(\cdot \mid h) \text{ if } h \in H(J) \text{ is not reached by } R_i^{n-1}. \end{aligned}$$

It can be verified that $\widetilde{q}_i \in Q_i$ and that $\widetilde{c}_i$ is a consistent updating system.

For k^* the maximal element in $\{0, 1, \ldots, n-1\}$ such that $R_{-i}^{k^*}$ reaches $h' \in H_i(J)$, it holds that $\widetilde{c}_i(h') \in \prod_{j\neq i} \Delta^0(R_j^{k^*})$, so $\widetilde{c}_i$ satisfies Condition (i) of Definition 4.

It remains to be shown that for all $h' \in H_i(J)$, if b_i reaches h', then b_i is a best response to $(\widetilde{c}_i(h'), \widetilde{q}_i)$ at h', that is, for all $\widehat{b}_i \in B_i$, $U_i(h')(b_i, \widetilde{c}_i, \widetilde{q}_i) \geq U_i(h')(b_i/\widehat{b}_i^{h'}, \widetilde{c}_i, \widetilde{q}_i)$. If h' is not reached by R_i^{n-1}, then b_i is a best response to $c_i(h') \in \prod_{j\neq i} \Delta^0(R_j^{k^*})$ and q_i, where $k^* \leq n-2$, and, by

definition of $\widetilde{c}_i, \widetilde{q}_i$, b_i is therefore a best response to $(\widetilde{c}_i(h'), \widetilde{q}_i)$ at h'. If h' is reached by R_i^{n-1}, then for all $\widehat{b}_i \in P_i^{n-1}$,

$$U_i(h')(b_i, \widehat{c}_i, \widehat{q}_i) \geq U_i(h')(b_i/\widehat{b}_i^{h'}, \widehat{c}_i, \widehat{q}_i),$$

and so, by definition of $(\widetilde{c}_i, \widetilde{q}_i)$,

$$U_i(h')(b_i, \widetilde{c}_i, \widetilde{q}_i) \geq U_i(h')(b_i/\widehat{b}_i^{h'}, \widetilde{c}_i, \widetilde{q}_i).$$

It remains to be shown that there is no $\widehat{b}_i \in B_i \setminus P_i^{n-1}$ such that

$$U_i(h')(b_i, \widetilde{c}_i, \widetilde{q}_i) < U_i(h')(b_i/\widehat{b}_i^{h'}, \widetilde{c}_i, \widetilde{q}_i).$$

Since h' is reached by b_i and R_i^{n-1}, h' occurs with positive probability. But then $U_i(x_0)(b_i, \widetilde{c}_i, \widetilde{q}_i) < U_i(x_0)(b_i/\widehat{b}_i^{h'}, \widetilde{c}_i, \widetilde{q}_i)$. Let

$$\widehat{B}_i = \{\widehat{b}_i \in B_i \mid \widehat{b}_i \text{ maximizes } U_i(x_0)(\widehat{b}_i, \widetilde{c}_i, \widetilde{q}_i)\}.$$

Notice that $\widehat{B}_i \subseteq B_i \setminus P_i^{n-1}$. Let $k \leq n-2$ be the smallest integer such that $\widehat{B}_i \cap P_i^k \neq \emptyset$. We will show that one of the elements of $\widehat{B}_i$ belongs to P_i^{k+1}. To do so we need a consistent updating system $\overline{c}_i$ such that $\overline{c}_i(h)$ belongs to $\prod_{j \neq i} \Delta^0(P_j^k)$ for all histories $h \in H_i(J)$ reached by P_{-i}^k against which some member of $\widehat{B}_i$ is a best response in P_i^k. Consider a perturbation $\overline{c}_i(h')$ of $\widetilde{c}_i(h')$ that belongs to $\prod_{j \neq i} \Delta^0(P_j^k)$, choose $\overline{c}_i(h) = \overline{c}_i(h')$ for all histories $h \in H_i(J)$ reached by P_{-i}^k and choose $\overline{c}_i(h)$ at other histories such that $\overline{c}_i$ is consistent. The perturbation $\overline{c}_i$ can be chosen small enough to guarantee that $U_i(x_0)(\widehat{b}_i, \overline{c}_i, \widetilde{q}_i) > U_i(x_0)(\overline{b}_i, \overline{c}_i, \widetilde{q}_i)$, for all $\widehat{b}_i \in \widehat{B}_i$, for all $\overline{b}_i \in B_i \setminus \widehat{B}_i$. Consider an optimal choice in P_i^k against $(\overline{c}_i, \widetilde{q}_i)$. Obviously it is an element of $\widehat{B}_i$, but then $\widehat{B}_i \cap P_i^{k+1} \neq \emptyset$, contradicting the definition of k. ■

Obviously, from Theorem 1, $R^\infty(J) = P^\infty(J)$. Let $S^\infty(J)$ denote the set of rationalizable social behaviors. A social behavior b belongs to $S^\infty(J)$ if there exists $(b_i)_{i \in I} \in P^\infty(J)$ such that $b(h) = m \in M(h)$ implies $b_i(m \mid h) = 1$, $\forall i \in m^c$, and $b(h) = (h^+h^+, \emptyset)$ implies $f(b_i, b_{-i} \mid h) = \emptyset$.

We denote by $Z_J^\infty(x_0)$ the set of rationalizable outcomes with original status-quo $x_0 \in Z$. It is given by $Z_J^\infty(x_0) = \{x \in Z \mid \exists (x_0, m_1, ..., m_k) \in h^{-1}(\{x\}),\ \exists b \in S^\infty(J)$ such that $\forall j = 1, ..., k$, $b(x_0, m_1, ..., m_{j-1}) = m_j\}$. The set of socially rationalizable outcomes, $Z^\infty(x_0)$, is obtained by letting J go to infinity, $Z^\infty(x_0) = \limsup_{J \to \infty} Z_J^\infty(x_0)$. The set of socially rationalizable outcomes is never empty.

Theorem 2 $Z^\infty(x_0) \neq \emptyset$.

