SOCIAL RATIONALIZABILITY WITH MEDIATION

P. Jean-Jacques Herings, Ana Mauleon, Vincent Vannetelbosch

REPRINT | 3213

UCLouvain

CORE | LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics

Dynamic Games and Applications
https://doi.org/10.1007/s13235-022-00465-9

Social Rationalizability with Mediation

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

Accepted: 15 July 2022

Abstract
We propose a solution concept for social environments called social rationalizability with mediation that identifies the consequences of common knowledge of rationality and farsightedness. In a social environment several coalitions may and could be willing to move at the same time. Individuals not only hold conjectures about the behaviors of other individuals but also about how a mediator is going to solve conflicts of interest. The set of socially rationalizable outcomes with mediation is shown to be non-empty for all social environments, and it can be computed by an iterative reduction procedure. We show that social rationalizability with mediation does not necessarily satisfy coalitional rationality when the number of coalition members is greater than two.

Keywords Social environments · Rationalizability · Mediation · Coalitional rationality

JEL Classification C70 · C72 · C78

1 Introduction

Social environments [4] constitute a framework in which it is possible to study how groups of agents interact in a society. It specifies what each coalition can do if and when it forms.

This article is part of the topical collection "Group Formation and Farsightedness" edited by Francis Bloch, Ana Mauleon and Vincent Vannetelbosch.

✉ P. Jean-Jacques Herings
P.J.J.Herings@tilburguniversity.edu

Ana Mauleon
ana.mauleon@usaintlouis.be

Vincent Vannetelbosch
vincent.vannetelbosch@uclouvain.be

1 Tilburg University, Tilburg, The Netherlands

2 CEREC, UCLouvain Saint-Louis, Brussels, Belgium

3 CORE/LIDAM, UCLouvain, Louvain-la-Neuve, Belgium

Published online: 23 August 2022

Social environments are general enough to encompass the representation of a cooperative game, an extensive-form game with perfect information, as well as a normal-form game.[1]

We propose a new solution concept for social environments called social rationalizability with mediation that identifies the consequences of common knowledge of rationality and farsightedness. Given that social environments mainly deal with the behavior of coalitions, whereas rationalizability is about the implications of rationality of individuals, we convert coalitional behavior into individual behavior. Individual participation in a coalition basically reverts either to agree to a coalitional move or to object to it and block it. In a social environment several coalitions may and could be willing to move at the same time. Conflicts of interest may arise: one coalition may try to preempt the move of another coalition or coordination problems in and between coalitions may arise. We assume that individuals not only hold conjectures about the behaviors of other individuals but also about how a mediator is going to solve conflicts of interest.

In the rationalizability approach, conjectures are not assumed to be correct, but are only constrained by considerations of rationality: individuals are rational and this is common knowledge. That is, each individual believes that the behavior of 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 believes that every other individual believes that every other individual's behavior is a best responses to some conjecture, and so on.

Central to social rationalizability with mediation are the notions of individual behavior and of conjectures about the mediator's behavior. An individual behavior describes, for each history, the coalitional moves the individual agrees to join and those she decides to block. The mediator (player 0 whose payoff is always zero) chooses a move for each possible set of moves on which the individuals could agree to join, and individuals hold conjectures about the behavior of the mediator. Our definition of social rationalizability is motivated by Pearce's [22] original extensive-form rationalizability.[2]

We show that the set of socially rationalizable outcomes with mediation is non-empty for all social environments and it can be computed by an iterative reduction procedure. Since social environments deal with coalitional moves, one may wonder if social rationalizability with mediation satisfies, in general, the property of coalitional rationality. That is, in a situation in which a coalition of two or more individuals can move from a status quo to different outcomes that are Pareto ranked, does social rationalizability with mediation prescribe that players coordinate on the outcome that Pareto dominates all others? We find that social rationalizability with mediation does not necessarily satisfy coalitional rationality when the number of coalition members is greater than two.

The most closely related paper to ours is Herings, Mauleon and Vannetelbosch [18] who also define rationalizability for social environments. There are two main differences. First, they do not define rationalizability directly on the social environment but rather embed the social environment in a multi-stage game and then use the notion of extensive-form game rationalizability by Pearce [22] to solve the multi-stage game. Second, their mediator is a dummy player whose payoff is always zero but who chooses an action consisting of a permutation on the set of feasible moves after each history. Such a permutation indicates

[1] Chwe [4] and Xue [27] propose, respectively, the largest consistent set and the optimistic or conservative stable standards of behavior as solution concepts for social environments. The largest consistent set may fail to satisfy individual rationality while the stable standards of behavior may be empty-valued or rule out too much.

[2] Related papers to extensive-form rationalizability are among others Bernheim [2], Shimoji and Watson [25], Vannetelbosch [26].

the order according to which moves are implemented. Suppose that, from a status quo, individuals can move to three outcomes x_1, x_2, and x_3. The mediator imposes a ranking over those three outcomes, for instance, (x_2, x_3, x_1). If the individuals find the moves to x_1, x_2, and x_3 acceptable, then x_2 is implemented. If they only agree on x_1 and x_2, then x_2 is still implemented. Such behavior of the dummy player guarantees that individuals coordinate on the Pareto-dominant outcome. However, with a more general mediator, it may happen that if the individuals find the moves to x_1, x_2, and x_3 acceptable, then the mediator chooses to implement x_2. But, if they only agree on x_1 and x_2, then she chooses to implement x_1 instead. We show that once the behavior of the mediator is not constrained to the choice of a permutation over alternatives, individuals may fail to coordinate on the Pareto-dominant outcome.

Besides the largest consistent set and the optimistic or conservative stable standards of behavior, another common notion for analyzing outcomes that emerge in the long run when individuals are farsighted is the farsighted stable set [4, 11, 19, 23].[3] However, the farsighted stable set suffers from a conceptual drawback: the maximality issue. For instance, while coalitional moves improve on existing outcomes along a farsighted objection, coalitions might do even better by an alternative deviation. Dutta and Vohra [8] propose the rational expectations farsighted stable set and the strong rational expectations farsighted stable set that restrict coalitions to hold common, history-independent expectations that incorporate maximality regarding the continuation path. More recently, Ray and Vohra [24] incorporate absolute maximality into the definition of the farsighted stable set. Absolute maximality requires immunity to all deviations, not just by the coalition that moves or by those coalitions that intersect the one that moves. Asking for maximality can be interpreted as imposing coalitions to play a form of coalitional best responses. We find that social rationalizability may violate coalitional rationality. In other words, the rationality of individuals is not enough to guarantee that coalitional best responses or maximality do emerge endogenously.

The paper is organized as follows. In Sect. 2 we define social environments and social rationalizability with mediation and we provide an illustration. In Sect. 3 we show that social rationalizability with mediation satisfies two-player coalitional rationality, while in Sect. 4 we show that coalitional rationality does not necessarily hold for larger coalitions. In Sect. 5 we provide an alternative definition of social rationalizability with mediation and we show the equivalence with our original definition. Finally, we show that, if we restrict the behavior of the mediator to be consistent with a permutation over alternatives, then we can guarantee that individuals coordinate on the Pareto-dominant outcome.

2 Rationalizable Social Behaviors with Mediation

2.1 Social Environments

We study a *social environment* $\Gamma = \langle I, Z, (u_i)_{i\in I}, \{\rightarrow_S\}_{S\subseteq I, S\neq\emptyset}, x_0\rangle$, where I is the finite set of *individuals*, Z is the finite set of *outcomes*, $\{\rightarrow_S\}_{S\subseteq I, S\neq\emptyset}$ are *effectiveness relations* defined on Z, for every individual $i \in I$, $u_i : Z \rightarrow \mathbb{R}$ is her *utility function*, and $x_0 \in Z$ is the *initial status quo*. The relation $\rightarrow_S$ represents what coalition S can do: $x \rightarrow_S y$ means that if x is the current status quo, then coalition S can make y the new status quo. It does

[3] Alternative notions of farsightedness are suggested by Bloch and van den Nouweland [3], Diamantoudi and Xue [5], Dutta et al. [6], Dutta and Vohra [8], Dutta and Vartianen [7], Herings et al. [11–14], Karos and Robles [15], Kimya [16], Luo et al. [17], Mauleon and Vannetelbosch [18], Page et al. [21], and Page and Wooders [20] among others.

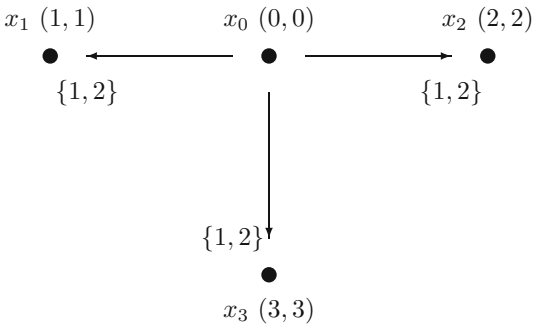

Fig. 1 An example of a social environment with two individuals

not mean that coalition S can enforce y no matter what anyone else does; after S moves to y from x, another coalition S' might move to z, where $y \to_{S'} z$. A priori no restrictions are imposed on the effectiveness relations $\{\to_S\}_{S\subseteq I, S\neq\emptyset}$. For example, the effectiveness relation can be empty, $x \to_S x$ is allowed for, and $x \to_S y$ does not imply $y \to_S x$. All *actions* or *moves* are public and the individuals care only about the end outcome. Both non-cooperative and cooperative games can be modeled as a social environment. The definition of a social environment follows Chwe [4], except that our theory allows the set of stable outcomes to depend on the initial status quo, which is therefore an explicit part of the description of our social environment.

