Rationalizability for social environments

P. Jean-Jacques Herings [a,*], Ana Mauleon [b],
Vincent J. Vannetelbosch [c]

[a] *Department of Economics, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands*
[b] *LABORES (URA 362, CNRS), Université Catholique de Lille, Boulevard Vauban 60, BP 109, 59016 Lille,
France*
[c] *FNRS, CORE and IRES, Université Catholique de Louvain, Voie du Roman Pays 34, 1348 Louvain-la-Neuve,
Belgium*

Abstract

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 analyze both non-cooperative and cooperative games. In order to remedy the shortcomings of existing solution concepts and to identify the consequences of common knowledge of rationality and farsightedness, we propose to apply extensive-form rationalizability to the framework of social environments. For us, the social environment is a primitive. On this social environment is defined a multistage game. An outcome of the social environment is socially rationalizable if and only if it is rationalizable in the multistage game. 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.
© 2004 Elsevier Inc. All rights reserved.

* Corresponding author.
 E-mail addresses: p.herings@algec.unimaas.nl (P.J.J. Herings), mauleon@ires.ucl.ac.be (A. Mauleon),
vannetelbosch@core.ucl.ac.be (V.J. Vannetelbosch).

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 by Chwe (1994) 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 (1990). Moreover, this framework allows us to study economic and social activities where the rules of the game are rather amorphous or the procedures are rarely pinned down (e.g. in sequential bargaining or coalition formation without a rigid protocol), and for which classical game theory could lead to a solution which relies heavily on an arbitrarily chosen procedure or rule (see Greenberg, 1990; Greenberg et al., 1996). For social environments where coalitions can form through binding or non-binding agreements and actions are public, Chwe (1994) and Xue (1998) 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 (1996).[2] No commitment assumption is imposed. Secondly, it incorporates the farsightedness of the coalitions.[3] 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 (1994) and Xue (1998) 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.[4]

---

[1] For a very specific social environment, namely the coalitional contingent threat situation, Mariotti (1997) 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] Other literature on endogenous coalition or network formation has examined farsightedness issues; see e.g. Aumann and Myerson (1988) who have developed an extensive-form game of links formation, and Ray and Vohra (1999) who have defined on a partition function an extensive-form bargaining game.

[4] In Chwe (1994), 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 (1998), it is formalized

Both approaches suffer from a number of drawbacks as well, some of them pointed out by the authors themselves. For instance, as indicated by Chwe (1994), 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 by Xue (1998) 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.

In order to remedy the problems mentioned above and to identify the consequences of common knowledge of rationality and farsightedness within the framework of social environments, we propose to apply extensive-form rationalizability to the framework of social environments. Extensive-form rationalizability, which has been first introduced by Pearce (1984), is a solution concept relying on the following assumptions:

  (i) each player always maximizes her expected payoff and she updates her expectations using Bayes' rule whenever possible;
 (ii) each player believes that her opponents are rational unless this is contradicted by their observed behaviors;
(iii) assumptions (i) and (ii) are common knowledge at the beginning of the game.

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 defining on a social environment a multi-stage game with observed actions[5] and 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 noncooperative 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.

---

by means of the theory of social situations developed by Greenberg (1990). A social situation allows for capturing perfect foresight (which strengthens farsightedness) by extending the Von Neumann and Morgenstern (1944) notion of stability to accommodate different behavior on the part of the individuals in terms of their Knightian (pessimism or optimism) attitude towards uncertainty.

[5] In this respect, our paper is related to the literature on non-cooperative foundations of cooperative solution concepts. For instance, Perry and Reny (1994) have developed a dynamic bargaining game of coalition formation which implements the core.

Two equivalent definitions of extensive-form rationalizability have been proposed in the literature.[6] The first one is Pearce's (1984) original extensive-form rationalizability and is based on a reduction procedure. The second one is due to Battigalli (1997). Battigalli's extensive-form rationalizability 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.

Although normal-form rationalizability is known to be a weak concept, extensive-form rationalizability incorporates elements of forward induction and has therefore quite some cutting power.[7] We will use a cautious version of extensive-form rationalizability.[8] Cautious rationalizability in the extensive form is appropriate if individuals are known to be not merely rational but also cautious (i.e. individuals will exercise prudence when it is costless to do so). Cautiousness is needed to eliminate the taking of risks that seem likely to be costly when there are no offsetting advantages for an individual to consider.

We take the social environment as a primitive. Associated to each social environment, we define a multistage game. An individual strategy describes, for each history, the coalitional moves the individual agrees to join and those she decides to block. It may happen that the individuals agree on more than one move. In this case it is the choice of a dummy player that will determine which move among the ones the individuals have agreed upon will be implemented. Central to our new concept is that individuals hold cautious conjectures about their opponents' strategies as well as about the choice of the dummy player, which reverts to hold beliefs about which agreement is realized within the set of agreements. An outcome of the social environment is said to be socially rationalizable if and only if it is supported by extensive-form rationalizability in the multi-stage game.

Our main results are the following. We show how to apply rationalizability in the extensive form to social environments, which is not straightforward since social environments are outside the realm of noncooperative games. By doing so, we are the first to provide a solution concept for social environments that leads to a non-empty set of stable outcomes that is consistent with individual rationality. Since social environments

---

[6] Among the papers related to extensive-form rationalizability are Bernheim (1984), who introduced subgame-perfect rationalizability, and Shimoji and Watson (1998), who studied the equivalence between conditional dominance and extensive-form rationalizability. See Bernheim (1984), Pearce (1984), Herings and Vannetelbosch (1999) for the definitions of rationalizability for normal-form games and of its refinements.

[7] Forward induction is the principle that a player, when trying to predict the future behavior of another player, should take into account the observed past behavior of this player, and, if possible, should base her predictions on strategies of the other player that are rational, and that prescribe the observed past choices of this player. See Battigalli (1996) for more details.

[8] in the definition of cautious rationalizability, a strategy of a player is said to be a cautious response if it is a best response against a completely mixed strategy combination. our version of cautious rationalizability in the extensive form is different from the one proposed by Pearce (1984). In Pearce's definition, cautious rationalizable strategies are obtained by eliminating strategies that are not best responses first, next those that are not cautious responses, then the ones that are not best responses, and so on. If one carries the logic behind cautious rationalizability one step further, one would like to consider a solution concept where players eliminate responses that are not cautious in each round, which leads to the concept of extensive-form rationalizability we use in this paper.

deal with coalitional moves, it is important that social rationalizability not only guarantees individual rationality, but also coalitional rationality. Coalitional rationality specifies that among a set of alternatives a coalition should be able to coordinate on the Pareto optimal one. Social rationalizability is shown to satisfy coalitional rationality.

The paper has been organized as follows. In Section 2 we introduce some notations and primitives. We present the solution concepts of Chwe (1994) and Xue (1998), and we give the motivation for introducing a new concept by means of a number of examples. In Section 3 we define on the social environment a multi-stage game and we apply extensive-form rationalizability to this multi-stage game. The examples are reconsidered and solved. An outcome of the social environment is socially rationalizable if and only if it is rationalizable in the multi-stage game. In Section 4 we study the property of coalitional rationality and show it is satisfied by social rationalizability. Finally, Section 5 concludes.