Proof. Consider the iterative procedure provided by Definition 5. For each iteration n, choose a $q_i \in Q_i$ and a consistent updating system c_i such that $c_i(h') \in \prod_{j \neq i} \Delta^0(P_j^{n-1})$ for all $h' \in H_i(J)$ reached by P^{n-1}. Consider any $b_i \in P_i^{n-1}$ such that $U_i(x_0)(b_i, c_i, q_i) \geq U_i(x_0)(\widehat{b}_i, c_i, q_i)$ for all

$\widehat{b}_i \in P_i^{n-1}$. If $h^{'}$ is reached by b_i and P_{-i}^{n-1} then it follows as in the proof of Theorem 1 that $U_i(h^{'})(b_i, c_i, q_i) \geq U_i(h^{'})(b_i/\widehat{b}_i^{h^{'}}, c_i, q_i)$ for all $\widehat{b}_i \in P_i^{n-1}$. It follows that $b_i \in P_i^n$, so $P^n \neq \emptyset$. Since P^0 is finite and $P^n \supseteq P^{n+1}$, there is N such that $P^n = P^{n^{'}}$ for all $n, n^{'} \geq N$. It follows that $P^\infty(J) = P^N \neq \emptyset$. Any $(b_i)_{i \in I} \in P^\infty(J)$ yields a social behavior $b \in S^\infty(J)$, so $S^\infty(J) \neq \emptyset$; and as a consequence $Z_J^\infty(x_0) \neq \emptyset$. As a subset of the finite set Z it holds that $Z_J^\infty(x_0)$ is finite. Now it follows from the definition of the limit superior that $Z^\infty(x_0) \neq \emptyset$. ■

We reconsider the five examples and we show that social rationalizability remedies the problems of the largest consistent set, the optimistic stable standard of behavior, and the conservative stable standard of behavior. Even though the definitions so far may seem rather complicated, the examples are easily solved for by the reduction procedure of Definition 5.

EXAMPLE 1: Consider again the social environment where $I = \{1\}$, $Z = \{x_0, x_1, x_2\}$, and the effectiveness relations as well as the payoffs are depicted in Figure 1. We have $H_1(J) = \{(x_0)\}$ and $M_1(x_0) = \{(x_0x_1, \{1\}), (x_0x_2, \{1\})\}$. Any behavior of individual 1 is such that $b_1((x_0x_1, \{1\}) \mid (x_0)) = 1$ or 0 and $b_1((x_0x_2, \{1\}) \mid (x_0)) = 1$ or 0. For simplicity, we denote the set of all behaviors of individual 1 as $B_1 = \{(0,0), (0,1), (1,0), (1,1)\}$ where $(0,1)$ means that $b_1((x_0x_1, \{1\}) \mid (x_0)) = 0$ and $b_1((x_0x_2, \{1\}) \mid (x_0)) = 1$. By Definition 5, $P^0 = B_1$. Obviously, the unique best response for individual 1 is her behavior $(0,1)$. Hence, this social environment has a unique rationalizable social behavior $b(x_0) = (x_0x_2, \{1\})$ and a unique rationalizable outcome $Z^\infty(x_0) = \{x_2\}$. So, contrary to the largest consistent set, social rationalizability satisfies individual rationality.

EXAMPLE 2: Consider again the social environment where $I = \{1,2,3\}$, $Z = \{x_0, x_1, x_2, x_3\}$, and the effectiveness relations as well as the payoffs are depicted in Figure 2. Let $h^1 = (x_0)$, $h^2 = (x_0, (x_0x_1, \{1\}))$ and $h^3 = (x_0, (x_0x_1, \{1\}), (x_1x_2, \{2\}))$. We have $H_i(J) = \{h^i\}$ and $M_i(h^i) = \{(x_{i-1}x_i, \{i\})\}$, $i = 1,2,3$. Any behavior of individual i is such that $b_i((x_{i-1}x_i, \{i\}) \mid h^i) = 1$ or 0. The set of all behaviors of individual i is $B_i = \{0,1\}$, $i \in I$. By Definition 5, $P^0 = B_1 \times B_2 \times B_3$. When individual 3 gets the move, she is really indifferent between moving and not moving, so $P_3^1 = B_3$. When individual 2 contemplates the move from x_1 to x_2, he conjectures a positive probability to individual 3 moving to x_3. Indeed, any $c_2(h^2) \in \Delta^0(B_1) \times \Delta^0(B_3)$ puts positive probability weight on both $b_3((x_2x_3, \{3\}) \mid h^3) = 1$ and $b_3((x_2x_3, \{3\}) \mid h^3) = 0$. Hence, the unique optimal behavior for individual 2 is $b_2((x_1x_2, \{2\}) \mid h^2) = 1$, and P_2^1 is a proper subset of B_2: $P_2^1 = \{1\}$. Initially, individual 1 puts positive probability weight on all behaviors of 2 and 3, and depending on her conjectures she decides to stay at x_0 or to move to x_1, so $P_1^1 = B_1$. However, in the second iteration she knows that individual 2 will move

to x_2 for sure when given the move: any $c_1(h^1) \in \Delta^0(P_2^1) \times \Delta^0(P_3^1)$ gives probability one to $b_2((x_1x_2, \{2\}) \mid h^2) = 1$. Therefore, the unique optimal behavior for individual 1 is to stay at x_0: $b_1((x_0x_1, \{1\}) \mid h^1) = 0$. So, $P_1^\infty = \{0\}$, $P_2^\infty = \{1\}$ and $P_3^\infty = B_3$. The unique rationalizable (or stable) outcome is the original status-quo, $Z^\infty(x_0) = \{x_0\}$.

EXAMPLE 3: Consider again the social environment where $I = \{1, 2\}$, $Z = \{x_0, x_1, x_2\}$, and the effectiveness relations as well as the payoffs are depicted in Figure 3. Let $h^1 = (x_0)$. We have $H_i(J) = \{h^1\}$ and $M_i(h^1) = \{(x_0x_i, \{i\})\}$, $i \in I$. Any behavior of individual i is such that $b_i((x_0x_i, \{i\}) \mid h^1) = 1$ or 0. The set of all behaviors of individual i is $B_i = \{0, 1\}$, $i \in I$. By Definition 5, $P^0 = B_1 \times B_2$. Given any $q_i \in Q_i$ and any $c_i(h^1) \in \Delta^0(B_{-i})$, individual i has a unique best response which is to move to x_i. So, $b_i((x_0x_i, \{i\}) \mid h^1) = 1$, $P_i^1 = P_i^\infty = \{1\}$, $i \in I$, and $Z^\infty(x_0) = \{x_1, x_2\}$.