Figure 1 presents an example of a social environment in which a coalition of two individuals may decide to move from the initial status quo x_0, where they both get a utility of 0, to outcome x_1 and getting both 1 unit of utility, or to outcome x_2 and obtaining 2 units of utility each, or to outcome x_3 and receiving both 3 units of utility. The social environment is therefore given by $I = \{1, 2\}$, $Z = \{x_0, x_1, x_2, x_3\}$, for $k = 1, 2, 3$, $x_0 \to_I x_k$ are the only possible moves, and, for $i = 1, 2$, for $k = 0, 1, 2, 3$, $u_i(x_k) = k$.

2.2 Individual and Social Behaviors

In what follows, we denote the move $x \to_S y$ of coalition S from x to y by (xy, S). When none of the coalitions is willing and able to move at x, then the *no-move* results, which 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 is given by $M = \{(xy, S) \mid x, y \in Z,\ x \to_S y\}$. The set of all possible no-moves is equal to $N = \{(xx, \emptyset) \mid x \in Z\}$. We denote by $h = (x_0, m_1, m_2, \ldots, m_{k-1})$ a *history* of *length* k, where $x_0 \in Z$ is the initial status quo, for $j = 1, \ldots, k-2$, $m_j = (m_j^- m_j^+, m_j^c) \in M$, $m_{k-1} \in M \cup N$, $m_1^- = x_0$, and $m_j^+ = m_{j+1}^-$. The length of a history h is denoted by $\ell(h)$.

The set of all histories is denoted by H^*. The set of all histories $h \in H^*$ such that $m_{\ell(h)-1} \in M$, i.e., the set of *non-terminal histories*, is denoted by H. The set of feasible moves after a non-terminal history $h \in H$ is given by $M(h) = \{m \in M \mid m^- = h^+\}$ and, for $i \in I$, $M_i(h) = \{(xy, S) \in M(h) \mid i \in S\}$ denotes the set of feasible moves after history h involving individual i. The set containing the no-move after a non-terminal history h is given by $N(h) = \{(h^+h^+, \emptyset)\}$. There are no feasible moves after a history h such that $m_{\ell(h)-1} \in N$, i.e., a *terminal history*.

To make the length of a history h explicit, we sometimes use the notation h^k, where k is the length of the history. Let $h^- = x_0$ be the *initial status quo* of h and $h^+ = m_{\ell(h)-1}^+$ be the

end outcome of h. Given h^k and $j, k \in \mathbb{N}$ with $j \leq k$, 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 $h^j \leq h^k$.[4] If we write $h^j < h^k$, then h^j is a proper sub-history of h^k, so $j < k$.

We denote by $H(J)$ the set of non-terminal histories with at most J moves. That is, $H(J) = \{h \in H \mid \ell(h) \leq J + 1\}$. Temporarily, we fix J and consider only histories in $H(J)$. Let $H_i(J) = \{h \in H(J) \mid M_i(h) \neq \emptyset\}$ be the set of non-terminal histories that contain at most J moves and after which individual i is involved in a move. A *social behavior* selects after any non-terminal history a move or the no-move. A social behavior is denoted by $s = (s(h))_{h \in H(J)}$, where $s(h) \in M(h) \cup N(h)$. Let SB 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. To do this, we examine individual behaviors first.

We model an *individual behavior* as, for each relevant 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. 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.

Let $H_0(J) = \{h \in H(J) \mid M(h) \neq \emptyset\}$ be the set of histories that contain at most J moves and after which there is at least one feasible move. It may happen that the individuals agree on more than one move. We denote by $\mathcal{M}(h) = \{\overline{M} \mid \overline{M} \subseteq M(h)\}$ the collection of sets of feasible moves after $h \in H_0(J)$. Notice that $\mathcal{M}(h)$ contains at least two elements, one of which is the empty set. For every history $h \in H_0(J)$, the so-called agreement function is a mapping $f(\cdot \mid h) : \prod_{i \in I} B_i \to \mathcal{M}(h)$ which associates to the profiles of individual behaviors the set of moves after history h on which there is agreement, so $f((b_i)_{i \in I} \mid h) = \overline{M}$ 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 $b_i((xy, S) \mid h) = 0$. Notice that by this definition we have $f((b_i)_{i \in I} \mid h) = \emptyset$ if there is no move on which there is agreement.

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, and the no-move when no agreement is possible. A profile of individual behaviors may induce, potentially, multiple social behaviors.

2.3 Beliefs, Conjectures, and Payoffs

A problem or a conflict may arise when there are several moves on which agreement is possible. We assume that there is a *mediator*, referred to as player 0, who always obtains a payoff of zero. The mediator chooses one move among any set of possible agreements after history $h \in H_0(J)$. Histories $h \in H(J) \setminus H_0(J)$ are automatically followed by the no-move in $N(h)$. Let $b_0 = (b_0(\cdot \mid h))_{h \in H_0(J)}$ be a behavior of player 0, where $b_0(\cdot \mid h) : \mathcal{M}(h) \to M(h) \cup N(h)$ and $b_0(\overline{M} \mid h) \in \overline{M}$ whenever $\overline{M} \neq \emptyset$. If $\overline{M} = \emptyset$, then $b_0(\overline{M} \mid h) \in N(h)$. Let B_0 be the set of behaviors of player 0.

Rationalizability assumes that individuals form conjectures about each others' behavior, including the behavior of the mediator, player 0, and then optimize subject to these con-

[4] 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 makes the move from one outcome to another.

jectures. We restrict the individuals to hold uncorrelated conjectures about the behaviors of their opponents and player 0. 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 $c_i : H_i(J) \to \prod_{j\neq i} \Delta(B_j) \times \Delta(B_0)$.[5] For $b_{-i} \in \prod_{j\neq i} B_j$, 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 by $c_i^j(h')(b_j) \in \Delta(B_j)$ the probability individual i conjectures at history h' that player j's behavior is b_j, and by $c_i^0(h')(b_0) \in \Delta(B_0)$ the probability individual i conjectures at history h' that player 0's behavior is b_0.

A profile $(b_i, b_{-i}, b_0) \in B_i \times \prod_{j\neq i} B_j \times B_0$ is said to *allow* for $h = (x_0, m_1, \ldots, m_k) \in H_i(J)$ if

(i) $\forall j \in \{1, \ldots, k\}$, $\forall i \in m_j^c$, $b_i(m_j \mid h^j) = 1$,
(ii) $\forall j \in \{1, \ldots, k\}$, $b_0(f((b_i)_{i\in I} \mid h^j) \mid h^j) = m_j$.

A conjecture c_i is said to *allow* for $h \in H_i(J)$ if there is $b_i \in B_i$ and (b_{-i}, b_0) in the support of c_i such that (b_i, b_{-i}, b_0) allows for h. A behavior $b_i \in B_i$ and set $A_{-i} \subseteq \prod_{j\neq i} B_j$ is said to *allow* for h if there is $(b_{-i}, b_0) \in A_{-i} \times B_0$ such that (b_i, b_{-i}, b_0) allows for h.

2.4 Social Rationalizability with Mediation

We next propose a definition of social rationalizability with mediation that is motivated by extensive-form rationalizability as defined in Pearce [22] and is based on a reduction procedure. Social rationalizability is derived from two assumptions: (1) individuals are rational, and (2) this is common knowledge at the initial 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.

Definition 1 A *consistent updating system* for individual i is a mapping $c_i : H_i(J) \to \prod_{j\neq i} \Delta(B_j) \times \Delta(B_0)$ such that, for all $g, h \in H_i(J)$,

(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 such that h is allowed for and that no conjecture is changed unless falsified. Notice that a conjecture of individual i at history h is an element of $\prod_{j\neq i} \Delta(B_j) \times \Delta(B_0)$, so describes the behavior of the other players for every possible sequence of moves. The conjecture at history h serves as the prior. To form the posterior, individuals update according to Bayes rule.

The following example illustrates Bayesian updating on the basis of a consistent updating system.

Example 1 Consider the social environment Γ, where $I = \{1, 2\}$, $Z = \{x_0, x_1, x_2, x_3\}$, and the feasible moves are given by $x_0 \to_{\{1\}} x_1$, $x_1 \to_{\{2\}} x_2$, and $x_2 \to_{\{1\}} x_3$. We do not specify utility functions as they are irrelevant for this illustration. The set H consists of four non-terminal histories, $h_0 = (x_0)$, $h_1 = (x_0, (x_0x_1, \{1\}))$, $h_2 = (x_0, (x_0x_1, \{1\}), (x_1x_2, \{2\}))$, and $h_3 = (x_0, (x_0x_1, \{1\}), (x_1x_2, \{2\}), (x_2x_3, \{1\}))$. After history h_3, the no-move is the only possibility. For $J \geq 3$, the set $H_0(J)$ of histories after which there is at least one feasible move consists of the histories h_0, h_1, and h_2. We have that $M(h_0) = M_1(h_0) = \{(x_0x_1, \{1\})\}$,

[5] As general notation, we denote by $\Delta(X)$ the set of all probability measures on a finite set X and by $\Delta^\circ(X)$ the set of all probability measures giving positive probability to each member of X.

$M(h_1) = M_2(h_1) = \{(x_1x_2, \{2\})\}$, and $M(h_2) = M_1(h_2) = \{(x_2x_3, \{1\})\}$. Since there is only one feasible move after each history, there is no need to introduce the mediator.