2. Social environments

2.1. Notations and primitives

As in Chwe (1994) and Xue (1998), 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 $I = \{1, 2, \ldots, \#I\}$ 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 $u_i : Z \rightarrow \mathbb{R}$ specifies the utility function of individual $i \in I$. 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, or on the time it takes to reach a particular end outcome. Conventional game theoretic situations can be modeled as a social environment (see for instance Chwe, 1994).

For social environments where coalitions can form through binding or non-binding agreements and actions are public, Chwe (1994) and Xue (1998) 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 (1994) 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. 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 $x_0, x_1, \ldots, x_m$, where $x_0 = x$ and $x_m = y$, and a sequence $S_0, S_1, \ldots, S_{m-1}$ such that $x_j \to_{S_j} x_{j+1}$ and $u_i(x_j) < u_i(y) \ \forall i \in S_j$, for $j = 0, 1, \ldots, m - 1$. 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 $x \to_S y$ 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 $x \to_S y$, $\exists z \in Y$, 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 (1994) 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 $(k = 1, 2, \ldots)$ is inductively obtained as follows: $x \in Z$ belongs to Y^k if and only if $\forall y, S$ such that $x \to_S y$, $\exists z \in Y^{k-1}$, 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 $\bigcap_{k \geqslant 1} Y^k$.

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 (1998). Some notations and definitions have to be introduced. A path is a sequence $(x_0, x_1, \ldots, x_m)$ where for all $j = 0, 1, \ldots, m - 1$, there exists a coalition $S_j \subseteq I$ such that $x_j \to_{S_j} x_{j+1}$ 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 (1998) defined a standard of behavior as a function $\sigma : Z \to 2^{\Pi}$ such that $\sigma(x) \subseteq \Pi_x$ for all $x \in Z$. A standard of behavior σ is said to be *internally stable* if for all outcomes $x \in Z$ and for all $\alpha \in \sigma(x)$, there do not exist $y \in \alpha$, $S \subseteq I$, and $z \in Z$ such that $y \to_S z$ and S "prefers" $\sigma(z)$ to α. A standard of behavior σ is said to be *externally stable* if for all outcomes $x \in Z$ and for all $\alpha \in \Pi_x \setminus \sigma(x)$, there exist $y \in \alpha$, $S \subseteq I$ and $z \in Z$ such that $y \to_S z$ and S "prefers" $\sigma(z)$ to α. A standard of behavior σ is stable if it is both internally and externally stable.

As in Greenberg (1990), Xue (1998) distinguished an optimistic and a conservative approach to define "prefers." In the optimistic approach a coalition S prefers $\sigma(z)$ to α if $\exists \beta \in \sigma(z)$, $u_i(\alpha) < u_i(\beta) \ \forall i \in S$.[9] 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

---

[9] We define the utility of a path α as the utility of the end outcome of α.

"prefers" is defined by the optimistic approach. A CSSB is a stable standard of behavior, where "prefers" is defined by the conservative approach.

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

(i) σ is an OSSB if $\forall x \in Z$, $\alpha \in \Pi_x \setminus \sigma(x) \iff \exists S \subseteq I$, $y \in \alpha$, and $z \in Z$ such that $y \to_S z$ and $\exists \beta \in \sigma(z)$: $u_i(\alpha) < u_i(\beta) \; \forall i \in S$.
(ii) σ is a CSSB if $\forall x \in Z$, $\alpha \in \Pi_x \setminus \sigma(x) \iff \exists S \subseteq I$, $y \in \alpha$, and $z \in Z$ such that $y \to_S z$ 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 (1994) 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. Therefore, 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 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 Fig. 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.[10]

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 Fig. 2, there are three individuals that have the 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
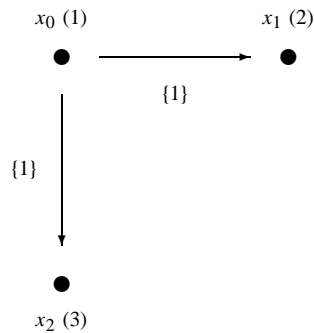

Fig. 1. Individual rationality.

---

[10] Two other problems have also been mentioned by Chwe (1994). First, the LCS does not incorporate the decision of subcoalitions to veto coalitional moves. second, a coalition considers what further moves other coalitions will 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 (1998) concepts) overcomes these problems.
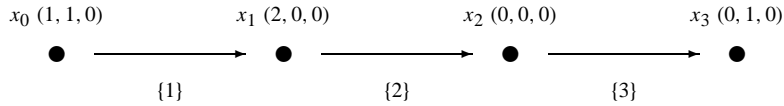
$x_0 \, (1, 1, 0)$ $x_1 \, (2, 0, 0)$ $x_2 \, (0, 0, 0)$ $x_3 \, (0, 1, 0)$

● $\longrightarrow$ ● $\longrightarrow$ ● $\longrightarrow$ ●

$\{1\}$ $\{2\}$ $\{3\}$

Fig. 2. LCS may rule out too much.

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 Fig. 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 (1998). 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 Fig. 2. But even though a CSSB is typically a very weak concept, it may also rule out too much. In the social environment of Fig. 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 (1998). Although a unique CSSB exists, it is empty-valued for some status quos. A standard of 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 Fig. 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 Fig. 4, the only sensible standard of behavior is $\sigma(x_0) = \{(x_0)\}$. Nevertheless, the standard of behavior $\sigma(x_0) = \{(x_0), (x_0, x_1), (x_0, x_2)\}$, $\sigma(x_1) = \{(x_1)\}$ and $\sigma(x_2) = \{(x_2)\}$ is both the unique CSSB and the unique OSSB. It may look like this phenom-
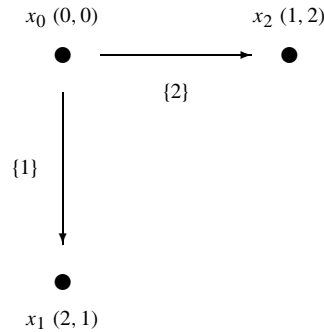
$x_0\ (0,0)$ $x_2\ (1,2)$

{2}

{1}

$x_1\ (2,1)$

Fig. 3. CSSB and OSSB may rule out too much.

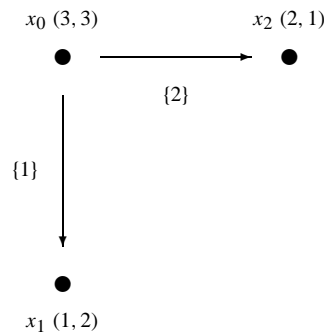$x_0\ (3,3)$ $x_2\ (2,1)$

{2}

{1}

$x_1\ (1,2)$

Fig. 4. OSSB and CSSB may rule out too little.

enon is caused by the absence of the no-move. But even if we add moves $x_0 \rightarrow_{\{1\}} x_0$, $x_0 \rightarrow_{\{2\}} x_0$, $x_0 \rightarrow_{\{1,2\}} x_0$, then the standard of behavior defined by $\sigma(x_0) = \{(x_0), (x_0, x_1), (x_0, x_2), (x_0, x_0), (x_0, x_0, x_1), (x_0, x_0, x_2), (x_0, x_0, x_0), \ldots\}$, $\sigma(x_1) = \{(x_1)\}$ and $\sigma(x_2) = \{(x_2)\}$ is a CSSB. OSSB seems to do better now, as the unique OSSB is given by $\sigma(x_0) = \{(x_0), (x_0, x_0), (x_0, x_0, x_0), \ldots\}$, $\sigma(x_1) = \{(x_1)\}$, 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. Social rationalizability

We denote by $(x \rightarrow_S y)$ the move from x to y by coalition S. The no-move at status quo x is denoted by $(x \rightarrow_\emptyset x)$. One has to distinguish between $(x \rightarrow_\emptyset x)$ and $(x \rightarrow_{\{i\}} x)$.