EXAMPLE 4: Consider again the social environment where $I = \{1, 2\}$, $Z = \{x_0, x_1, x_2\}$, and the effectiveness relations as well as the payoffs are depicted in Figure 4. Let $h^1 = (x_0)$. We have $H_i(J) = \{h^1\}$ and $M_i(h^1) = \{(x_0x_i, \{i\})\}$, $i \in I$. Any behavior of individual i is such that $b_i((x_0x_i, \{i\}) \mid h^1) = 1$ or 0. The set of all behaviors of individual i is $B_i = \{0, 1\}$, $i \in I$. By Definition 5, $P^0 = B_1 \times B_2$. Given any $q_i \in Q_i$ and any $c_i(h^1) \in \Delta^0(B_{-i})$, individual i has a unique best response which is not to move. So, $b_i((x_0x_i, \{i\}) \mid h^1) = 0$, $P_i^1 = P_i^\infty = \{0\}$, $i \in I$, and $Z^\infty(x_0) = \{x_0\}$.

EXAMPLE 5: It is possible to describe the classical prisoners' dilemma as a social environment. The set of pure strategies of individual i is $Z_i = \{\text{cooperate, defect}\}$, $i = 1, 2$, and $Z = Z_1 \times Z_2$ is the set of strategy profiles. Assume that coalitions cannot form. Then, one way to represent

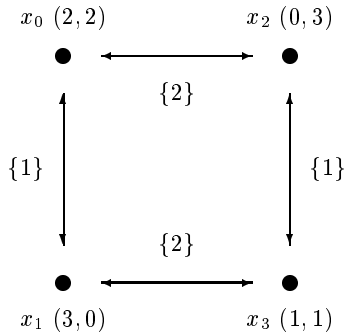

Figure 5: The prisoners' dilemma

this normal-form game as a social environment is Greenberg's [5] individual contingent threats

situation, where $\forall x, y \in Z$, $x \to_S y$ if $S = \{i\}$ and $x_{I\setminus\{i\}} = y_{I\setminus\{i\}}$. This social environment is depicted in Figure 5, where x_0 = (cooperate, cooperate), x_1 = (defect, cooperate), x_2 = (cooperate, defect), and x_3 = (defect, defect).

Consider first the case $J = 1$. Let $h^1 = (x_0)$. We have $H_i(J) = \{h^1\}$ and $M_i(h^1) = \{(x_0 x_i, \{i\})\}$, $i = 1, 2$. Any behavior of individual i is such that $b_i((x_0 x_i, \{i\}) \mid h^1) = 1$ or 0. The set of all behaviors of individual i is $B_i = \{0, 1\}$. By Definition 5, $P^0 = B_1 \times B_2$. Given any $q_i \in Q_i$ and any $c_i(h^1) \in \Delta^0(B_{-i})$, individual i has a unique best response which is to move: $b_i((x_0 x_i, \{i\}) \mid h^1) = 1$. So, $P_i^1 = P_i^\infty = \{1\}$, $i = 1, 2$, and $Z_1^\infty(x_0) = \{x_1, x_2\}$.

Consider now the case $J = 2$. Let $h^1 = (x_0)$, $h^2 = (x_0, (x_0 x_1, \{1\}))$, $h^3 = (x_0, (x_0 x_2, \{2\}))$. We have $H_i(J) = \{h^1, h^2, h^3\}$, $M_1(h^1) = \{(x_0 x_1, \{1\})\}$, $M_1(h^2) = \{(x_1 x_0, \{1\})\}$, $M_1(h^3) = \{(x_2 x_3, \{1\})\}$, $M_2(h^1) = \{(x_0 x_2, \{2\})\}$, $M_2(h^2) = \{(x_1 x_3, \{2\})\}$ and $M_2(h^3) = \{(x_2 x_0, \{2\})\}$. The set of behaviors of individual i is $B_i = \{(0,0,0), (1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)\}$, where $(1,0,1)$ simply means $b_1((x_0 x_1, \{1\}) \mid h^1) = 1$, $b_1((x_1 x_0, \{1\}) \mid h^2) = 0$, $b_1((x_2 x_3, \{1\}) \mid h^3) = 1$ for individual 1, and $b_2((x_0 x_2, \{2\}) \mid h^1) = 1$, $b_2((x_1 x_3, \{2\}) \mid h^2) = 0$, $b_2((x_2 x_0, \{2\}) \mid h^3) = 1$ for individual 2. Let $q_i = (q_i(h^1), q_i(h^2), q_i(h^3))$ where $q_i(h^k) = (q_i^{h^k}, 1 - q_i^{h^k})$ and $q_i^{h^k}$ is the probability assigned by the implementability prior-belief of individual i that her move will be implemented after h^k if both individuals decide to move.

One can show that, for all $b_i \in B_i$ there exists a consistent updating system c_i with $c_i(h^1) \in \Delta^0(B_j)$ and there exists a belief $q_i \in Q_i$ such that b_i is the unique best response among B_i; and so $b_i \in P_i^1$. For instance, $b_1 = (0,1,1)$ is the unique best response against the belief q_1 such that $q_1^{h^1} = q_1^{h^2} = q_1^{h^3} = \frac{1}{2}$ and the conjecture c_1 such that

$$c_1^2(h^1)(b_2) = \begin{cases} \frac{4}{8} & \text{if } b_2 = (0,1,0) \\ \frac{3}{8} & \text{if } b_2 = (0,1,1) \\ \frac{1}{48} & \text{otherwise} \end{cases}.$$

One can verify that, after each history, b_1 is the unique best response, and hence, $b_1 = (0,1,1) \in P_1^1$. In Table 1 we give for each behavior $b_1 \in B_1$ beliefs and conjectures against which it is the unique best response. For example, the fifth column means that $b_1 = (1,1,1)$ is the unique best response against the conjecture c_1 such that

$$c_1^2(h^1)(b_2) = \begin{cases} \frac{2}{8} & \text{if } b_2 = (1,0,0) \text{ or } b_2 = (1,1,0) \\ \frac{3}{8} & \text{if } b_2 = (0,1,1) \\ \frac{1}{40} & \text{otherwise} \end{cases}.$$

and the implementability prior-belief q_1 such that $q_1^{h^1} = q_1^{h^3} = \frac{1}{2}$ and $q_1^{h^2} = \frac{3}{4}$.