There are two possible behaviors for individual 2, b_2 and b_2', where $b_2(x_1x_2, \{2\}) = 0$ and $b_2'(x_1x_2, \{2\}) = 1$, which indicate that individual 2 is willing, respectively not willing, to move from state x_1 to state x_2. Individual 1 forms conjectures at histories h_0 and h_2.

Consider the case where $c_1^2(h_0)(b_2) = 1/2$ and $c_1^2(h_0)(b_2') = 1/2$, so at h_0, individual 1 considers both behaviors of individual 2 equally likely. Since c_1^2 allows for h_2, consistency of the updating system requires that the conjectures of individual 1 at h_2 coincide with those at h_0, so $c_1^2(h_2) = c_1^2(h_0)$, which implies $c_1^2(h_2)(b_2) = 1/2$ and $c_1^2(h_2)(b_2') = 1/2$. Applying Bayes rule using the conjecture $c_1^2(h_2)$ as prior yields that at h_2 individual 1 puts weight 1 on b_2 and weight 0 on b_2'.

Consider next the case where $c_1^2(h_0)(b_2) = 1$ and $c_1^2(h_0)(b_2') = 0$, so at h_0 individual 1 is certain that individual 2 behaves in conformity with b_2. Since c_1^2 does not allow for h_2, consistency of the updating system imposes no restrictions on $c_1^2(h_2)$. For instance, one possibility is $c_1^2(h_2)(b_2) = 1/3$ and $c_1^2(h_2)(b_2') = 2/3$. Applying Bayes rule using the conjecture $c_1^2(h_2)$ as prior implies that at h_2 individual 1 puts weight 1 on b_2 and weight 0 on b_2'.

Formally, social rationalizability with mediation is the result of a reduction procedure that is defined as follows.

Definition 2 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 a consistent updating system c_i such that for all $h' \in H_i(J)$ that are allowed by b_i and P_{-i}^{n-1} it holds that

(iia) $c_i(h') \in \prod_{j\neq i} \Delta^\circ(P_j^{n-1}) \times \Delta^\circ(B_0)$,

(iib) for all $\widehat{b_i} \in P_i^{n-1}$, $U_i(h')(b_i, c_i) \geq U_i(h')(b_i/\widehat{b}_i^{h'}, c_i)$, where $U_i(h')(b_i, c_i)$ denotes the expected payoff of individual i given (b_i, c_i) conditional on reaching history h' and $b_i/\widehat{b}_i^{h'}$ 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 $P^\infty(J) = \lim_{n\to\infty} P^n$ is the set of *rationalizable individual behaviors* where histories contain at most J moves.

In Definition 2 individuals are cautious, meaning that they assign positive probability to all behaviors of their opponents in P_{-i}^{n-1} and of player 0 in B_0.

Let $S^\infty(J)$ denote the set of *rationalizable social behaviors*. A social behavior $s \in SB$ belongs to $S^\infty(J)$ if there exists $(b_i)_{i\in I} \in P^\infty(J)$ such that, for every $h \in H(J)$, $s(h) \in M(h)$ implies $s(h) \in f((b_i)_{i\in I} \mid h)$ and $s(h) \in N(h)$ implies $f((b_i)_{i\in I} \mid h) = \emptyset$.

Let $h^{-1}(\{x\}) = \{h \in H(J+1) \mid \ell(h) = J+2 \text{ and } h^+ = x\} \cup \{h \in H^* \setminus H(J+1) \mid \ell(h) \leq J+2 \text{ and } h^+ = x\}$ be the set of histories of length at most $J+2$ ending at $x \in Z$. We denote by $Z_J^\infty(x_0)$ the set of *rationalizable outcomes* with initial status quo $x_0 \in Z$. It is given by $Z_J^\infty(x_0) = \{x \in Z \mid \exists(x_0, m_1, \ldots, m_k) \in h^{-1}(\{x\}), \exists s \in S^\infty(J) \text{ such that } \forall j = 1, \ldots, k,\ s(x_0, m_1, \ldots, m_{j-1}) = m_j\}$. 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 allowed to depend on the initial state x_0. This is different from the largest consistent set as defined in Chwe [4], which does not allow for such a dependence.

The set of socially rationalizable outcomes is never empty.

Theorem 1 $Z^{\infty}(x_0) \neq \emptyset$.

The proof of this theorem is similar to the proof of Theorem 2 in Herings et al. [10] and is therefore omitted.

2.5 An Illustration

Remember that individuals hold conjectures about how a mediator (or a player whose payoff is always zero) is going to choose a move among any set of feasible moves after any history. That is, each individual who has the possibility of moving after a certain history holds beliefs about the move chosen by the mediator (i.e., player 0) for each possible set of moves on which the individuals could agree to join. Then, given the conjecture of individual i about the others' behavior and her belief on the moves chosen by player 0 among any set of feasible moves, individual i chooses the behavior that maximizes her expected utility.

For the social environment of Fig. 1, social rationalizability with mediation works as follows. To simplify notation, we denote by $(1, 0, 1)$ for instance the behavior of player i when $b_i((x_1, \{1, 2\}) \mid (x_0)) = 1$, $b_i((x_2, \{1, 2\}) \mid (x_0)) = 0$, and $b_i((x_3, \{1, 2\}) \mid (x_0)) = 1$. In the first iteration, we can see that the behaviors $(0, 0, 0)$, $(1, 0, 0)$, and $(0, 1, 0)$ are never best responses whatever the conjecture of individual i about the behavior of individual j and whatever the belief on the choice of player 0. In fact, the behavior $(1, 0, 0)$ gives always a higher expected utility for player i than $(0, 0, 0)$, the behavior $(1, 1, 0)$ gives always a higher expected utility for player i than $(1, 0, 0)$, and the behavior $(0, 1, 1)$ gives always a higher expected utility for player i than $(0, 1, 0)$. However, the behavior $(1, 1, 0)$ cannot be eliminated since it is the unique best response against the conjecture that player j will have the behavior $(1, 0, 0)$ with probability $3/7$, the behavior $(0, 1, 0)$ with probability $3/7$, and the behavior $(1, 1, 1)$ with probability $1/7$, and assuming that the mediator chooses the move to the best outcome when the set of moves on which there is agreement is formed by the first two moves (i.e., the move to x_1 and to x_2), while she chooses the move to the worst outcome for any other set of possible agreements. In Table 1 we give conjectures against which each behavior b_i, different from the four behaviors already discussed, is the unique best response, assuming that the mediator only chooses the move to the best outcome when the set of possible agreements coincides with the moves that b_i does not block. The uniqueness of the best response guarantees that there are also cautious conjectures against which the behavior is the unique best response.

Hence, after the first iteration, we can only eliminate the behaviors $(0, 0, 0)$, $(1, 0, 0)$, and $(0, 1, 0)$. In the second iteration, we can show that the behavior $(1, 1, 0)$ is never a best response whatever the conjecture of individual i about the behaviors of individual j not eliminated in the first iteration, and whatever the belief on the choice of the mediator. Notice that the behavior $(0, 1, 1)$ gives always a weakly greater expected utility for player i than the behavior $(1, 1, 0)$ given that $(1, 0, 0)$ has been eliminated in the first iteration. For the other behaviors, it can be shown that there are conjectures about the behavior of individual j and beliefs on the choice of the mediator such that each of them is the unique best response against that conjecture and belief. In Table 2 we give conjectures on the behavior of player j against which each of these behaviors b_i is the unique best response, assuming that the mediator only chooses the move to the best outcome when the set of possible agreements coincides with the moves that b_i does not block when $b_i \neq (1, 1, 1)$. For $b_i = (1, 1, 1)$, the mediator only chooses the move to the best outcome when there is agreement on the moves to x_1 and x_2 or there is agreement on any move. As before, the uniqueness of the best response

Table 1 Unique best response and conjecture

b_j	b_i			
	(0, 0, 1)	(1, 0, 1)	(0, 1, 1)	(1, 1, 1)
(0, 0, 0)	0	0	0	0
(0, 0, 1)	$\frac{3}{4}$	$\frac{3}{7}$	$\frac{3}{7}$	$\frac{3}{10}$
(1, 0, 0)	0	$\frac{3}{7}$	0	$\frac{3}{10}$
(0, 1, 0)	0	0	$\frac{3}{7}$	$\frac{3}{10}$
(1, 1, 0)	0	0	0	0
(0, 1, 1)	0	0	0	0
(1, 0, 1)	0	0	0	0
(1, 1, 1)	$\frac{1}{4}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{10}$

Table 2 Unique best response and conjecture

b_j	b_i			
	(0, 0, 1)	(1, 0, 1)	(0, 1, 1)	(1, 1, 1)
(0, 0, 1)	$\frac{3}{4}$	0	0	0
(1, 1, 0)	0	$\frac{1}{7}$	$\frac{3}{7}$	$\frac{1}{2}$
(0, 1, 1)	0	0	$\frac{3}{7}$	0
(1, 0, 1)	0	$\frac{3}{7}$	0	0
(1, 1, 1)	$\frac{1}{4}$	$\frac{3}{7}$	$\frac{1}{7}$	$\frac{1}{2}$

guarantees that there are also cautious conjectures against which the behavior is the unique best response.

Next, in the third iteration, once individual i knows that player j will play a behavior that never blocks the move to x_3, her behavior $(0, 0, 1)$ will be the unique best response against any cautious conjecture about the behavior of player j and the mediator. Hence, $(0, 0, 1)$ is the unique socially rationalizable behavior and the Pareto-dominant outcome, x_3, is the unique socially rationalizable outcome.