Indeed, $(x \to_{\{i\}} x)$ means that individual i can move from x to x. The set of all possible moves and no-move is given by $M = \{(x \to_S y) \mid x, y \in Z, x \to_S y\} \cup \{(x \to_\emptyset x) \mid x \in Z\}$. An original status quo is given, and it is denoted x_0. To determine which coalitional moves are going to be implemented and which outcomes are rationalizable in a social environment Γ, we define on Γ a multistage game with finite horizon and with observed actions, $G(\Gamma)$.

The multistage game starts at status quo x_0. In the first stage all players $i \in I$ choose simultaneously the coalitional moves from x_0 they agree to join and those they decide 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). It may also happen that the individuals agree on more than one move. In such a circumstance it is the action chosen by a dummy player that will determine which move among the ones the individuals have agreed upon will be implemented. In case all moves are blocked, then the game ends at the current status quo. Otherwise, a move to a new status quo is implemented. In case coalitional moves are possible from the new status quo then the game proceeds to the second stage. Otherwise, it ends at the new status quo. In the second stage all players again choose simultaneously the coalitional moves from the new status quo they agree to join and those they decide to block; and so on. The game ends when each possible move after a certain stage is blocked, or when the game has reached its final stage.

We give now a precise definition of the multi-stage game $G(\Gamma)$ defined on the social environment Γ. We denote by $I^+ = \{0, 1, \ldots, n\}$ the original set of players together with a dummy player, denoted by player 0. Individual i's opponents are denoted by $-i$. We let $h^0 = \emptyset$ be the history at the start of the play and $x_0 \in Z$ the original status quo. At the end of each stage, all players observe the stage's action profile. For $k = 0, 1, \ldots$, we denote by h^k the history at the beginning of stage k. It consists of the actions taken by the players at all the previous stages. At stage k, $k = 0, 1, \ldots$, of $G(\Gamma)$, all players $i \in I^+$ simultaneously choose actions. Let $M(h^k) = \{(x \to_S y) \in M \mid z(h^k) = x\}$ be the set of feasible moves after history h^k, where $z(h^k)$ is the current status quo after history h^k. Let $M_i(h^k) = \{(x \to_S y) \in M(h^k) \mid i \in S\}$ be the set of feasible moves after history h^k involving individual i. An action of player $i \in I$ is a mapping $a_i^k : M_i(h^k) \to \{0, 1\}$. If $a_i^k((x \to_S y)) = 1$ then $i \in S$ agrees to join in the potential move of coalition S from x to y. If $a_i^k((x \to_S y)) = 0$ then $i \in S$ blocks the move of coalition S from x to y. An action a_0^k of the dummy player 0 is a permutation of $M(h^k)$ that indicates the order according to which moves are implemented. More precisely, let $f(a_{-0}^k) = \{(x \to_S y) \in M(h^k) \mid \forall i \in S, a_i^k((x \to_S y)) = 1\}$ be the set of moves on which all players agree. If $f(a_{-0}^k)$ is non-empty then the first element of a_0^k contained in $f(a_{-0}^k)$ is implemented. Let $A_i(h^k)$ be the set of actions for $i \in I^+$ after history h^k, and let $a^k \equiv (a_0^k, a_1^k, \ldots, a_n^k)$ be the stage-k action profile, where $a_i^k \in A_i(h^k)$, $i \in I^+$. Then the history at the beginning of stage $k+1$ is given by $h^{k+1} = (a^0, a^1, \ldots, a^k)$.

In social environments, negotiations on outcomes have to stop once in order to realize payoffs. We model this by assuming that the game has to stop for sure after K periods, and consider all outcomes that could be an end outcome when K approaches infinity. The total number of stages in the game is therefore given by $K + 1$. The set H^k denotes the set of all stage-k histories, H the set of all possible non-terminal histories, $H = \bigcup_{k=0}^{K-1} H^k$, and $H^* = \bigcup_{k=0}^{K} H^k$ the set of all possible histories. Let $z : H^* \to Z$ be a function that gives for

each possible history the current status quo; $z(h^k)$ is the status quo after history h^k. If the game ends at $z(h^k)$, then the payoff of player $i \in I$ is $u_i(z(h^k))$. The payoff of the dummy player is defined to be always equal to zero, irrespective of the outcome of the game.

We denote by H_i the set of histories after which player $i \in I$ has a move, $H_i = \{h \in H \mid M_i(h) \neq \emptyset\}$. A pure strategy of player $i \in I$ is a mapping $s_i : H_i \to A_i$, where $A_i = \bigcup_{h \in H_i} A_i(h)$. A pure strategy of player 0 is a mapping $s_0 : H \to A_0$, where $A_0 = \bigcup_{h \in H} A_0(h)$. Let S_i be the set of pure strategies of player i. Let $S = \prod_{i \in I^+} S_i$ be the set of strategy profiles.

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. 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[11] on the behaviors of their opponents, $c_i : H_i \to \prod_{j \in I^+ \setminus \{i\}} \Delta(S_j)$, with $c_i(h')(s_{-i})$ being the probability individual $i \in I^+$ conjectures at history h' that her opponents strategies are s_{-i}. We denote by $U_i(s_i, c_i)$ the expected payoff given (s_i, c_i). A conjecture c_i allows for a history $h \in H_i$ if there are a strategy s_i and a profile s_{-i} in the support of c_i such that the path induced by (s_i, s_{-i}) generates h. A strategy combination $(s_i)_{i \in \widehat{I}, \widehat{I} \subseteq I^+}$ allows for h if there is a profile $(s_j)_{j \notin \widehat{I}}$ such that the path induced by $(s_i)_{i \in I^+}$ generates h. A set $\widehat{S}_{-i} \subseteq S_{-i}$ allows for h if there is a profile $s_{-i} \in \widehat{S}_{-i}$ which allows for h. At each history h' after which individual i is involved in a move, she uses her conjecture at h' to determine the likelihood of future histories by means of Bayesian updating. Given conjectures and strategies, individuals can compute their expected utility. Notice that a conjecture may change as the course of the social environment unfolds, and that there is only a need for an individual to form conjectures when an individual is potentially involved in a move.

Two alternative definitions of extensive-form rationalizability, which are equivalent (see Theorem 1 in Battigalli, 1997), have been proposed in the literature. The first one is Pearce's (1984) original extensive-form rationalizability and is based on a reduction procedure. The second one is Battigalli's (1997) extensive-form rationalizability and is based on the notion of a hierarchy of nested hypotheses.[12] The definitions we present next deviate from the original ones because we assume that players hold *cautious* conjectures. Battigalli's definition of extensive-form rationalizability 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.

---

[11] The analysis where individuals hold correlated conjectures about the behaviors of their opponents is fully parallel and does not create any difficulties.

[12] Pearce's (1984) extensive-form rationalizability, like most extensive-form theories, does not adequately deal with counterfactuals and strategic manipulations of conjectures. Battigalli (1997) overcomes such drawbacks by providing an alternative characterization of extensive-form rationalizability 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 (1984) iterative deletion procedure.