Using the symmetry of the prisoners' dilemma it is straightforward that all $b_2 \in B_2$ belong to P_2^1. So, applying social rationalizability to the prisoners' dilemma, we obtain for $J = 2$ that all behaviors are rationalizable: $P_i^1 = B_i = P_i^\infty$, $i = 1, 2$, and $Z_2^\infty(x_0) = \{x_0, x_1, x_2, x_3\}$.

	b_1							
b_2	$(0,0,1)$	$(1,0,1)$	$(0,1,1)$	$(1,1,1)$	$(0,0,0)$	$(0,1,0)$	$(1,0,0)$	$(1,1,0)$
$(0,0,0)$	$\frac{1}{40}$	$\frac{1}{48}$	$\frac{1}{48}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{1}{48}$	$\frac{2}{48}$	$\frac{1}{40}$
$(0,0,1)$	$\frac{1}{40}$	$\frac{1}{48}$	$\frac{1}{48}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{1}{48}$	$\frac{3}{8}$	$\frac{1}{40}$
$(1,0,0)$	$\frac{1}{40}$	$\frac{4}{8}$	$\frac{1}{48}$	$\frac{2}{8}$	$\frac{1}{40}$	$\frac{1}{48}$	$\frac{2}{48}$	$\frac{2}{8}$
$(0,1,0)$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{4}{8}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{1}{48}$	$\frac{2}{48}$	$\frac{1}{40}$
$(1,1,0)$	$\frac{1}{40}$	$\frac{1}{48}$	$\frac{1}{48}$	$\frac{2}{8}$	$\frac{1}{40}$	$\frac{1}{48}$	$\frac{2}{48}$	$\frac{1}{40}$
$(0,1,1)$	$\frac{1}{40}$	$\frac{1}{48}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{4}{8}$	$\frac{2}{48}$	$\frac{4}{8}$
$(1,0,1)$	$\frac{3}{8}$	$\frac{1}{48}$	$\frac{1}{48}$	$\frac{1}{40}$	$\frac{3}{8}$	$\frac{1}{48}$	$\frac{3}{8}$	$\frac{1}{40}$
$(1,1,1)$	$\frac{1}{8}$	$\frac{1}{48}$	$\frac{1}{48}$	$\frac{1}{40}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{2}{48}$	$\frac{1}{8}$
$q_1^{h^1}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$
$q_1^{h^2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$
$q_1^{h^3}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$

Table 1: Best responses, conjectures and beliefs in the prisoners' dilemma.

4 Coalitional Rationality

Social rationalizability is based on common knowledge of individual rationality. An interesting theory of social behavior should also be expected to satisfy at least some rudimentary forms of coalitional rationality. It is conceivable that coalitions fail to choose between a set of outcomes, because of internal disputes on the outcome on which to coordinate. If, on the other hand, the outcomes are Pareto ranked, then a sensible concept of coalitional rationality should prescribe coordination on the outcome that Pareto dominates all the others. We can formalize this within the theory of social environments.

Consider the social environment Γ^* where $I = \{1, 2, ..., \#I\}$, $Z = \{x_0, x_1, ..., x_N\}$, the outcomes are Pareto ranked: $u_i(x_N) > u_i(x_{N-1}) > ... > u_i(x_1) > u_i(x_0) = 0 \ \forall i \in I$, and only $x_0 \rightarrow_I x_k$, $k = 1, ..., N$, are possible moves. A two-individual case with $N = 3$ is depicted in Figure 6. We say that social rationalizability satisfies coalitional rationality if it selects the Pareto-dominant outcome, x_N.

In this social environment Γ^*, we have $I(x_0) = I$, $H_i = \{(x_0)\}$ and $M(x_0) = M_i(x_0) = \{(x_0x_1, I), (x_0x_2, I), ..., (x_0x_N, I)\}$, $\forall i \in I$. A behavior of individual i is denoted by $b_i = (b_{i1}, ..., b_{ik}, ..., b_{iN})$ where $b_{ik} = b_i((x_0x_k, I) \mid (x_0))$; so, b_{ik} is component k of b_i. A belief of individual i over the implementability of agreements is denoted by $q_i = (q_{i1}, ..., q_{ik}, ..., q_{iN})$ where q_{ik} is the probability assigned by the implementability prior-belief of individual i to the move (x_0x_k, I). From now on we denote the history (x_0) by h^1.

EXAMPLE 6: Consider the two-individual and three-move case, $I = \{1, 2\}$, $Z = \{x_0, x_1, x_2, x_3\}$, $x_0 \rightarrow_I x_k$, $k = 1, 2, 3$, are the only possible moves, and the special case where $u_i(x_k) = k$, $\forall k \in \{0, 1, 2, 3\}$, $\forall i \in \{1, 2\}$. This social environment is depicted in Figure 6. The behaviors of individual i are such that $b_i((x_0x_k, \{1, 2\}) \mid h^1) = 1$ or $b_i((x_0x_k, \{1, 2\}) \mid h^1) = 0$, $k = 1, 2, 3$. The set of all behaviors of individual i is $B_i = \{(0, 0, 0),\ (1, 0, 0),\ (0, 1, 0),\ (0, 0, 1),\ (1, 1, 0),\ (1, 0, 1),\ (0, 1, 1),\ (1, 1, 1)\}$, where $(1, 0, 1)$ simply means $b_i((x_0x_1, \{1, 2\}) \mid h^1) = 1$, $b_i((x_0x_2, \{1, 2\}) \mid h^1) = 0$, $b_i((x_0x_3, \{1, 2\}) \mid h^1) = 1$ for individual i, $i = 1, 2$. Which outcomes are socially rationalizable? Is the Pareto-dominant outcome the unique socially rationalizable one?

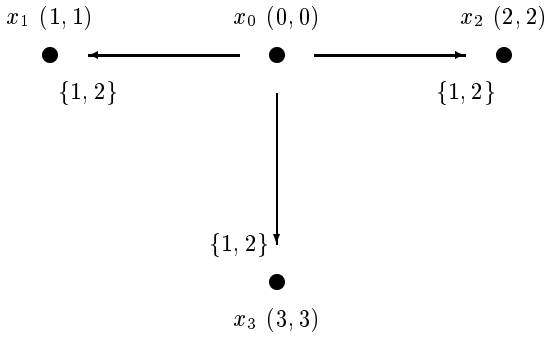

Figure 6: Coalitional rationality.

By Definition 5, $P_i^0 = B_i$. We show first that $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$, $(1, 1, 0)$ do not belong to P_i^1, $i = 1, 2$. Take any $b_i \in B_i$ such that $b_{i3} = 0$ and take $b_i^{'} \in B_i$ such that $b_{i1}^{'} = b_{i1}$, $b_{i2}^{'} = b_{i2}$ and $b_{i3}^{'} = 1$. It is quite straightforward that, for all $c_i(h^1) \in \Delta^0(B_j)$ and for all $q_i \in Q_i$, $U_i(h^1)(b_i, c_i, q_i) < U_i(h^1)(b_i^{'}, c_i, q_i)$. Indeed, the behaviors b_i and $b_i^{'}$ give the same payoffs to individual i against the opponent's behaviors b_j with $b_{j3} = 0$, but $b_i^{'}$ does strictly better than b_i against the opponent's behaviors with $b_{j3} = 1$.