The next two sections study the case of two players who can move to an arbitrary number of Pareto-ranked outcomes and the general case of more than two players who can move to an arbitrary number of Pareto-ranked outcomes.

3 Two-Player Coalitional Rationality

We investigate if social rationalizability with mediation satisfies, in general, the property of coalitional rationality. That is, in a situation in which a coalition of two or more individuals can move from a status quo x_0 to different outcomes that are Pareto ranked, does social rationalizability with mediation prescribe that players coordinate on the outcome that Pareto dominates all others? In the example of Fig. 1 with two players and three possible moves, we have seen that the Pareto-dominant outcome is the unique socially rationalizable outcome. Does this hold for the general case of two players and arbitrary Pareto-ranked payoffs?

For the case with two possible moves, using similar arguments as in the case with three possible moves, the behaviors $(0, 0)$ and $(1, 0)$ are eliminated in the first iteration and the

behavior $(1, 1)$ in the second iteration, leaving the Pareto-dominant outcome again as the unique socially rationalizable outcome. This section therefore focuses on the case with at least four possible moves.

Consider the social environment Γ^1, where $I = \{1, 2\}$, $Z = \{x_0, x_1, \ldots, x_K\}$ with $K \geq 4$, the outcomes are Pareto ranked,

$$u_i(x_K) > u_i(x_{K-1}) > \cdots > u_i(x_1) > u_i(x_0) = 0, \quad i \in I,$$

and the feasible moves are given by $x_0 \rightarrow_I x_k$, where $k = 1, \ldots, K$. We say that social rationalizability with mediation satisfies coalitional rationality if it selects the Pareto-dominant outcome x_K as the unique solution.

In the social environment Γ^1, we have, for every $i \in I$, $H_i = \{(x_0)\}$ and $M(x_0) = M_i(x_0) = \{(x_0x_1, I), (x_0x_2, I), \ldots, (x_0x_K, I)\}$. Since there is only one non-terminal history, in this section we drop histories from the notation for behaviors, conjectures, and utilities. A behavior of individual $i \in I$ is denoted by $b_i = (b_{i1}, \ldots, b_{iK})$, where, for $k \in \{1, \ldots, K\}$, $b_{ik} = b_i(x_0x_k, I)$. A behavior of player 0 is of the form $b_0 = (b_0(\overline{M}))_{\emptyset \neq \overline{M} \subseteq M}$ with $b_0(\overline{M}) \in \overline{M}$.

We introduce some additional notation. In this section, from now on, we fix an individual $i \in I$ and take j to be the other individual in I. Given $b_i \in B_i$, let $A_i(b_i) = \{m_i \in M_i \mid b_i(m_i) = 1\}$ be the set of moves on which individual i agrees and let $a_i(b_i) = \#A_i(b_i)$ be the cardinality of this set. For $b_i \in B_i$ with $a_i(b_i) \geq 1$, we define $\overline{k} = \max\{k \in \{1, \ldots, K\} \mid b_{ik} = 1\}$ and $\underline{k} = \min\{k \in \{1, \ldots, K\} \mid b_{ik} = 1\}$ as the number of the best and the worst outcome, respectively, on which individual i agrees. For $k \in \{1, \ldots, K\}$, we denote by $e(k)$ the individual behavior such that the kth component is 1 and the other components are 0, i.e., the individual only agrees with the move (x_0x_k, I), and by $\mathbf{1}$ the vector of all ones, that is, the behavior where the individual agrees to join every move.

We now show that coalitional rationality holds in general in the two-player social environment Γ^1. In order to do so, we use Lemmas 1–8. Lemma 1 states that if a behavior of individual i is the unique best response against a conjecture $c_i \in \Delta(\widetilde{B}_j) \times \Delta(B_0)$, where $\widetilde{B}_j$ is some non-empty subset of B_j, then it is also the unique best response against some cautious conjecture $c_i^* \in \Delta^\circ(\widetilde{B}_j) \times \Delta^\circ(B_0)$. The proof of Lemma 1 follows from the continuity of U^i and is left to the reader.

Lemma 1 *Take any $b_i \in B_i$. Let $\widetilde{B}_j$ be a non-empty subset of B_j. If there exists $c_i \in \Delta(\widetilde{B}_j) \times \Delta(B_0)$ such that, for every $b_i' \in B_i \setminus \{b_i\}$, $U_i(b_i, c_i) > U_i(b_i', c_i)$, then there exists $c_i^* \in \Delta^\circ(\widetilde{B}_j) \times \Delta^\circ(B_0)$ such that, for every $b_i' \in B_i \setminus \{b_i\}$, $U_i(b_i, c_i^*) > U_i(b_i', c_i^*)$.*

Lemma 2 claims that the individual behavior $b_i = (0, \ldots, 0)$, so individual i blocks all moves, is never a best response whatever the cautious conjecture $c_i \in \Delta^\circ(B_j) \times \Delta^\circ(B_0)$. Indeed, the behavior $b_i' = e(K)$, so b_i' is the same as b_i except that individual i joins the move to x_K, is always a strictly better response. All proofs that are not in the main text can be found in the "Appendix".

Lemma 2 *Let $b_i = (0, \ldots, 0)$. For $b_i' = e(K)$, for every $c_i \in \Delta^\circ(B_j) \times \Delta^\circ(B_0)$, it holds that $U_i(b_i', c_i) > U_i(b_i, c_i)$.*

Lemma 3 states that any individual behavior $b_i = e(k)$ with $k < K$, so individual i only agrees to join a single move different from the move to x_K, is never a best response whatever the cautious conjecture $c_i \in \Delta^\circ(B_j) \times \Delta^\circ(B_0)$. Indeed, the behavior b_i', where b_i' is the same as b_i except that individual i joins the move to x_{k+1} is always a strictly better response.

Lemma 3 *Let $b_i = e(k)$ for some $k < K$. For $b_i' = e(k) + e(k+1)$, for every $c_i \in \Delta^\circ(B_j) \times \Delta^\circ(B_0)$, it holds that $U_i(b_i', c_i) > U_i(b_i, c_i)$.*

Lemma 4 establishes that for any behavior b_i where individual i agrees to move to at least two outcomes or to move only to x_K there exists a conjecture $c_i \in \Delta(B_j) \times \Delta(B_0)$ such that b_i is her unique best response. This conjecture is such that for every k such that $b_{ik} = 1$ it puts positive weight on $b_j = e(k)$ as well as on $b_j = \mathbf{1}$ and puts zero weight on any other behavior. The positive weights on $b_j = e(k)$ guarantee that b_i gives higher utility than a behavior b_i' which blocks moves that are not blocked by b_i. The positive weight on $b_j = \mathbf{1}$, together with a suitably chosen conjecture on the behavior of the mediator, implies that b_i outperforms any b_i' that agrees to strictly more moves than b_i.

Lemma 4 *Take any $b_i \in B_i$ such that either $a_i(b_i) \geq 2$ or $b_i = e(K)$. Then, for all $b_i' \in B_i \setminus \{b_i\}$, we have $U_i(b_i, c_i) > U_i(b_i', c_i)$, where $c_i \in \Delta(B_j) \times \Delta(B_0)$ is such that*

$$c_i^j(b_j) = \begin{cases} \frac{u_i(x_K)}{[a_i(b_i)\cdot u_i(x_K)+u_i(x_1)]} & \text{if there is } k \in \{1,\dots,K\} \text{ such that } b_j = e(k) \text{ and } b_{ik} = 1 \\ \frac{u_i(x_1)}{[a_i(b_i)\cdot u_i(x_K)+u_i(x_1)]} & \text{if } b_j = \mathbf{1} \\ 0 & \text{otherwise} \end{cases}$$

and $c_i^0(b_0) = 1$ where b_0 selects the move to the best outcome in $\overline{M}$ if $\overline{M} = A_i(b_i)$ and the move to the worst outcome in $\overline{M}$ in any other non-empty set $\overline{M} \subseteq M$.

From Lemmas 2, 3, and 4 we have that $P_i^1 = \{b_i \in B_i \mid a_i(b_i) \geq 2\} \cup \{e(K)\}$.

For the second iteration, we first show that the behavior $b_i = e(1) + e(2) \in P_i^1$ does not belong to P_i^2. In fact, the behavior $b_i' = (0, 1, \dots, 1) \in P_i^1$ gives higher utility than b_i against all relevant cautious conjectures.

Lemma 5 *Consider the behavior $b_i = e(1) + e(2) \in P_i^1$. Take the behavior $b_i' = (0, 1, \dots, 1) \in P_i^1$. Then, for every $c_i \in \Delta^\circ(P_j^1) \times \Delta^\circ(B_0)$, we have $U_i(b_i', c_i) > U_i(b_i, c_i)$.*

We continue by showing that any behavior $b_i \in P_i^1$ different from $e(1) + e(2)$ belongs to P_i^2, i.e., is the best response of player i in P_i^1 against some cautious conjecture in $\Delta^\circ(P_j^1) \times \Delta^\circ(B_0)$. We achieve this by showing that b_i is the unique best response against a particular conjecture $c_i \in \Delta(P_j^1) \times \Delta(B_0)$.