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. Given two histories $g, h \in H$, we write $g < h$ if g is a strict subhistory of h.

Definition 3. A consistent updating system for individual $i \in I^+$ is a mapping $c_i : H_i \to \prod_{j \in I^+ \setminus \{i\}} \Delta(S_j)$ such that for all $g, h \in H_i$:

 (i) $c_i(h)$ allows for h,
 (ii) if $g < h$ and $c_i(g)$ allows for 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.[13] That is, individuals update according to Bayes rule whenever possible. A strategy s_i is individually rational if it is a best response to some cautious consistent updating system c_i. In Definition 4, R_i^1 is the set of strategies of i that are individually rational. Higher degrees of rationality are constructed recursively.

Definition 4. Let $R^0 = \prod_{i \in I^+} S_i$. For $n \geqslant 1$, $R^n = \prod_{i \in I^+} R_i^n$ is inductively defined as follows: for all $i \in I^+$, $s_i \in R_i^n$ if there exists a consistent updating system c_i such that

 (i) for all $h' \in H_i$, $c_i(h') \in \prod_{j \in I^+ \setminus \{i\}} \Delta^0(R_j^{k^*})$ where k^* is the maximal element in $\{0, 1, \ldots, n-1\}$ such that $R_{-i}^{k^*}$ allows for h',
 (ii) for all $h' \in H_i$, if s_i allows for h', then s_i is a best response to $c_i(h')$ at h', that is, for all $\hat{s}_i \in S_i$, $U_i(h')(s_i, c_i) \geqslant U_i(h')(s_i/\hat{s}_i^{h'}, c_i)$, where $s_i/\hat{s}_i^{h'}$ is the strategy which results from s_i when behavior at h' and its followers $g > h'$ is specified by $\hat{s}_i$.

The set $R^\infty(K) = \lim_{n \to \infty} R^n$ is the set of rationalizable strategy profiles where the game $G(\Gamma)$ consists of at most $K + 1$ stages.

Definition 4 can be interpreted as follows. The sequence $R_j^1, R_j^2, R_j^3, \ldots$, represents for individual i a hierarchy of increasingly strong hypotheses about the behavior of individual j. When individual i implements a strategy $s_i \in R_i^\infty(K)$, she always holds the strongest hypothesis which is consistent with the history reached (part (i) in Definition 4) and optimizes accordingly. The important distinction to original extensive form rationalizability is that conjectures are cautious.

Notice that since the payoffs of the dummy player are trivial, it is never possible to eliminate any of her actions. For any n it holds that $R_0^n = S_0$. From the perspective of the players in I, the dummy player randomly implements a move on which there is agreement.

---

[13] Battigalli (1996) has shown that the structural consistency condition incorporated into extensive-form rationalizability does not appropriately model strategic independence. Indeed, according to extensive-form rationalizability, player i is allowed to change her conjecture about opponent j, just because she observed a subjectively unexpected behavior by another opponent k, while intuition suggests that in such case i's conjecture about j should not change if strategic independence holds.

The conjectured probability by which a particular move is implemented is not restricted, however, apart from being positive because of cautiousness.

The cautious version of the extensive-form rationalizability concept due to Pearce (1984) is a reduction procedure and is defined as follows.

Definition 5. Let $P^0 = \prod_{i \in I^+} S_i$. For $n \geqslant 1$, $P^n = \prod_{i \in I^+} P_i^n$ is inductively defined as follows: for all $i \in I^+$, $s_i \in P_i^n$ if:

(i) $s_i \in P_i^{n-1}$,
(ii) there exists a consistent updating system c_i such that for all $h' \in H_i$ that are allowed by s_i and P_{-i}^{n-1} it holds:
   (a) $c_i(h') \in \prod_{j \in I^+ \setminus \{i\}} \Delta^0(P_j^{n-1})$,
   (b) for all $\hat{s}_i \in P_i^{n-1}$, $U_i(h')(s_i, c_i) \geqslant U_i(h')(s_i/\hat{s}_i^{h'}, c_i)$.

The set $P^\infty(K) = \lim_{n \to \infty} P^n$ is the set of rationalizable strategy profiles where the game $G(\Gamma)$ consists of at most $K + 1$ stages.

Again, since the payoffs of the dummy player are trivial, it is never possible to eliminate any of her actions. For any n it holds that $P_0^n = S_0$.

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

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

The proof of this theorem is similar to the proof of Theorem 1 in Battigalli (1997), and is therefore omitted. The interested reader is referred to Herings et al. (2000) for details of the proof.

It follows as a corollary to Theorem 1, $R^\infty(K) = P^\infty(K)$.

Given a social environment Γ, we define an outcome $x \in Z$ to be socially rationalizable if it is supported by extensive-form rationalizability in $G(\Gamma)$. We denote by Z_K^∞ the set of rationalizable outcomes. It is given by

$$Z_K^\infty = \big\{ x \in Z \mid \exists (a^0, a^1, \ldots, a^k) \in z^{-1}(\{x\}),$$
$$\exists s \in P^\infty(K) \text{ such that } \forall j = 1, \ldots, k, \; s(a^0, a^1, \ldots, a^{j-1}) = a^j \big\},$$

where $z^{-1}(\{x\})$ are histories leading to the outcome x. The set of socially rationalizable outcomes, Z^∞, is obtained by letting K go to infinity, $Z^\infty = \limsup_{K \to \infty} Z_K^\infty$. It captures the set of outcomes that are stable if the game could go on infinitely long, but stops in finite time with probability one. That is, we consider that the game has to stop for sure after K periods, and consider all outcomes that could be an end outcome when K approaches infinity. This corresponds well to the idea of social environments, where negotiations on outcomes have to stop once, but at an indefinite time period, and there is no discounting of payoffs.

The set of socially rationalizable outcomes is never empty.

Theorem 2. $Z^\infty \neq \emptyset$.

The proof of this theorem, as well as the other proofs not in the main text, may be found in the appendix.

We reconsider the four 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 Fig. 1. The extensive-form game defined on this social environment is depicted in Fig. 5. It is a one-stage game where individual 1 and the dummy player 0 simultaneously choose an action. We have $M_1(h^0) = M(h^0) = \{(x_0 \to_{\{1\}} x_1), (x_0 \to_{\{1\}} x_2)\}$. Any strategy of player 1 is such that $a_1^0((x_0 \to_{\{1\}} x_1))$ equals 1 or 0 and $a_1^0((x_0 \to_{\{1\}} x_2))$ equals 1 or 0. For simplicity, we denote the set of strategies of individual 1 as $S_1 = \{(0, 0), (0, 1), (1, 0), (1, 1)\} = A_1$ where $(0, 1)$ means that $a_1^0((x_0 \to_{\{1\}} x_1)) = 0$ and $a_1^0((x_0 \to_{\{1\}} x_2)) = 1$. Any strategy of the dummy player is a permutation of $M(h^0) = M$. For simplicity, the set of strategies of player 0 is denoted $S_0 = \{(x_1, x_2), (x_2, x_1)\}$ where (x_1, x_2) means that the first element ranked is the move from x_0 to x_1 and the second element ranked is the move from x_0 to x_2. So, the strategy (x_1, x_2) means that the first element of (x_1, x_2) contained in $f(a_1^0)$ will be the outcome of the game. For example, if $a_0^0 = (x_2, x_1)$ and $a_1^0 = (1, 1)$, then $f(a_1^0) = \{(x_0 \to_{\{1\}} x_1), (x_0 \to_{\{1\}} x_2)\}$ and the outcome of the game will be the move to x_2 leading to a payoff of 3 for player 1. By Definition 5, $P^0 = S$. Obviously, the unique best response for individual 1 is her action $(0, 1)$. Against any cautious conjecture on the behavior of the dummy player, her action $(0, 1)$ will give her a payoff of 3, while any other action will give her an expected payoff strictly less than 3. Hence, $(0, 1)$ is the unique rationalizable action for individual 1 and $Z^\infty = \{x_2\}$ is the unique rationalizable outcome. Contrary to the largest consistent set, social rationalizability satisfies individual rationality.