Next it is shown that all $b_i \in B_i$ with $b_{i3} = 1$ belong to P_i^1, $i = 1, 2$. For any b_i with $b_{i3} = 1$, there exists $c_i(h^1) \in \Delta(B_j)$ and $q_i \in Q_i$ such that b_i is the unique best response among B_i. For instance, the behavior $b_i = (1, 0, 1)$ is the unique best response against the belief $q_i = (\frac{1}{81}, \frac{71}{81}, \frac{1}{9})$ and the conjecture $c_i(h^1) \in \Delta(B_j)$ such that

$$c_i^j(h^1)(b_j) = \begin{cases} \frac{3}{7} & \text{if } b_j = (1, 0, 0) \text{ or } b_j = (0, 0, 1) \\ \frac{1}{7} & \text{if } b_j = (1, 1, 1) \\ 0 & \text{otherwise} \end{cases} .$$

In Table 2 we give beliefs and conjectures against which each behavior b_i with $b_{i3} = 1$ is the unique best response. By a continuity argument, see also Lemma 1 below, b_i is also the unique best response against the belief q_i and a cautious conjecture that puts weight on all behaviors $b_j \in B_j$. So, $P_i^1 = \{(0, 0, 1),\ (1, 0, 1),\ (0, 1, 1),\ (1, 1, 1)\}$, $i = 1, 2$.

	b_i			
b_j	$(0,0,1)$	$(1,0,1)$	$(0,1,1)$	$(1,1,1)$
$(0,0,0)$	0	0	0	0
$(0,0,1)$	$\frac{3}{4}$	$\frac{3}{7}$	$\frac{3}{7}$	$\frac{1}{3}$
$(1,0,0)$	0	$\frac{3}{7}$	0	$\frac{1}{3}$
$(0,1,0)$	0	0	$\frac{3}{7}$	$\frac{1}{3}$
$(1,1,0)$	0	0	0	0
$(0,1,1)$	0	0	0	0
$(1,0,1)$	0	0	0	0
$(1,1,1)$	$\frac{1}{4}$	$\frac{1}{7}$	$\frac{1}{7}$	0
q_{i1}	$\frac{4}{9}$	$\frac{1}{81}$	$\frac{71}{81}$	$\frac{1}{3}$
q_{i2}	$\frac{4}{9}$	$\frac{71}{81}$	$\frac{1}{81}$	$\frac{1}{3}$
q_{i3}	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$

Table 2: Unique best response, conjecture and belief.

In the second iteration, individual i knows that individual j will play a behavior in P_j^1. Hence, for all $c_i(h^1) \in \Delta^0(P_j^1)$ and for all $q_i \in Q_i$, the unique best response of individual i is the behavior $b_i = (0,0,1)$ which gives her a payoff of 3. Indeed, for all $c_i(h^1) \in \Delta^0(P_j^1)$ and for all $q_i \in Q_i$, any $b_i^{'} \neq b_i$ belonging to P_i^1 will give her a payoff less than 3, because $c_i(h^1)$ puts positive probability on $b_j^{'} = b_i^{'}$ and q_i has full support. So, $P_i^2 = \{(0,0,1)\} = P_i^\infty$, $i = 1,2$, and $Z^\infty(x_0) = \{x_3\}$. In Example 6, the case with two individuals and three Pareto ranked moves, the property of coalitional rationality is satisfied. There is a unique socially rationalizable outcome and it is the Pareto-dominant one.

We show that the coalitional rationality property holds in general in the social environment Γ^*. In order to do so we use the following five lemmas. Lemma 1 tells us that if a behavior of individual i is the unique best response against a conjecture c_i (possibly degenerate) and a belief q_i, then it is also the unique best response against some cautious conjecture c_i^* and the belief q_i.

Lemma 1 *Take any $b_i \in B_i$. If there exists $q_i \in Q_i$ and c_i such that (i) $c_i(h^1) \in \prod_{j \neq i} \Delta(B_j)$ and (ii) for all $b_i^{'} \in B_i$, $b_i^{'} \neq b_i$, $U_i(h^1)(b_i, c_i, q_i) > U_i(h^1)(b_i^{'}, c_i, q_i)$, then there exists c_i^* such that (iii) $c_i^*(h^1) \in \prod_{j \neq i} \Delta^0(B_j)$ and (iv) for all $b_i^{'} \in B_i$, $b_i^{'} \neq b_i$, $U_i(h^1)(b_i, c_i^*, q_i) > U_i(h^1)(b_i^{'}, c_i^*, q_i)$.*

Proof. Let c_i^* be a conjecture that puts probability $\varepsilon / \# B_j$ on each behavior $b_j \in B_j$ plus $(1-\varepsilon)$ times the probabilities put by conjecture $c_i(h^1)$ on the behaviors in B_j, $j \neq i$. Then $c_i^*(h^1) \in \prod_{j \neq i} \Delta^0(B_j)$, and using that B_i is a finite set, and that $U_i(h^1)$ varies continuously with ε, it follows that $\varepsilon > 0$ can be chosen small enough that $U_i(h^1)(b_i, c_i^*, q_i) > U_i(h^1)(b_i', c_i^*, q_i)$. ■

Lemma 2 is useful to prove Lemma 3 and Lemma 4.

Lemma 2 *Take $y_1, y_2, y_3, y_4 \in \mathbb{R}_+$ with $y_2, y_4 > 0$. Then,*

$$\frac{y_1}{y_2} < \frac{y_1 + y_3}{y_2 + y_4} \text{ if and only if } \frac{y_1}{y_2} < \frac{y_3}{y_4}.$$

Proof. Follows from a straightforward manipulation of the formula. ■

Lemma 3 tells us that any individual behavior b_i such that individual i blocks the move to x_N (i.e. $b_i((x_0x_N, I) \mid h^1) = 0$) is never a best response whatever the conjecture c_i and the implementability prior-belief q_i. Indeed, the behavior $b_i^{'}$, where $b_i^{'}$ is the same as b_i except that individual i joins the move to x_N, is always a strictly better response.