Lemma 6 *Let $b_i \in P_i^1 \setminus \{e(1) + e(2)\}$. Then, for all $b_i' \in P_i^1 \setminus \{b_i\}$, we have $U_i(b_i, c_i) > U_i(b_i', c_i)$, where, for $\varepsilon > 0$ sufficiently small, $c_i \in \Delta(P_j^1) \times \Delta(B_0)$ is such that*

$$c_i^j(b_j) = \begin{cases} 1 - \varepsilon - \varepsilon^2 & \text{if } b_j = b_i \\ \varepsilon & \text{if } b_j = \mathbf{1} \\ \varepsilon^2 & \text{if } b_j = e(1) + e(\max\{2, \underline{k}\}) \end{cases}$$

and $c_i^0(b_0) = 1$ where b_0 selects the move to the best outcome in $\overline{M}$ if $\overline{M} = A_i(b_i)$ and the move to the worst outcome in $\overline{M}$ for any other non-empty set $\overline{M} \subseteq M$.

From Lemmas 5 and 6, it follows that $P_i^2 = P_i^1 \setminus \{e(1) + e(2)\}$. The set P_i^2 consists of the behaviors b_i such that $a_i(b_i) \geq 2$ and $\overline{k} \geq 3$ as well as the behavior $b_i = e(K)$.

Lemma 7 shows that the behaviors $b_i \in P_i^2$ that block the moves to the best $K - 3$ outcomes do not belong to P_i^3. The behavior $b_i' = (0, 0, 1, \dots, 1) \in P_i^2$ gives a greater utility against any cautious conjecture in $\Delta^\circ(P_j^2) \times \Delta^\circ(B_0)$.

Lemma 7 *Consider a behavior $b_i \in P_i^2$ such that $\overline{k} = 3$. Take the behavior $b_i' = (0, 0, 1, \ldots, 1) \in P_i^2$. Then, for every $c_i \in \Delta^\circ(P_j^2) \times \Delta^\circ(B_0)$, we have $U_i(b_i', c_i) > U_i(b_i, c_i)$.*

We continue by showing that any behavior $b_i \in P_i^2$ such that $\overline{k} \geq 4$ belongs to P_i^3, i.e., is the best response of player i in P_i^2 against some cautious conjecture in $\Delta^\circ(P_j^2) \times \Delta^\circ(B_0)$. We achieve this by showing that b_i is the unique best response against a particular conjecture $c_i \in \Delta(P_j^2) \times \Delta(B_0)$.

Lemma 8 *Let $b_i \in P_i^2$ be such that $\overline{k} \geq 4$. Then, for all $b_i' \in P_i^2 \setminus \{b_i\}$, we have $U_i(b_i, c_i) > U_i(b_i', c_i)$, where, for $\varepsilon > 0$ sufficiently small, $c_i \in \Delta(P_j^2) \times \Delta(B_0)$ is such that*

$$c_i^j(b_j) = \begin{cases} 1 - \varepsilon - \varepsilon^2 & \text{if } b_j = b_i \\ \varepsilon & \text{if } b_j = \mathbf{1} \\ \varepsilon^2 & \text{if } b_j = e(1) + e(2) + e(\max\{3, \underline{k}\}) \end{cases}$$

and $c_i^0(b_0) = 1$ where b_0 selects the move to the best outcome in $\overline{M}$ if $\overline{M} = A_i(b_i)$ and the move to the worst outcome in $\overline{M}$ for any other non-empty set $\overline{M} \subseteq M$.

Hence, by Lemmas 7 and 8 we have that every behavior b_i in P_i^3 is such that $a_i(b_i) \geq 2$ and $\overline{k} \geq 4$ or $b_i = e(K)$. Proceeding in this way, we obtain the following proposition.

Proposition 1 *For $1 \leq k \leq K - 1$, it holds that $P_i^k = \{b_i \in B_i \mid a_i(b_i) \geq 2 \text{ and } \overline{k} \geq k + 1\} \cup \{e(K)\}$.*

Proof The proposition has already been shown for $k = 1, 2, 3$. Assume the proposition is true for some $k \leq K - 2$. We show the proposition to hold for $k + 1$.

We eliminate any behavior $b_i \in P_i^k$ such that $\overline{k} = k+1$ by the behavior $b_i' = \sum_{\ell=\overline{k}+1}^{K} e(\ell)$. The proof follows the steps of the proof of Lemma 7.

The other behaviors b_i in P_i^k are such that $a_i(b_i) \geq 2$ and $\overline{k} > k + 1$ or $b_i = e(K)$. Such a behavior b_i is the unique best response, for $\varepsilon > 0$ sufficiently small, against the conjecture $c_i \in \Delta(P_j^k) \times \Delta(B_0)$ defined by

$$c_i^j(b_j) = \begin{cases} 1 - \varepsilon - \varepsilon^2 & \text{if } b_j = b_i \\ \varepsilon & \text{if } b_j = \mathbf{1} \\ \varepsilon^2 & \text{if } b_j = \sum_{\ell=1}^{k} e(\ell) + e(\max\{k + 1, \underline{k}\}) \end{cases}$$

and $c_i^0(b_0) = 1$ where b_0 selects the move to the best outcome in $\overline{M}$ if $\overline{M} = A_i(b_i)$ and the move to the worst outcome in $\overline{M}$ for any other non-empty set $\overline{M} \subseteq M$. The proof follows the steps of the proof of Lemma 8. □

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

Theorem 2 *Consider the social environment Γ^1. There is a unique behavior of individual i that is socially rationalizable, $P_i^K = P_i^\infty = \{e(K)\}$.*

Proof From Proposition 1, we have $P_i^{K-1} = \{b_i \in B_i \mid \overline{k} = K\}$. Finally, for every $c_i \in \Delta^\circ(P_j^{K-1}) \times \Delta^\circ(B_0)$, the behavior $b_i = e(K)$ gives to individual i a utility equal to $U_i(b_i, c_i) = u_i(x_K)$. For every $b_i' \in P_i^{K-1} \setminus \{b_i\}$, for every $c_i \in \Delta^\circ(P_j^{K-1}) \times \Delta^\circ(B_0)$, $U_i(b_i', c_i) < u_i(x_K)$ because for some $k < K$, $b_{ik}' = 1$, and the cautiousness of c_i implies that with positive probability the opponent of i follows a behavior b_j such that $b_{jk} = 1$ and the mediator chooses $b_0(f(b_i, b_j)) = x_k$, which leads to utility $u_i(x_k) < u_i(x_K)$. So, $P_i^K = \{e(K)\} = P_i^\infty$. □

The above result implies that social rationalizability with mediation satisfies the property of two-player coalitional rationality. When the outcomes can be Pareto ranked, a coalition of two players always 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^1. *We have* $Z^\infty(x_0) = \{x_K\}$.

Throughout the paper we consider the case where players hold uncorrelated conjectures about the behavior of the other players and the mediator. It can be verified that the conclusion of Corollary 1 remains valid when players are allowed to hold correlated conjectures. Whenever a behavior is eliminated from P_i^k for some k, this is achieved by a behavior where the worst possible agreement is at least as good as the best possible agreement under the eliminated behavior. Such elimination still takes place when conjectures are allowed to be correlated. Whenever a behavior survives against uncorrelated conjectures, it remains to do so against correlated conjectures, since uncorrelated conjectures are a special case of correlated ones.

4 Coalitional Rationality for More Than Two Players

Does social rationalizability with mediation satisfy, in general, the property of coalitional rationality? We now provide an example of a social environment with three players which violates this property.

Example 2 Consider the social environment Γ^2 in which the coalition of three individuals may decide to move from the status quo x_0, where they all get a utility equal to 0, to outcome x_1 obtaining each 1 unit of utility, or to outcome x_2 all getting 2 units of utility, or to outcome x_3 and receive 3 units of utility each. That is, $I = \{1, 2, 3\}$, $Z = \{x_0, x_1, x_2, x_3\}$, for every $k \in \{1, 2, 3\}$, $x_0 \rightarrow_I x_k$ are the only possible moves, and, for every $i \in I$, for every $k \in \{0, 1, 2, 3\}$, $u_i(x_k) = k$.

In the social environment Γ^2, we have, for every $i \in I$, $H_i = \{(x_0)\}$ and $M_i(x_0) = M(x_0) = \{(x_0x_1, I), (x_0x_2, I), (x_0x_3, I)\}$. As in Sect. 3, since there is only one non-terminal history, in this section we drop histories from the notation for behaviors, conjectures, and utilities. In this section, from now on, we fix an individual $i \in I$.

By Definition 2, $P_i^0 = B_i$. We show that the behaviors $(0, 0, 0)$, $(1, 0, 0)$, and $(0, 1, 0)$ do not belong to P_i^1. Let b_i be any such behavior. Take $b_i' \in B_i$ such that $b_i' = b_i + (0, 0, 1)$. It is quite straightforward that for all $c_i \in \prod_{j \in I \setminus \{i\}} \Delta^\circ(B_j) \times \Delta^\circ(B_0)$, $U_i(b_i, c_i) < U_i(b_i', c_i)$. Indeed, the behaviors b_i and b_i' give the same payoff to individual i against the opponents' behaviors b_{-i} whenever the set of moves on which the opponents of individual i agree does not include the move to outcome x_3, i.e., when some opponents behavior b_j is such that $b_{j3} = 0$. But, b_i' does strictly better than b_i against the opponents' behaviors b_{-i} such that $b_{j3} = 1$ for all $j \in I \setminus \{i\}$ and a mediator that chooses the move to x_3 whenever that move belongs to $\overline{M}$.