Example 2. Consider again the social environment with $I = \{1, 2, 3\}$, $Z = \{x_0, x_1, x_2, x_3\}$, and the effectiveness relations, as well as the payoffs, being depicted as in Fig. 2. The
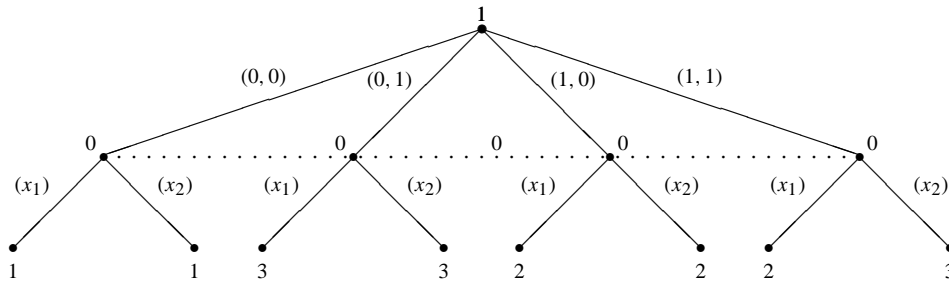
Fig. 5. The extensive-form game for the social environment of Example 1.

extensive-form game defined on this social environment is a three-stage game. At each stage of the game only one individual belonging to I has a non-empty set of actions. That is, any strategy $s_i \in S_i$ is such that $a_i^{i-1}((x_{i-1} \to_{\{i\}} x_i))$ equals 1 or 0. The set of strategies of an individual $i \in I$ is denoted for simplicity by $S_i = \{0, 1\}$. At each stage of the game, the dummy player has a trivial action since $M(h)$ is always a singleton. By Definition 5, it holds that $P^0 = S$. When individual 3 gets the choice, she is really indifferent between moving and not moving, so $P_3^1 = S_3$. 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 $c_2(h^1) \in \prod_{i \in I^+ \setminus \{2\}} \Delta^0(S_i)$ puts positive probability weight on both $a_3^2((x_2 \to_{\{3\}} x_3)) = 1$ and $a_3^2((x_2 \to_{\{3\}} x_3)) = 0$. Hence, the unique optimal behavior for individual 2 is $a_2^1((x_1 \to_{\{2\}} x_2)) = 1$, and P_2^1 is a proper subset of S_2, $P_2^1 = \{1\}$. Initially, individual 1 puts positive probability weight on all strategies of her opponents, and depending on her cautious conjectures she decides to stay at x_0 or to move to x_1, so $P_1^1 = S_1$. However, in the second iteration she knows that individual 2 will move to x_2 for sure when given the move: any $c_1(h^0) \in \prod_{i \in I^+ \setminus \{1\}} \Delta^0(P_i^1)$ gives probability one to $a_2^1((x_1 \to_{\{2\}} x_2)) = 1$. Therefore, the unique optimal behavior for individual 1 is to stay at x_0: $a_1^0((x_0 \to_{\{1\}} x_1)) = 0$. So, $P_1^\infty = \{0\}$, $P_2^\infty = \{1\}$, and $P_3^\infty = S_3$. The unique rationalizable (or stable) outcome is the original status quo, $Z^\infty = \{x_0\}$.

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 Fig. 3. The extensive-form representation of this social environment is depicted in Fig. 6. For $i \in I$, we have $H_i = \{h^0\}$ and $M_i(h^0) = \{(x_0 \to_{\{i\}} x_i)\}$. Any strategy of individual $i \in I$ is such that $a_i^0((x_0 \to_{\{i\}} x_i))$ equals 1 or 0. The set of strategies of individual $i \in I$ is $S_i = \{0, 1\}$. Any strategy of the dummy player is a permutation of $M(h^0) = \{(x_0 \to_{\{1\}} x_1), (x_0 \to_{\{2\}} x_2)\}$. The set of strategies of the dummy player is $S_0 = \{(x_1, x_2), (x_2, x_1)\}$ where (x_1, x_2) means that the first element of (x_1, x_2) contained in $f(a_{-0}^0)$ will be the outcome of the game whenever $f(a_{-0}^0) \neq \emptyset$. By Definition 5, $P^0 = S$. Given any cautious conjecture $c_i(h^0) \in \prod_{j \in I^+ \setminus \{i\}} \Delta^0(S_j)$, individual i has a unique best response which is to move to x_i. So, $a_i^0((x_0 \to_{\{i\}} x_i)) = 1$, $P_i^1 = P_i^\infty = \{1\}$, $i \in I$, and $Z^\infty = \{x_1, x_2\}$.
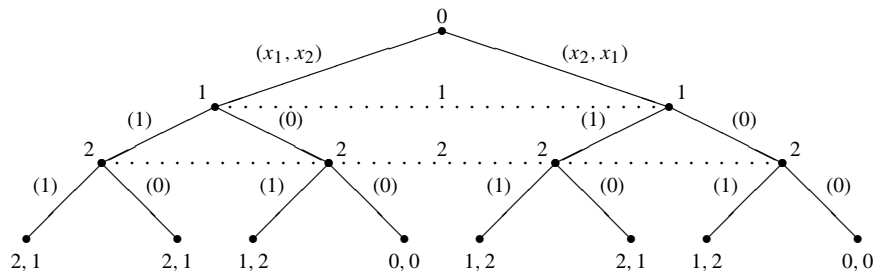

Fig. 6. The extensive-form game for the social environment of Example 3.

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 Fig. 4. The extensive-form representation of this social environment is a one-stage game. We have $H_i = \{h^0\}$ and $M_i(h^0) = \{(x_0 \to_{\{i\}} x_i)\}$, $i \in I$. Any strategy of individual i is such that $a_i^0((x_0 \to_{\{i\}} x_i))$ equals 1 or 0. The set of strategies of individual $i \in I$ is $S_i = \{0, 1\}$. Any strategy of the dummy player is a permutation of $M(h^0) = \{(x_0 \to_{\{1\}} x_1), (x_0 \to_{\{2\}} x_2)\}$. The set of strategies of the dummy player is $S_0 = \{(x_1, x_2), (x_2, x_1)\}$. By Definition 5, $P^0 = S$. Given any cautious conjecture $c_i(h^0) \in \prod_{j \in I^+ \backslash \{i\}} \Delta^0(S_j)$, individual i has a unique best response which is not to move. So, for $i \in I$, $a_i^0((x_0 \to_{\{i\}} x_i)) = 0$, $P_i^1 = P_i^\infty = \{0\}$, and $Z^\infty = \{x_0\}$.

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. In general, what can be said about social environments in which a strictly Pareto-dominant outcome exists, but not all outcomes are Pareto-ranked? Does social rationalizability satisfy the condition that it always selects a Pareto-dominant outcome when one exists? We can formalize this within the theory of social environments.