Lemma 3 *Take any $b_i \in B_i$ with $b_{iN} = 0$. Take $b_i^{'} \in B_i$ such that $b_{ik}^{'} = b_{ik}$ for $k = 1, ..., N-1$ and $b_{iN}^{'} = 1$. Then, $U_i(h^1)(b_i^{'}, c_i, q_i) > U_i(h^1)(b_i, c_i, q_i)$ for all $c_i \in \prod_{j \neq i} \Delta^0(B_j)$ and all $q_i \in Q_i$.*

Proof. Consider any profile $b_{-i} \in \prod_{j \neq i} B_j$. Let $\overline{f}(b_{-i} \mid h^1)$ be the agreement set without individual i, that is, all the moves after h^1 on which the opponents of individual i agree when their behavior is b_{-i}.
(i) For all $b_{-i} \in \prod_{j \neq i} B_j$ and $q_i \in Q_i$, if $(x_0x_N, I) \notin \overline{f}(b_{-i} \mid h^1)$ then $U_i(h^1)(b_i^{'}, b_{-i}, q_i) = U_i(h^1)(b_i, b_{-i}, q_i)$.
(ii) For all $b_{-i} \in \prod_{j \neq i} B_j$ and $q_i \in Q_i$, if $(x_0x_N, I) \in \overline{f}(b_{-i} \mid h^1)$ then
- if for every k such that $b_{ik} = 1$ we have $(x_0x_k, I) \notin \overline{f}(b_{-i} \mid h^1)$ then $0 = U_i(h^1)(b_i, b_{-i}, q_i) < U_i(h^1)(b_i^{'}, b_{-i}, q_i)$,
- otherwise,

$$U_i(h^1)(b_i, b_{-i}, q_i) = \frac{\sum_{(x_0x_k, I) \in f(b_i, b_{-i} \mid h^1)} u_i(x_k) \cdot q_{ik}}{\sum_{(x_0x_k, I) \in f(b_i, b_{-i} \mid h^1)} q_{ik}} \qquad (= \frac{y_1}{y_2})$$

and

$$U_i(h^1)(b_i^{'}, b_{-i}, q_i) = \frac{\sum_{(x_0x_k, I) \in f(b_i, b_{-i} \mid h^1)} u_i(x_k) \cdot q_{ik} + u_i(x_N) \cdot q_{iN}}{\sum_{(x_0x_k, I) \in f(b_i, b_{-i} \mid h^1)} q_{ik} + q_{iN}} \qquad (= \frac{y_1 + y_3}{y_2 + y_4});$$

since $u_i(x_N) > \sum_{(x_0x_k, I) \in f(b_i, b_{-i} \mid h^1)} u_i(x_k) \cdot q_{ik} / \sum_{(x_0x_k, I) \in f(b_i, b_{-i} \mid h^1)} q_{ik}$, by Lemma 2 we have $U_i(h^1)(b_i^{'}, b_{-i}, q_i) > U_i(h^1)(b_i, b_{-i}, q_i)$.
Hence, $U_i(h^1)(b_i^{'}, c_i, q_i) > U_i(h^1)(b_i, c_i, q_i)$ for all $c_i \in \prod_{j \neq i} \Delta^0(B_j)$ and all $q_i \in Q_i$. ■

We introduce some additional notations. Given $b_i \in B_i$, let $K_i = \#\{k \mid b_{ik} = 1\} \leq N$, $e(k)$ is the individual behavior such that the kth component is 1 and the other components are 0, and $\mathbf{1}$ is the unit vector, that is, the behavior where the individual agrees to join every move. Lemma 4 establishes that there exists a conjecture c_i and an implementability belief q_i such that any behavior $b_i \neq \mathbf{1}$ where individual i agrees to move to x_N is her unique best response. This conjecture is such that it puts weight on $b_j = e(k)$ whenever $b_{ik} = 1$ and on $b_j = \mathbf{1}$. The former part of the conjecture guarantees that b_i gives higher utility than $b_i' \neq b_i$ whenever b_i' blocks moves that are not blocked by b_i. The latter part, together with a suitably chosen implementability prior-belief, implies that b_i outperforms any b_i' that agrees to strictly more moves than b_i.

Lemma 4 *Take any $b_i \in B_i \setminus \{\mathbf{1}\}$ such that $b_{iN} = 1$. Then, for all $b_i^{'} \in B_i$ ($b_i^{'} \neq b_i$), we have $U_i(h^1)(b_i, c_i, q_i) > U_i(h^1)(b_i^{'}, c_i, q_i)$, where $c_i(h^1) \in \prod_{j \neq i} \Delta(B_j)$ is such that*

$$c_i^j(h^1)(b_j) = \begin{cases} u_i(x_N) \cdot [K_i \cdot u_i(x_N) + u_i(x_1)]^{-1} & \text{if } b_j = e(k) \text{ and } b_{ik} = 1 \\ u_i(x_1) \cdot [K_i \cdot u_i(x_N) + u_i(x_1)]^{-1} & \text{if } b_j = \mathbf{1} \\ 0 & \text{otherwise} \end{cases}$$

and $q_i \in Q_i$ is such that

$$q_{ik} = \begin{cases} \varepsilon & \text{if } k = N \\ \varepsilon^2 & \text{if } b_{ik} = 1 \text{ and } k \neq N \\ (1 - \varepsilon - \sum_{(x_0 x_k, I) \mid b_{ik} = 1, k \neq N} \varepsilon^2) \cdot [\#\{(x_0 x_k, I) \mid b_{ik} = 0\}]^{-1} & \text{if } b_{ik} = 0 \end{cases}$$

with $0 < \varepsilon \leq (u_i(x_N) - u_i(x_{N-1})) \cdot [N \cdot u_i(x_{N-1})]^{-1}$.

Proof. Let $p(\overline{f}(b_{-i} \mid h^1))$ be the probability the opponents of individual i agree on $\overline{f}(b_{-i} \mid h^1)$. Notice that $\overline{f}(b_{-i} \mid h^1)$ could be empty. Then,

$$\begin{aligned} p(M(h^1)) &= \prod_{j \neq i} \frac{u_i(x_1)}{K_i \cdot u_i(x_N) + u_i(x_1)}, \\ p(\overline{f}(b_{-i} \,|\, h^1)) &= 0 \text{ if } \#\overline{f}(b_{-i} \mid h^1) \geq 2 \text{ and } \overline{f}(b_{-i} \mid h^1) \neq M(h^1), \\ p(\{(x_0 x_k, I)\}) &= \prod_{j \neq i} \frac{u_i(x_N) + u_i(x_1)}{K_i \cdot u_i(x_N) + u_i(x_1)} - \prod_{j \neq i} \frac{u_i(x_1)}{K_i \cdot u_i(x_N) + u_i(x_1)} \text{ if } b_{ik} = 1, \end{aligned}$$

and $p(\{(x_0 x_0, \emptyset\})$ is the remainder. The probability $p(\{(x_0 x_k, I)\})$ follows from the observation that the agreement set is $\{(x_0 x_k, I)\}$ if and only if all opponents of i choose $e(k)$ or $\mathbf{1}$, and not all of them choose $\mathbf{1}$. Then,

$$\begin{aligned} U_i(h^1)(b_i, c_i, q_i) &= \sum_{(x_0 x_k, I) \mid b_{ik} = 1} p(\{(x_0 x_k, I)\}) \cdot u_i(x_k) \\ &\quad + \left[\frac{u_i(x_1)}{K_i \cdot u_i(x_N) + u_i(x_1)} \right]^{\#I - 1} \left[\frac{\sum_{(x_0 x_k, I) \mid b_{ik} = 1} q_{ik} \cdot u_i(x_k)}{\sum_{(x_0 x_k, I) \mid b_{ik} = 1} q_{ik}} \right]. \end{aligned}$$