Next, we show that for every $b_i \in B_i \setminus \{(0, 0, 0), (1, 0, 0), (0, 1, 0)\}$ there exists $c_i \in \prod_{j \in I \setminus \{i\}} \Delta(B_j) \times \Delta(B_0)$ such that b_i is the unique best response against c_i. We can extend Lemma 1 to the setting of Example 2 and conclude that $b_i \in P_i^1$.

In fact, we can use conjectures that are similar to the ones used in Lemma 4. We define the conjecture $c_i \in \prod_{j\in I\setminus\{i\}} \Delta(P_j^1) \times \Delta(B_0)$ by

$$c_i^j(b_j) = \begin{cases} \frac{u_i(x_K)}{[a_i(b_i)\cdot u_i(x_K)+u_i(x_1)]} & \text{if there is } k \in \{1,\dots,K\} \text{ such that } b_j = e(k) \text{ and } b_{ik} = 1 \\ \frac{u_i(x_1)}{[a_i(b_i)\cdot u_i(x_K)+u_i(x_1)]} & \text{if } b_j = \mathbf{1} \\ 0 & \text{otherwise} \end{cases}$$

and $c_i^0(b_0) = 1$ where b_0 selects the move to the best outcome in $\overline{M}$ if $\overline{M} = A_i(b_i)$ and the move to the worst outcome in $\overline{M}$ in any other non-empty set $\overline{M} \subseteq M$.

The possible sets of moves on which both opponents agree are respectively equal to $\emptyset$, for any $k \in \{1,\dots,K\}$ such that $b_{ik} = 1$, $\{(x_0x_k, I)\}$, and M. The definition of b_0 implies that under b_i the move to the best feasible outcome results, irrespective of the realization of $\overline{M}$. Any other behavior b_i' that is a best response should therefore also result in the move to the best feasible outcome. Since opponents may only agree on the move to $\{(x_0x_k, I)\}$ for any $k \in \{1,\dots,K\}$ such that $b_{ik} = 1$, this implies that $A_i(b_i) \subseteq A_i(b_i')$. Since the opponents may also agree on the set of all moves and $b_i' \neq b_i$, b_0 selects the move to the worst outcome in $A_i(b_i')$ in this case. Since $b_i \in B_i \setminus \{(0,0,0), (1,0,0), (0,1,0)\}$ and $A_i(b_i) \subseteq A_i(b_i')$, the move to the worst outcome in $A_i(b_i')$ is inferior to the move to the best outcome in $A_i(b_i)$. This shows that b_i is the unique best response to c_i. Hence, $P_i^1 = \{(0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)\}$.

In the second iteration, individual i knows that any other individual j will play a behavior in P_j^1. We continue by defining for each behavior $b_i \in P_i^1$ a conjecture $c_i \in \prod_{j\in I\setminus\{i\}} \Delta(P_j^1) \times \Delta(B_0)$ such that b_i is the unique best response against c_i.

(i) The behavior $b_i = (0,0,1)$ is the unique best response against the conjecture c_i such that, for $j \in I \setminus \{i\}$,

$$c_i^j(b_j) = \begin{cases} 3/4 \text{ if } b_j = (0,0,1) \\ 1/4 \text{ if } b_j = \mathbf{1} \end{cases}$$

and $c_i^0(b_0) = 1$ where b_0 is such that the move to the worst outcome in $\overline{M}$ is selected for any non-empty set $\overline{M} \subseteq M$. Indeed, for every $b_i' \in P_1^i \setminus \{b_i\}$, we have that $U_i(b_i, c_i) = 3 > U_i(b_i', c_i)$.

(ii) Let $I \setminus \{i\} = \{j, j'\}$. The behavior $b_i = (1,1,0)$ is the unique best response against the conjecture c_i such that

$$c_i^j(b_j) = \begin{cases} 7/8 \text{ if } b_j = (1,1,0) \\ 1/8 \text{ if } b_j = \mathbf{1} \end{cases}$$

$$c_i^{j'}(b_{j'}) = \begin{cases} 7/8 \text{ if } b_{j'} = (1,0,1) \\ 1/8 \text{ if } b_{j'} = \mathbf{1} \end{cases}$$

and $c_i^0(b_0) = 1$ where b_0 selects the move to the best outcome in $\overline{M}$ if $\overline{M} = A_i(b_i)$ and the move to the worst outcome in $\overline{M}$ in any other non-empty set $\overline{M} \subseteq M$. Indeed, for every $b_i' \in P_1^i \setminus \{b_i\}$, we have that $U_i(b_i, c_i) = 72/64 > U_i(b_i', c_i)$.

(iii) Let $I \setminus \{i\} = \{j, j'\}$. The behavior $b_i = (1,0,1)$ is the unique best response against the conjecture c_i such that

$$c_i^j(b_j) = \begin{cases} 7/8 \text{ if } b_j = (1,0,1) \\ 1/8 \text{ if } b_j = \mathbf{1} \end{cases}$$

$$c_i^{j'}(b_{j'}) = \begin{cases} 7/8 \text{ if } b_{j'} = (1,1,0) \\ 1/8 \text{ if } b_{j'} = \mathbf{1} \end{cases}$$

and $c_i^0(b_0) = 1$ where b_0 selects the move to the best outcome in $\overline{M}$ if $\overline{M} = A_i(b_i)$ and the move to the worst outcome in $\overline{M}$ in any other non-empty set $\overline{M} \subseteq M$. Indeed, for every $b_i' \in P_1^i \setminus \{b_i\}$, we have that $U_i(b_i, c_i) = 80/64 > U_i(b_i', c_i)$.

(iv) Let $I \setminus \{i\} = \{j, j'\}$. The behavior $b_i = (0, 1, 1)$ is the unique best response against the conjecture c_i such that

$$c_i^j(b_j) = \begin{cases} 7/8 \text{ if } b_j = (0,1,1) \\ 1/8 \text{ if } b_j = \mathbf{1} \end{cases}$$

$$c_i^{j'}(b_{j'}) = \begin{cases} 7/8 \text{ if } b_{j'} = (1,1,0) \\ 1/8 \text{ if } b_{j'} = \mathbf{1} \end{cases}$$

and $c_i^0(b_0) = 1$ where b_0 selects the move to the best outcome in $\overline{M}$ if $\overline{M} = A_i(b_i)$ and the move to the worst outcome in $\overline{M}$ in any other non-empty set $\overline{M} \subseteq M$. Indeed, for every $b_i' \in P_1^i \setminus \{b_i\}$, we have that $U_i(b_i, c_i) = 136/64 > U_i(b_i', c_i)$.

(v) The behavior $b_i = (1, 1, 1)$ is the unique best response against the conjecture c_i such that, for $j \in I \setminus \{i\}$,

$$c_i^j(b_j) = \begin{cases} 6/13 \text{ if } b_j = (1,1,0) \\ 3/13 \text{ if } b_j = (1,0,1) \\ 3/13 \text{ if } b_j = (0,1,1) \\ 1/13 \quad \text{ if } b_j = \mathbf{1} \end{cases}$$

and $c_i^0(b_0) = 1$ where b_0 selects the move to the best outcome in $\overline{M}$ in any non-empty set $\overline{M} \subseteq M$. Indeed, for every $b_i' \in P_1^i \setminus \{b_i\}$, we have that $U_i(b_i, c_i) = 351/169 > U_i(b_i', c_i)$.

We find that $P_i^1 = P_i^2 = P_i^\infty$. The set of socially rationalizable outcomes with mediation coincides with the set of initial outcomes, $Z^\infty(x_0) = \{x_0, x_1, x_2, x_3\}$. Therefore, social rationalizability with mediation does not satisfy the property of coalitional rationality when the number of players is greater than two. This conclusion remains valid when one allows for correlated beliefs, since such beliefs can only expand the set of socially rationalizable outcomes.

5 Discussion

5.1 An Equivalent Definition of Social Rationalizability

An alternative definition of social rationalizability with mediation is obtained by adapting Battigalli's [1] notion of extensive-form rationalizability to social environments. Social rationalizability based on the approach of Battigalli is derived from two assumptions: (1) individuals are rational and endowed with a hierarchy of hypotheses, and (2) this is common knowledge at the initial status quo. In Definition 3, R_i^1 is the set of individual behaviors of $i \in I$ that are individually rational. Higher degrees of rationality are constructed recursively.

Definition 3 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 a consistent updating system c_i such that

(i) For all $h' \in H_i(J)$, $c_i(h') \in \prod_{j \neq i} \Delta^\circ(R_j^{k^*}) \times \Delta^\circ(B_0)$ where k^* is the maximal element in $\{0, 1, \ldots, n-1\}$ such that $R_{-i}^{k^*}$ allows for h',

(ii) For all $h' \in H_i(J)$, if b_i allows for h', then b_i is a best response to $c_i(h')$ at h', that is, for all $\widehat{b}_i \in B_i$, $U_i(h')(b_i, c_i) \geq U_i(h')(b_i/\widehat{b}_i^{h'}, c_i)$, 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.

The sequence $(R_j^1)_{j\neq i}, (R_j^2)_{j\neq i}, (R_j^3)_{j\neq i}, \ldots$ in Definition 3 represents for individual i a hierarchy of increasingly strong hypotheses about the behavior of individuals $j \neq i$. 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 3) and optimizes accordingly.

Theorem 3 *For all $n \geq 0$, $R^n = P^n$.*

Theorem 3 claims that both definitions of social rationalizability are equivalent. The proof of this theorem is similar to the proof of Theorem 1 in Battigalli [1] and is therefore omitted. From Theorem 3, we have that $R^\infty(J) = P^\infty(J)$. Notice that the computation of the set of socially rationalizable outcomes is greatly simplified by using the reduction procedure of Definition 2.