Consider the social environment Γ^* where $I = \{1, 2, \ldots, \#I\}$, $Z = \{x_0, x_1, \ldots, x_N\}$ and there is one outcome which strictly dominates all other outcomes, for $i \in I$, for $k \neq 0, N$, $u_i(x_N) > u_i(x_k) > u_i(x_0) = 0$. Only $x_0 \to_I x_k$, $k = 1, \ldots, N$, are possible moves. A two-individual case with $N = 3$ is depicted in Fig. 7. We say that social rationalizability satisfies coalitional rationality if it selects the Pareto-dominant outcome, x_N.

In the extensive-form game $G(\Gamma^*)$, we have, for $i \in I$, $H_i = \{h^0\}$ and $M(h^0) = M_i(h^0) = \{(x_0 \to_I x_1), (x_0 \to_I x_2), \ldots, (x_0 \to_I x_N)\}$. $G(\Gamma^*)$ is a one-stage game where
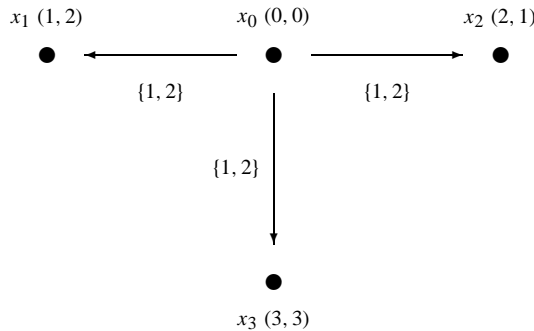

Fig. 7. Coalitional rationality.

all players simultaneously choose an action. A strategy or action of individual i is denoted by $s_i = a_i = (a_{i1}, \ldots, a_{ik}, \ldots, a_{iN})$ where $a_{ik} = a_i((x_0 \rightarrow_I x_k))$. A strategy or action of the dummy player is a permutation of $M(h^0)$.

Example 5. Consider the two-individual and three-move case, so $I = \{1, 2\}$, $Z = \{x_0, x_1, x_2, x_3\}$, and $x_0 \rightarrow_I x_k$, $k = 1, 2, 3$, are the only possible moves. Moreover, consider the special case where $u_1(x_0) = u_2(x_0) = 0$, $u_1(x_1) = u_2(x_2) = 1$, $u_1(x_2) = u_2(x_1) = 2$, $u_1(x_3) = u_2(x_3) = 3$. This social environment is depicted in Fig. 7. The strategies of individual $i \in I$ are such that $a_i^0((x_0 \rightarrow_{\{1,2\}} x_k)) = 1$ or $a_i^0((x_0 \rightarrow_{\{1,2\}} x_k)) = 0$. The set of strategies of individual $i \in I$ is $S_i = \{(0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1), (1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)\} = A_i$, where $(1, 0, 1)$ simply means $a_i^0((x_0 \rightarrow_{\{1,2\}} x_1)) = 1$, $a_i^0((x_0 \rightarrow_{\{1,2\}} x_2)) = 0$, $a_i^0((x_0 \rightarrow_{\{1,2\}} x_3)) = 1$. The strategies of the dummy player are permutations of $M(h^0)$, where $M(h^0) = \{(x_0 \rightarrow_{\{1,2\}} x_1), (x_0 \rightarrow_{\{1,2\}} x_2), (x_0 \rightarrow_{\{1,2\}} x_3)\}$. The set of strategies of the dummy player is $S_0 = \{(x_1, x_2, x_3), (x_2, x_1, x_3), (x_2, x_3, x_1), (x_1, x_3, x_2), (x_3, x_1, x_2), (x_3, x_2, x_1)\} = A_0$. Which outcomes are socially rationalizable? Is the Pareto-dominant outcome the unique socially rationalizable one?

By Definition 5, $P_i^0 = S_i$. We show first that $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$, $(1, 1, 0)$ do not belong to P_i^1. Take any $s_i \in S_i$ such that $a_{i3} = 0$ and take $s_i' \in S_i$ such that $a_{i1}' = a_{i1}$, $a_{i2}' = a_{i2}$ and $a_{i3}' = 1$. It is quite straightforward that, for any cautious conjecture $c_i(h^0) \in \prod_{j \in I^+ \setminus \{i\}} \Delta^0(S_j)$, $U_i(h^0)(s_i, c_i) < U_i(h^0)(s_i', c_i)$. Indeed, the strategies s_i and s_i' give the same payoffs to individual i against the opponent's strategies s_j with $a_{j3} = 0$. But if individual j's strategy is such that $a_{j3} = 1$ we have that either $a_0^0 \in \{(x_3, x_1, x_2), (x_3, x_2, x_1)\}$ and s_i' does strictly better than s_i, or $a_0^0 \notin \{(x_3, x_1, x_2), (x_3, x_2, x_1)\}$ and s_i' does at least as well as s_i.

Next it is shown that all $s_i \in S_i$ with $a_{i3} = 1$ belong to P_i^1 by proving that for any s_i with $a_{i3} = 1$, there exists $c_i(h^0) \in \prod_{j \in I^+ \setminus \{i\}} \Delta(S_j)$ such that s_i is the unique best response among S_i. For instance, $(1, 0, 1)$ is the unique best response against the conjecture $c_i(h^0) \in \prod_{j \in I^+ \setminus \{i\}} \Delta(S_j)$, where, for $j \neq \{0, i\}$,

$$c_i^j(h^0)(s_j) = \begin{cases} 1/3 & \text{if } s_j = (1, 0, 0) \text{ or } s_j = (0, 0, 1) \text{ or } s_j = (1, 1, 1), \\ 0 & \text{otherwise,} \end{cases}$$

and

$$c_i^0(h^0)(s_0) = \begin{cases} 1 & \text{if } s_0 = (x_2, x_3, x_1), \\ 0 & \text{otherwise.} \end{cases}$$

In Table 1 we give the conjectures against which each strategy s_i with $a_{i3} = 1$ is the unique best response. By a continuity argument, see also Lemma 3 below, s_i is also the unique best response against a cautious conjecture that puts weight on all strategies $s_j \in S_j$. So, $P_i^1 = \{(0, 0, 1), (1, 0, 1), (0, 1, 1), (1, 1, 1)\}$.

In the second iteration, individual i knows that individual j will play a strategy in P_j^1. Hence, for all $c_i(h^0) \in \Delta^0(P_j^1) \times \Delta^0(S_0)$, the unique best response of individual i is $s_i = (0, 0, 1)$ which gives her a payoff of 3. Indeed, for all $c_i(h^0) \in \Delta^0(P_j^1) \times \Delta^0(S_0)$, any $s_i' \neq s_i$ belonging to P_i^1 will give her a payoff less than 3, because $c_i(h^0)$ puts positive

Table 1
Unique best response and conjectures

	s_j	s_i			
		$(0, 0, 1)$	$(1, 0, 1)$	$(0, 1, 1)$	$(1, 1, 1)$
	$(0, 0, 0)$	0	0	0	0
	$(0, 0, 1)$	1/2	1/3	1/3	1/3
	$(1, 0, 0)$	0	1/3	0	1/3
	$(0, 1, 0)$	0	0	1/3	1/3
	$(1, 1, 0)$	0	0	0	0
	$(0, 1, 1)$	0	0	0	0
	$(1, 0, 1)$	0	0	0	0
	$(1, 1, 1)$	1/2	1/3	1/3	0
s_0	$(x_1, x_2, x_3), (x_2, x_1, x_3)$	1	0	0	0
	$(x_3, x_1, x_2), (x_3, x_2, x_1)$	0	0	0	1
	(x_2, x_3, x_1)	0	1	0	0
	(x_1, x_3, x_2)	0	0	1	0

probability on $s_j' = s_i'$ and $c_i^0(h^0)$ has full support. So, $P_i^2 = \{(0, 0, 1)\} = P_i^\infty$, $i = 1, 2$, and $Z^\infty = \{x_3\}$. In Example 5, the case with two individuals and a Pareto-dominant outcome, 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, some lemmata are used. Lemma 3, whose proof is obvious and left to the reader, tells us that if a strategy of individual i is the unique best response against a conjecture c_i (possibly degenerate), then it is also the unique best response against some cautious conjecture c_i^*.