Two cases have to be considered. In Case 1 we consider $b_i^{'}$ such that, for some k, $b_{ik} = 1$ and $b_{ik}^{'} = 0$. In Case 2 we take $b_i^{'} \neq b_i$ such that $b_{ik} = 1$ implies $b_{ik}^{'} = 1$.

Case 1. Since $p(\{(x_0x_k, I)\}) = 0$ if $b_{ik} = 0$, and there is k such that $b_{ik} = 1$ and $b_{ik}^{'} = 0$, it follows that

$$\begin{aligned} U_i(h^1)(b_i^{'}, c_i, q_i) \quad \leq \quad & \sum_{(x_0x_k,I)|b_{ik}=1} p(\{(x_0x_k, I)\}) \cdot u_i(x_k) - p(\{(x_0x_1, I)\}) \cdot u_i(x_1) \\ & + \left[\frac{u_i(x_1)}{K_i \cdot u_i(x_N) + u_i(x_1)}\right]^{\#I-1} \left[\frac{\sum\limits_{(x_0x_k,I)|b_{ik}^{'}=1} q_{ik} \cdot u_i(x_k)}{\sum\limits_{(x_0x_k,I)|b_{ik}^{'}=1} q_{ik}}\right]. \end{aligned}$$

Since

$$\left[\frac{u_i(x_1)}{K_i \cdot u_i(x_N) + u_i(x_1)}\right]^{\#I-1} \left[\frac{\sum\limits_{(x_0x_k,I)|b_{ik}^{'}=1} q_{ik} \cdot u_i(x_k)}{\sum\limits_{(x_0x_k,I)|b_{ik}^{'}=1} q_{ik}}\right] \leq \left[\frac{u_i(x_1)}{K_i \cdot u_i(x_N) + u_i(x_1)}\right]^{\#I-1} \cdot u_i(x_N),$$

and

$$\left[\frac{u_i(x_1)}{K_i \cdot u_i(x_N) + u_i(x_1)}\right]^{\#I-1} \cdot u_i(x_N) < \frac{(u_i(x_N) + u_i(x_1))^{\#I-1} - (u_i(x_1))^{\#I-1}}{(K_i \cdot u_i(x_N) + u_i(x_1))^{\#I-1}} \cdot u_i(x_1)$$

which equals $p(\{(x_0x_1, I)\}) \cdot u_i(x_1)$, it follows that $U_i(h^1)(b_i^{'}, c_i, q_i) < \sum_{(x_0x_k,I)|b_{ik}=1} p(\{(x_0x_k, I)\}) \cdot u_i(x_k)$. Hence, $U_i(h^1)(b_i, c_i, q_i) > U_i(h^1)(b_i^{'}, c_i, q_i)$.

Case 2. It holds that

$$\begin{aligned} U_i(h^1)(b_i^{'}, c_i, q_i) \quad = \quad & \sum_{(x_0x_k,I)|b_{ik}=1} p(\{(x_0x_k, I)\}) \cdot u_i(x_k) + \left[\frac{u_i(x_1)}{K_i \cdot u_i(x_N) + u_i(x_1)}\right]^{\#I-1} \\ & \cdot \left[\frac{\sum\limits_{(x_0x_k,I)|b_{ik}=1} q_{ik} \cdot u_i(x_k) + \sum\limits_{(x_0x_k,I)|b_{ik}=0,b_{ik}^{'}=1} q_{ik} \cdot u_i(x_k)}{\sum\limits_{(x_0x_k,I)|b_{ik}=1} q_{ik} + \sum\limits_{(x_0x_k,I)|b_{ik}=0,b_{ik}^{'}=1} q_{ik}}\right]. \end{aligned}$$

Since $b_{ik}^{'} = 1$ while $b_{ik} = 0$ for some $k \leq N-1$ (and $b_{iN} = 1$), we have

$$(\frac{y_3}{y_4} =) \qquad \frac{\sum\limits_{(x_0x_k,I)|b_{ik}=0,b_{ik}^{'}=1} q_{ik} \cdot u_i(x_k)}{\sum\limits_{(x_0x_k,I)|b_{ik}=0,b_{ik}^{'}=1} q_{ik}} \leq u_i(x_{N-1}).$$

Also, notice that

$$(\frac{y_1}{y_2} =) \qquad \frac{\sum\limits_{(x_0x_k,I)|b_{ik}=1} q_{ik} \cdot u_i(x_k)}{\sum\limits_{(x_0x_k,I)|b_{ik}=1} q_{ik}} = \frac{\sum\limits_{(x_0x_k,I)|b_{ik}=1,k\neq N} \varepsilon^2 \cdot u_i(x_k) + \varepsilon \cdot u_i(x_N)}{\sum\limits_{(x_0x_k,I)|b_{ik}=1,k\neq N} \varepsilon^2 + \varepsilon} > \frac{\varepsilon u_i(x_N)}{N\varepsilon^2 + \varepsilon} = \frac{u_i(x_N)}{N\varepsilon + 1}.$$

Since, by definition of ε, $0 < \varepsilon \leq (u_i(x_N) - u_i(x_{N-1})) \cdot [N \cdot u_i(x_{N-1})]^{-1}$, we have

$$\frac{y_1}{y_2} > \frac{u_i(x_N)}{N\varepsilon + 1} \geq u_i(x_{N-1}) \geq \frac{y_3}{y_4}.$$

Hence, by Lemma 2 it follows that $U_i(h^1)(b_i, c_i, q_i) > U_i(h^1)(b_i^{'}, c_i, q_i)$. ■

The next lemma shows that the behavior where i agrees to join every move is individually rational.