5.2 A Permutational Mediator

Assume the mediator, player 0, is known to behave as follows. A behavior $b_0 = (b_0(\cdot \mid h))_{h\in H(J)}$ of player 0 is such that after each history h she chooses a permutation of $M(h)$ that indicates the order according to which moves are implemented. For $\overline{M} \in \mathcal{M}(h)$, the highest ordered element in $\overline{M}$ according to this permutation is implemented. We refer to such a mediator as a *permutational mediator*. We demonstrate next that for the social environment Γ^2 of Example 2 it is not possible that $(1, 1, 0)$ survives the first round of elimination in case of a permutational mediator.

Example 3 Consider the social environment Γ^2 of Example 2. The conjecture against which $b_i = (1, 1, 0)$ is the best response is such that the mediator selects the move to the best outcome in $\overline{M}$ if $\overline{M} = A_i(b_i)$ and the move to the worst outcome in $\overline{M}$ in any other nonempty set $\overline{M} \subseteq M$. Such a mediator cannot be permutational, as a permutational mediator that selects the move to x_2 when there is agreement on both the move to x_1 and the move to x_2 cannot select the move to x_1 when there is agreement on all moves. The behavior $b_i' = (1, 1, 1)$ is at least as good against any conjecture than b_i and strictly better against some conjectures when there is a permutational mediator. Indeed, when the opponents agree on a set containing the move to x_3, then under behavior b_i' a permutational mediator either selects the move to x_3 or the same move as under behavior b_i, and when the opponents agree on a set not containing the move to x_3, then a permutational mediator selects the same move under b_i' and b_i.

Consider the social environment Γ^3 where I contains a finite number of individuals, $Z = \{x_0, x_1, \ldots, x_K\}$, and there is one outcome which strictly Pareto dominates all other outcomes,

$$u_i(x_K) > u_i(x_k) > u_i(x_0), \quad i \in I,\ k \in \{1, \ldots, K-1\}.$$

The possible moves are given by $x_0 \to_I x_k$ for $k = 1, \ldots, K$. We say that social rationalizability with mediation satisfies coalitional rationality if it selects the Pareto-dominant outcome x_K.

Theorem 4 *Consider the social environment Γ^3 with a permutational mediator. There is a unique behavior of individual $i \in I$ that is socially rationalizable, $P_i^\infty = \{e(K)\}$.*

The proof of this theorem is similar to the proof of Theorem 3 in Herings et al. [9] and Theorem 6 in Herings et al. [18] and is therefore omitted.

Corollary 2 *Consider the social environment Γ^3 with a permutational mediator. There is a unique socially rationalizable outcome, $Z^\infty = \{x_K\}$.*

It can be shown that, in the case of two players, social rationalizability with mediation requires $K-2$ additional rounds of elimination to obtain coordination on the Pareto-dominant outcome compared to social rationalizability with a permutational mediator.

In the case of the social environment Γ^2 (Example 1), social rationalizability with a permuational mediator satisfies the property of coalitional rationality while social rationalizability with mediation does not. The reason behind this fact is that once $P_i^1 = \{(0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)\}$, the behavior $(1,1,0)$ that blocks the move to the Pareto-dominant outcome cannot be eliminated when the mediator can arbitrarily select different moves for different sets of possible agreements. On the contrary, a permutational mediator holds a ranking over the feasible moves and chooses, for any set of possible agreements, the agreement that is ranked highest.

Finally, instead of having a permutational mediator, one could simply 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. Assuming that the implementability prior-beliefs of the individuals are cautious, Herings et al. [9] show that social rationalizability ends up selecting the Pareto-dominant outcome.

5.3 Conclusion

Social environments constitute a framework in which it is possible to study how groups of agents interact in a society. We have proposed a new solution concept for social environments that is based on individual rationality, called social rationalizability with mediation. 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 have conjectures about how a mediator, a player whose payoff is always zero, is going to solve the conflicts of interest.

Social rationalizability with mediation identifies which coalitions are likely to form and which outcomes might occur when the individuals are rational and this is common knowledge at the initial status quo. We have shown that for all social environments the set of socially rationalizable outcomes with mediation is non-empty. Social rationalizability with mediation 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, we have analyzed if social rationalizability with mediation is 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. We have shown that

social rationalizability with mediation does not satisfy the property of coalitional rationality for coalitions of more than two players. In fact, restrictions on the behavior of the mediator are needed to guarantee that individuals coordinate on the Pareto-dominant outcome. So, coalitional rationality does not necessarily follow from individual behaviors of rational individuals.

We have made our point in the simplest social environment possible that enables us to study coalitional rationality. There is an initial status quo that is Pareto dominated by a number of Pareto-ranked alternatives and the grand coalition can move from the initial status quo to any of the other states. An interesting variation would be where the grand coalition can move between any two states. We expect that our conclusions survive in such a more complicated social environment.

We have treated the mediator as an unbiased player 0. Our analysis can be extended to cover the situation where the mediator is favoring one of the players. This is not necessarily detrimental for the other players. For instance, in the social environments with a common interest as studied in this paper, the mediator would behave like a permutational mediator for one particular permutation, and coordination on the Pareto-dominant outcome would be easier to sustain. What would happen in social environments without a common interest is an interesting question for future research.

Acknowledgements Ana Mauleon and Vincent Vannetelbosch are, respectively, Research Director and Senior Research Associate of the National Fund for Scientific Research (FNRS). Financial support from the Fonds de la Recherche Scientifique - FNRS research Grant T.0143.18 is gratefully acknowledged.

Appendix

Proof of Lemma 2 (i) For all $b_j \in B_j$ with $b_{jK} = 0$, we have that

$$U_i\left(b_i', b_j, c_i^0\right) = U_i\left(b_i, b_j, c_i^0\right) = 0.$$

(ii) For all $b_j \in B_j$ with $b_{jK} = 1$, we have that

$$u_i(x_K) = U_i\left(b_i', b_j, c_i^0\right) > U_i\left(b_i, b_j, c_i^0\right) = 0.$$

It follows that, for every $c_i \in \Delta^\circ(B_j) \times \Delta^\circ(B_0)$, $U_i(b_i', c_i) > U_i(b_i, c_i)$.

□

Proof of Lemma 3 (i) For all $b_j \in B_j$ with $b_{jk} = b_{jk+1} = 0$, we have

$$U_i\left(b_i, b_j, c_i^0\right) = U_i\left(b_i', b_j, c_i^0\right) = u_i(x_0).$$

(ii) For all $b_j \in B_j$ with $b_{jk} = 1$ and $b_{jk+1} = 0$, we have

$$U_i\left(b_i, b_j, c_i^0\right) = U_i\left(b_i', b_j, c_i^0\right) = u_i(x_k).$$

(iii) For all $b_j \in B_j$ with $b_{jk} = 0$ and $b_{jk+1} = 1$, we have

$$U_i\big(b_i, b_j, c_i^0\big) = 0 < u_i(x_{k+1}) = U_i\big(b_i', b_j, c_i^0\big).$$

(iv) For all $b_j \in B_j$ with $b_{jk} = b_{jk+1} = 1$, we have

$$U_i\big(b_i, b_j, c_i^0\big) = u_i(x_k) < U_i\big(b_i', b_j, c_i^0\big).$$

Hence, for every $c_i \in \Delta^\circ(B_j) \times \Delta^\circ(B_0)$, it holds that $U_i(b_i', c_i) > U_i(b_i, c_i)$.

□

Proof of Lemma 4 (i) Consider the behavior $b_i = e(K)$. Then,

$$U_i(b_i, c_i) = \left[\frac{u_i(x_K)}{u_i(x_K) + u_i(x_1)}\right] \cdot u_i(x_K) + \left[\frac{u_i(x_1)}{u_i(x_K) + u_i(x_1)}\right] \cdot u_i(x_K) = u_i(x_K).$$

For any $b_i' \in B_i \setminus \{e(K)\}$ it holds that

$$U_i(b_i', c_i) \leq \left[\frac{u_i(x_K)}{u_i(x_K) + u_i(x_1)}\right] \cdot u_i(x_K) + \left[\frac{u_i(x_1)}{u_i(x_K) + u_i(x_1)}\right] \cdot u_i(x_{K-1}) < u_i(x_K),$$

where the expression after the weak inequality uses the fact that $f(b_i', \mathbf{1})$ contains a move leading to an outcome different from x_K, so b_0 selects a move leading to an outcome worse than x_K, or $f(b_i', \mathbf{1})$ is equal to the empty set and outcome x_0 results.