Lemma 3. *Take any $s_i \in S_i$. If there exists c_i such that* (i) $c_i(h^0) \in \prod_{j \in I^+ \setminus \{i\}} \Delta(S_j)$ *and* (ii) *for all $s_i' \in S_i$, $s_i' \neq s_i$, $U_i(h^0)(s_i, c_i) > U_i(h^0)(s_i', c_i)$, then there exists c_i^* such that* (iii) $c_i^*(h^0) \in \prod_{j \in I^+ \setminus \{i\}} \Delta^0(S_j)$ *and* (iv) *for all $s_i' \in S_i \setminus \{s_i\}$, $U_i(h^0)(s_i, c_i^*) > U_i(h^0)(s_i', c_i^*)$.*

Lemma 4 tells us that any strategy s_i such that individual i blocks the move to x_N (i.e. $a_i^0((x_0 \to_I x_N)) = 0$ or $a_{iN} = 0$) is never a best response whatever the cautious conjecture c_i. Indeed, the strategy s_i', where s_i' is the same as s_i except that individual i joins the move to x_N, is always a strictly better response.

Lemma 4. *For $i \in I$, take any $s_i \in S_i$ with $a_{iN} = 0$. Take $s_i' \in S_i$ such that $a_{ik}' = a_{ik}$ for $k = 1, \ldots, N - 1$ and $a_{iN}' = 1$. Then, $U_i(h^0)(s_i', c_i) > U_i(h^0)(s_i, c_i)$ for all $c_i \in \prod_{j \in I^+ \setminus \{i\}} \Delta^0(S_j)$.*

We introduce some additional notations. For $i \in I$, given $s_i \in S_i$, we define $K_i(s_i) = \#\{k \mid a_{ik} = 1\}$. Moreover, we denoted by $e(k)$ the strategy such that the kth component is 1 and the other components are 0, and **1** is the unit vector, that is, the strategy where the individual agrees to join every move. Lemma 5 establishes that there exists a conjecture c_i such that any strategy s_i where individual i agrees to move to x_N is her unique best

response. This conjecture is such that it puts weight on $s_j = e(k)$ whenever $a_{ik} = 1$ and on $s_j = \mathbf{1}$. The former part of the conjecture guarantees that s_i gives higher utility than $s_i' \neq s_i$ whenever s_i' blocks moves that are not blocked by s_i. The latter part, together with a belief on the choice of the dummy player, implies that s_i outperforms any s_i' that agrees to strictly more moves than s_i.

Lemma 5. *For $i \in I$, take any $s_i \in S_i$ such that $a_{iN} = 1$. Then, for all $s_i' \in S_i \setminus \{s_i\}$, we have $U_i(h^0)(s_i, c_i) > U_i(h^0)(s_i', c_i)$, where $c_i(h^0) \in \prod_{j \in I^+ \setminus \{i\}} \Delta(S_j)$ is such that, for $j \neq 0$,*

$$c_i^j(h^0)(s_j) = \begin{cases} \frac{1}{K_i + 1} & \text{if } s_j = e(k) \text{ and } a_{ik} = 1, \text{ or } s_j = \mathbf{1}, \\ 0 & \text{otherwise}, \end{cases}$$

and

$$c_i^0(h^0)(s_0) = \begin{cases} 1 & \text{if } s_0 = \big((x_k)_{a_{ik} \neq 1}, x_N, (x_k)_{a_{ik} = 1}\big), \\ 0 & \text{otherwise}. \end{cases}$$

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

Theorem 6. *Consider the social environment Γ^*. There is a unique strategy of individual $i \in I$ that is socially rationalizable, $P_i^\infty = \{e(N)\}$.*

Proof. By Definition 5, $P_i^0 = S_i$ and $P^0 = \prod_{i \in I^+} S_i$. In the first iteration, by Lemma 4, all $s_i \in P_i^0$ such that $a_{iN} = 0$ do not belong to P_i^1. By Lemma 3 and Lemma 5, all s_i such that $a_{iN} = 1$ do belong to P_i^1. So, $P_i^1 = \{s_i \mid a_{iN} = 1\}$. As always, $P_0^1 = S_0$.

In the second iteration, for all $c_i(h^0) \in \prod_{j \in I^+ \setminus \{i\}} \Delta^0(P_j^1)$, the strategy s_i such that $a_{iN} = 1$ and $a_{ik} = 0$ if $k \neq N$ gives to individual $i \in I$ a utility $U_i(h^0)(s_i, c_i) = u_i(x_N)$. However, for all $s_i' \in P_i^1 \setminus \{s_i\}$, $U_i(h^0)(s_i', c_i) < u_i(x_N)$ for all c_i, because for some $k < N$, $a_{ik}' = 1$, and the cautiousness of c_i implies that with positive probability the opponents of i agree on $\{(x_0 \rightarrow_I x_k)\}$ and x_k is the first element of the permutation chosen by player 0. This would lead to utility $u_i(x_k) < u_i(x_N)$. So, $P_i^2 = \{e(N)\} = P_i^\infty$, $i \in I$. $\quad\square$

The above result implies that social rationalizability satisfies the property of coalitional rationality. When there is a Pareto-dominant outcome it is selected by a coalition. Each individual only agrees to move to the Pareto dominating outcome, and blocks all other moves.

Corollary 7. *Consider the social environment Γ^*. We have $Z^\infty = \{x_N\}$.*

Finally, notice that $LCS(\Gamma^*) = Z \setminus \{x_0\}$; the unique OSSB is $\sigma(x_0) = \{(x_0, x_N)\}$; and the unique CSSB is $\sigma(x_0) = \{(x_0, x_k) \mid x_k$ is not Pareto-dominated by all other $x \in Z \setminus \{x_0, x_k\}\}$. In case the outcomes are Pareto ranked: $u_i(x_N) > u_i(x_{N-1}) > \cdots > u_i(x_1) > u_i(x_0) = 0$, then the unique CSSB is $\sigma(x_0) = \{(x_0, x_k) \mid x_k \in Z \setminus \{x_0, x_1\}\}$.

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 associate a particular multistage game with a social environment. This enables us to model individual behavior in a social environment. Moreover, it 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 conjectures on actions taken by a dummy player which will determine which move in the set of moves on which there is agreement will be carried out.

We have shown that for all social environments the set of socially rationalizable outcomes is non-empty. The non-emptiness of social rationalizability 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 and incorporates forward induction elements. 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 moves, it should select a Pareto dominating one for sure. It is shown that social rationalizability is consistent with coalitional rationality.