Lemma 5 *Take $b_i \in B_i$ such that $b_{ik} = 1$, $k = 1, ..., N$. Then, for all $b_i^{'} \in B_i$ $(b_i^{'} \neq b_i)$ and for all $q_i \in Q_i$, we have $U_i(h^1)(b_i, c_i, q_i) > U_i(h^1)(b_i^{'}, c_i, q_i)$, where $c_i(h^1) \in \prod_{j \neq i} \Delta(B_j)$ is such that*

$$c_i^j(h^1)(b_j) = \begin{cases} \frac{1}{N} & \text{if } b_j = e(k),\ k = 1, ..., N \\ 0 & \text{otherwise} \end{cases}.$$

Proof. For $b_i = \mathbf{1}$, we have

$$\begin{aligned} U_i(h^1)(b_i^{'}, c_i, q_i) &\leq \sum_{(x_0 x_k, I) | b_{ik} = 1} u_i(x_k) \cdot \left[\frac{1}{N}\right]^{\#I-1} - u_i(x_1) \cdot \left[\frac{1}{N}\right]^{\#I-1} \\ &< \sum_{(x_0 x_k, I) | b_{ik} = 1} u_i(x_k) \cdot \left[\frac{1}{N}\right]^{\#I-1} = U_i(h^1)(b_i, c_i, q_i) \quad \forall b_i^{'} \neq b_i,\ \forall q_i \in Q_i. \end{aligned}$$

■

Putting these results together, we are able to show the following main result.

Theorem 3 *Consider the social environment Γ^*. There is a unique behavior of individual i that is socially rationalizable, $P_i^\infty = \{e(N)\}$, $i \in I$.*

Proof. By Definition 5, $P_i^0 = B_i$ and $P^0 = \prod_{i \in I} B_i$. In the first iteration, by Lemma 3, all $b_i \in P_i^0$ such that $b_{iN} = 0$ do not belong to P_i^1, $i \in I$. By Lemma 1, Lemma 4 and Lemma 5, all b_i such that $b_{iN} = 1$ belong to P_i^1, $i \in I$. So, $P_i^1 = \{b_i \mid b_{iN} = 1\}$ $i \in I$.

In the second iteration, for all $c_i(h^1) \in \prod_{j \neq i} \Delta^0(P_j^1)$ and for all $q_i \in Q_i$, the behavior b_i such that $b_{iN} = 1$ and $b_{ik} = 0$ if $k \neq N$ gives to individual i a utility $U_i(h^1)(b_i, c_i, q_i) = u_i(x_N)$. However, for all $b_i^{'} \in P_i^1 \setminus \{b_i\}$, $U_i(h^1)(b_i^{'}, c_i, q_i) < u_i(x_N)$ for all c_i and all q_i, because for some $k < N$, $b_{ik}^{'} = 1$, and the cautiousness of c_i implies that with positive probability the opponents of i have an agreement set $\{(x_0 x_k, I)\}$, which leads to utility $u_i(x_k) < u_i(x_N)$. So, $P_i^2 = \{e(N)\} = P_i^\infty$, $i \in I$. ■

The above result implies that social rationalizability satisfies the property of coalitional rationality. When the outcomes can be Pareto ranked, a coalition selects the Pareto-dominant

outcome. Each individual only agrees to move to the Pareto dominating outcome, and blocks all other moves.

Corollary 1 *Consider the social environment* Γ^*. *We have* $Z^\infty(x_0) = \{x_N\}$.

5 Conclusion

Social environments constitute a framework in which it is possible to study how groups of agents interact in a society. We have argued for the need of a new solution concept for social environments that is based on individual rationality, called social rationalizability. One of the basic steps in our construction is to model individual behavior in a social environment, which makes a social environment apt to an analysis based on individual rationality. Individual behavior within a coalition is modeled as the decision to agree to a coalitional move or to block it. Since a coalition may have several moves available, and more than one coalition may have the option to move at the same time, there can be many moves on which there is agreement. Individuals therefore also form beliefs on which move in the set of moves on which there is agreement will be carried out.

Social rationalizability identifies which coalitions are likely to form and which outcomes might occur when (1) the individuals are rational and endowed with a hierarchy of hypotheses, and (2) this is common knowledge at the original status-quo. We have shown that for all social environments the set of socially rationalizable outcomes is non-empty. The computation of the set of socially rationalizable outcomes is greatly simplified by using a reduction procedure, which we show to be equivalent to the formal definition of social rationalizability.

Social rationalizability aims to be a weak concept that rules out with confidence. Its non-emptiness makes it applicable to cases where traditional solution concepts fail to make predictions. It is also not too weak in the sense that it satisfies individual rationality. As a theory of social behavior, social rationalizability should also be consistent with elementary notions of coalitional rationality. For instance, when a coalition has to choose between a number of Pareto ranked moves, it should select the Pareto dominating one for sure. It is shown that social rationalizability is consistent with coalitional rationality.

References

[1] Battigalli, P., "On Rationalizability in Extensive Games," *Journal of Economic Theory* 74 (1997), 40-61.

[2] Bernheim, D., "Rationalizable Strategic Behavior," *Econometrica* 52 (1984), 1007-1028.

[3] Bloch, F., "Sequential Formation of Coalitions in Games with Externalities and Fixed Payoff Division," *Games and Economic Behavior* 14 (1996), 90-123.

[4] Chwe, M.S., "Farsighted Coalitional Stability," *Journal of Economic Theory* 63 (1994), 299-325.

[5] Greenberg, J., *The Theory of Social Situations: An Alternative Game-Theoretic Approach*, Cambridge University Press, 1990.

[6] Herings, P.J.J. and V.J. Vannetelbosch, "Refinements of Rationalizability for Normal-Form Games," *International Journal of Game Theory* 28 (1999), 53-68.

[7] Mariotti, M., "A Model of Agreements in Strategic Form Games," *Journal of Economic Theory* 74 (1997), 196-217.

[8] Pearce, D.G., "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica* 52 (1984), 1029-1050.

[9] Rosenthal, R., "Cooperative Games in Effectiveness Form," *Journal of Economic Theory* 5 (1972), 88-101.

[10] Shimoji, M. and J. Watson, "Conditional Dominance, Rationalizability, and Game Forms," *Journal of Economic Theory* 83 (1998), 161-195.

[11] Vannetelbosch, V.J., "Rationalizability and Equilibrium in N-Person Sequential Bargaining," *Economic Theory* 14 (1999), 353-371.

[12] Vannetelbosch, V.J., "Alternating-Offer Bargaining and Common Knowledge of Rationality," *Theory and Decision* 47 (1999), 111-137.

[13] Von Neumann, J. and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press (1944).

[14] Xue, L., "Coalitional Stability under Perfect Foresight," *Economic Theory* 11 (1998), 603-627.