Finally, we would like to mention a recent related contribution. Ambrus (2002) has investigated the implications of groups or coalitions of players acting in their collective interest in noncooperative normal-form games. It is assumed that players are unable to make binding agreements, and pre-play communication is neither precluded nor assumed. The main idea is that each member of a coalition will confine play to a subset of their strategies if it is in their mutual interest to do so. This leads to an iterative procedure of restricting players' beliefs and actions in the game. The iterative procedure defines the set of coalitionally rationalizable strategies. One similarity between our paper and Ambrus' paper is that both deal with coalitional reasoning within a non-equilibrium framework. The main difference is that Ambrus' coalitional rationalizability concept is defined for normal-form games, while we look at extensive-form games defined on social environments. Another difference is that coalitional moves are publicly observed in social environments but in Ambrus' paper it is assumed that a player moves secretly.

Acknowledgments

Appendix A

Proof of Theorem 2. Consider the iterative procedure provided by Definition 5. For each iteration n, choose a consistent updating system c_i such that $c_i(h') \in \prod_{j \in I^+ \setminus \{i\}} \Delta^0(P_j^{n-1})$ for all $h' \in H_i$ allowed by P^{n-1}. Consider any $s_i \in P_i^{n-1}$ such that $U_i(h^0)(s_i, c_i) \geqslant U_i(h^0)(\hat{s}_i, c_i)$ for all $\hat{s}_i \in P_i^{n-1}$. If h' is allowed by s_i and P_{-i}^{n-1} then it follows that $U_i(h')(s_i, c_i) \geqslant U_i(h')(s_i/\hat{s}_i^{h'}, c_i)$ for all $\hat{s}_i \in P_i^{n-1}$. It follows that $s_i \in P_i^n$, 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 $n, n' \geqslant N$. It follows that $P^\infty(K) = P^N \neq \emptyset$. Any $(s_i)_{i \in I} \in P^\infty(K)$ yields an outcome; so $Z_K^\infty \neq \emptyset$. As a subset of the finite set Z it holds that Z_K^∞ is finite. Now it follows from the definition of the limit superior that $Z^\infty \neq \emptyset$. $\quad \square$

Proof of Lemma 4. Consider any profile $s_{-i} \in \prod_{j \in I^+ \setminus \{i\}} S_j$. We denote by $f(s_{-0,i})$ the moves on which the opponents of individual i agree when their strategies are given by $s_{-0,i}$.

(i) For all $s_{-i} \in \prod_{j \in I^+ \setminus \{i\}} S_j$, if $(x_0 \to_I x_N) \notin f(s_{-0,i})$ then $U_i(h^0)(s_i', s_{-i}) = U_i(h^0)(s_i, s_{-i})$.
(ii) For all $s_{-i} \in \prod_{j \in I^+ \setminus \{i\}} S_j$, if $(x_0 \to_I x_N) \in f(s_{-0,i})$ and x_N is the first element of s_0, then $U_i(h^0)(s_i', s_{-i}) > U_i(h^0)(s_i, s_{-i})$.
(iii) For all $s_{-i} \in \prod_{j \in I^+ \setminus \{i\}} S_j$, if $(x_0 \to_I x_N) \in f(s_{-0,i})$ and x_N is not the first element of s_0, then $U_i(h^0)(s_i', s_{-i}) \geqslant U_i(h^0)(s_i, s_{-i})$.

Hence, $U_i(h^0)(s_i', c_i) > U_i(h^0)(s_i, c_i)$ for all $c_i \in \prod_{j \in I^+ \setminus \{i\}} \Delta^0(S_j)$. $\quad \square$

Proof of Lemma 5. Let $\overline{M}$ be the set of moves on which the opponents of i could potentially agree. Then, either $\overline{M} = M$, or $\overline{M} = \{(x_0 \to_I x_k) \in M \mid$ for some k such that $a_{ik} = 1\}$, or $\overline{M} = \emptyset$. Two cases have to be considered.

Case 1. Consider s_i' such that, for some k, $a_{ik} = 1$ and $a_{ik}' = 0$. Then, against $\overline{M} = \{(x_0 \to_I x_k)\}$, s_i' gives a payoff of 0 to i and s_i gives a strictly positive payoff to i. It is straightforward that s_i performs at least as good as s_i' against any other potential $\overline{M}$. Hence, $U_i(h^0)(s_i, c_i) > U_i(h^0)(s_i', c_i)$.

Case 2. Consider $s_i' \neq s_i$ such that $a_{ik} = 1$ implies $a_{ik}' = 1$. Then, against $\overline{M} = M$, s_i' gives a payoff of $u_i(x_k)$ for k the highest ranked element in s_0 such that $a_{ik} = 0$ and $a_{ik}' = 1$, which is strictly less than $u_i(x_N)$, the payoff of s_i against M. It is straightforward that s_i performs at least as good as s_i' against any other potential $\overline{M}$. Hence, $U_i(h^0)(s_i, c_i) > U_i(h^0)(s_i', c_i)$. $\quad \square$

References

Ambrus, A., 2002. Coalitional rationality. Working paper. Princeton University, Princeton.

Aumann, R., Myerson, R., 1988. Endogenous formation of links between players and coalitions: an application of the Shapley value. In: Roth, A. (Ed.), The Shapley Value: Essays in Honor of Lloyd Shapley. Cambridge Univ. Press, Cambridge, pp. 175–191.

Battigalli, P., 1996. Strategic rationality orderings and the best rationalization principle. Games Econ. Behav. 13, 178–200.

Battigalli, P., 1997. On rationalizability in extensive games. J. Econ. Theory 74, 40–61.

Bernheim, D., 1984. Rationalizable strategic behavior. Econometrica 52, 1007–1028.

Bloch, F., 1996. Sequential formation of coalitions in games with externalities and fixed payoff division. Games Econ. Behav. 14, 90–123.

Chwe, M.S., 1994. Farsighted coalitional stability. J. Econ. Theory 63, 299–325.

Greenberg, J., 1990. The Theory of Social Situations: An Alternative Game-Theoretic Approach. Cambridge Univ. Press, Cambridge.

Greenberg, J., Monderer, D., Shitovitz, B., 1996. Multistage situations. Econometrica 64, 1415–1437.

Herings, P.J.J., Vannetelbosch, V.J., 1999. Refinements of rationalizability for normal-form games. Int. J. Game Theory 28, 53–68.

Herings, P.J.J., Mauleon, A., Vannetelbosch, V.J., 2000. Social rationalizability. METEOR research memorandum 00/17. Universiteit Maastricht.

Mariotti, M., 1997. A model of agreements in strategic form games. J. Econ. Theory 74, 196–217.

Pearce, D.G., 1984. Rationalizable strategic behavior and the problem of perfection. Econometrica 52, 1029–1050.

Perry, M., Reny, P.J., 1994. A noncooperative view of coalition formation and the core. Econometrica 62, 795–817.

Ray, D., Vohra, R., 1999. A theory of endogenous coalition structures. Games Econ. Behav. 26, 286–336.

Shimoji, M., Watson, J., 1998. Conditional dominance, rationalizability, and game forms. J. Econ. Theory 83, 161–195.

Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton Univ. Press, Princeton.

Xue, L., 1998. Coalitional stability under perfect foresight. J. Econ. Theory 11, 603–627.