(ii) Consider any behavior $b_i \in B_i$ such that $a_i(b_i) \geq 2$. We have that

$$U_i(b_i, c_i) = \left[\tfrac{u_i(x_K)}{a_i(b_i)\cdot u_i(x_K) + u_i(x_1)}\right] \cdot \sum_{\{k \in \{1,\ldots,K\} \mid b_{ik} = 1\}} u_i(x_k) + \left[\tfrac{u_i(x_1)}{a_i(b_i)\cdot u_i(x_K) + u_i(x_1)}\right] \cdot u_i(x_{\underline{k}}).$$

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

Case 1. It follows that

$$\begin{aligned} U_i(b_i', c_i) &\leq \left[\tfrac{u_i(x_K)}{a_i(b_i)\cdot u_i(x_K) + u_i(x_1)}\right] \cdot \left[\textstyle\sum_{\{k \in \{1,\ldots,K\} \mid b_{ik} = 1\}} u_i(x_k) - u_i(x_{\underline{k}})\right] \\ &\quad + \left[\tfrac{u_i(x_1)}{a_i(b_i)\cdot u_i(x_K) + u_i(x_1)}\right] \cdot u_i(x_K) \\ &\leq \left[\tfrac{u_i(x_K)}{a_i(b_i)\cdot u_i(x_K) + u_i(x_1)}\right] \cdot \left[\textstyle\sum_{\{k \in \{1,\ldots,K\} \mid b_{ik} = 1\}} u_i(x_k)\right] \\ &< U_i(b_i, c_i). \end{aligned}$$

Case 2. It holds that

$$\begin{aligned} U_i(b_i', c_i) &\leq \left[\tfrac{u_i(x_K)}{a_i(b_i)\cdot u_i(x_K) + u_i(x_1)}\right] \cdot \left[\textstyle\sum_{\{k \in \{1,\ldots,K\} \mid b_{ik} = 1\}} u_i(x_k)\right] \\ &\quad + \left[\tfrac{u_i(x_1)}{a_i(b_i)\cdot u_i(x_K) + u_i(x_1)}\right] \cdot u_i(x_{\underline{k}}) \\ &< U_i(b_i, c_i), \end{aligned}$$

where the expression after the weak inequality uses the fact that $f(b_i', \mathbf{1})$ is not equal to $f(b_i, \mathbf{1})$ and contains the move to outcome $x_{\underline{k}}$ as an element, so b_0 selects a move leading to an outcome which is at best equal to $x_{\underline{k}}$. □

Proof of Lemma 5 Since the behaviors $(0, \ldots, 0)$ and $e(1)$ of individual j do not belong to P_j^1, it follows that, for every conjecture $c_i \in \Delta^\circ(P_j^1) \times \Delta^\circ(B_0)$,

$$U_i\big(b_i', c_i\big) \geq u_i(x_2) \geq U_i(b_i, c_i).$$

Since such a conjecture c_i puts positive weight on the behavior $b_j = e(K)$ and

$$U_i\big(b_i', e(K), c_i^0\big) = u_i(x_K) > u_i(x_0) = U_i\big(b_i, e(K), c_i^0\big),$$

we conclude that $U_i(b_i', c_i) > U_i(b_i, c_i)$. □

Proof of Lemma 6 It holds that

$$U_i(b_i, c_i) = \big(1 - \varepsilon - \varepsilon^2\big) \cdot u_i(x_{\overline{k}}) + \varepsilon \cdot u_i(x_{\overline{k}}) + \varepsilon^2 \cdot u_i(x_{\underline{k}}),$$

where for $\underline{k} = 1$ the expression in the last term follows from the fact that

$$(x_0 x_1, I) \in f(b_i, e(1) + e(2)) = f(b_i, e(1) + e(\max\{2, \underline{k}\})) \neq A_i(b_i),$$

so b_i selects the worst move in $f(b_i, e(1) + e(2))$, which is equal to $(x_0 x_1, I)$.

Let $b_i' \in P_i^1 \setminus \{b_i\}$.

If $b_{i\overline{k}}' = 0$, then, for $\varepsilon > 0$ sufficiently small,

$$U_i\big(b_i', c_i\big) \leq \big(1 - \varepsilon - \varepsilon^2\big) \cdot u_i\big(x_{\overline{k}-1}\big) + \big(\varepsilon + \varepsilon^2\big) \cdot u_i(x_K) < U_i(b_i, c_i),$$

where the strict inequality makes use of the fact that ε is sufficiently small.

Assume $b_{i\overline{k}}' = 1$. If there is $k < \overline{k}$ such that $b_{ik}' = 1$, then, for $\varepsilon > 0$ sufficiently small,

$$U_i(b_i', c_i) \leq \big(1 - \varepsilon - \varepsilon^2\big) \cdot u_i\big(x_{\overline{k}}\big) + \varepsilon \cdot u_i\big(x_{\overline{k}-1}\big) + \varepsilon^2 \cdot u_i(x_K) < U_i(b_i, c_i),$$

where the strict inequality makes use of the fact that ε is sufficiently small.

Let the smallest k for which $b_{ik}' = 1$ be equal to $\overline{k}$. It follows that $b_i \neq e(K)$, since $\overline{k} = K$ together with the assumption that the smallest k for which $b_{ik}' = 1$ is equal to $\overline{k}$ implies $b_i' = e(K)$. Since $b_i' \neq b_i$ we have that $b_i \neq e(K)$. Since $b_i \in P_i^1$ and $b_i \neq e(1) + e(2)$, it also follows that $\overline{k} \geq 3$. We have that

$$U_i(b_i', c_i) = \big(1 - \varepsilon - \varepsilon^2\big) \cdot u_i\big(x_{\overline{k}}\big) + \varepsilon \cdot u_i\big(x_{\overline{k}}\big) < U_i(b_i, c_i),$$

where the second term in the expression after the equality follows from the fact that $f(b_i', \mathbf{1}) = A_i(b_i') \neq A_i(b_i)$, so the worst move $(x_0 x_{\overline{k}}, I)$ in $A_i(b_i')$ is selected by b_0. The expression after the equality also uses that $b_i \neq e(K)$, so $\underline{k} < \overline{k}$, and $f(b_i', e(1) + e(\max\{2, \underline{k}\})) = \emptyset$. □

Proof of Lemma 7 Since, for every behavior $b_j \in P_j^2$, there is $k \geq 3$ such that $b_{jk} = 1$, it follows that, for every conjecture $c_i \in \Delta^\circ(P_j^2) \times \Delta^\circ(B_0)$,

$$U_i\big(b_i', c_i\big) \geq u_i(x_3) \geq U_i(b_i, c_i).$$

Since such a conjecture c_i puts positive weight on the behavior $b_j = e(K)$ and

$$U_i\big(b_i', e(K), c_i^0\big) = u_i(x_K) > u_i(x_0) = U_i\big(b_i, e(K), c_i^0\big),$$

we conclude that $U_i(b_i', c_i) > U_i(b_i, c_i)$. □

Proof of Lemma 8 We have that

$$U_i(b_i, c_i) = \big(1 - \varepsilon - \varepsilon^2\big) \cdot u_i\big(x_{\overline{k}}\big) + \varepsilon \cdot u_i\big(x_{\overline{k}}\big) + \varepsilon^2 \cdot u_i\big(x_{\underline{k}}\big),$$

where if $\underline{k} \leq 2$ the expression in the last term follows from the fact that

$$\big(x_0 x_{\underline{k}}, I\big) \in f(b_i, e(1) + e(2) + e(3)) = f(b_i, e(1) + e(2) + e(\max\{3, \underline{k}\})) \neq A_i(b_i),$$

so we find that b_i selects the worst move in $f(b_i, e(1) + e(2) + e(3))$, which is equal to $(x_0 x_{\underline{k}}, I)$. The inequality in the last expression makes use of the fact that $\overline{k} \geq 4$.

Let $b'_i \in P_i^2 \setminus \{b_i\}$. If $b'_{i\overline{k}} = 0$, then, for $\varepsilon > 0$ sufficiently small,

$$U_i\left(b'_i, c_i\right) \leq \left(1 - \varepsilon - \varepsilon^2\right) \cdot u_i\left(x_{\overline{k}-1}\right) + \left(\varepsilon + \varepsilon^2\right) \cdot u_i(x_K) < U_i(b_i, c_i),$$

where the strict inequality makes use of the fact that ε is sufficiently small.

Assume $b'_{i\overline{k}} = 1$. If there is $k < \overline{k}$ such that $b'_{ik} = 1$, then, for $\varepsilon > 0$ sufficiently small,

$$U_i\left(b'_i, c_i\right) \leq \left(1 - \varepsilon - \varepsilon^2\right) \cdot u_i\left(x_{\overline{k}}\right) + \varepsilon \cdot u_i\left(x_{\overline{k}-1}\right) + \varepsilon^2 \cdot u_i(x_K) < U_i(b_i, c_i),$$

where the strict inequality makes use of the fact that ε is sufficiently small.

Let the smallest k for which $b'_{ik} = 1$ be equal to $\overline{k}$. It follows that $b_i \neq e(K)$, since $\overline{k} = K$ together with the assumption that the smallest k for which $b'_{ik} = 1$ is equal to $\overline{k}$ implies $b'_i = e(K)$. Since $b'_i \neq b_i$ it follows that $b_i \neq e(K)$. We have that

$$U_i\left(b'_i, c_i\right) = \left(1 - \varepsilon - \varepsilon^2\right) \cdot u_i\left(x_{\overline{k}}\right) + \varepsilon \cdot u_i\left(x_{\overline{k}}\right) < U_i(b_i, c_i),$$

where the second term in the expression after the equality follows from the fact that $f(b'_i, \mathbf{1}) = A_i(b'_i) \neq A_i(b_i)$, so the worst move $(x_0 x_{\overline{k}}, I)$ in $A_i(b'_i)$ is selected by b_0. The expression after the equality also uses that $b_i \neq e(K)$, so $\underline{k} < \overline{k}$, and $f(b'_i, e(1) + e(2) + e(\max\{3, \underline{k}\})) = \emptyset$. □

References

1. Battigalli P (1997) On rationalizability in extensive games. J Econ Theory 74:40–61
2. Bernheim D (1984) Rationalizable strategic behavior. Econometrica 52:1007–1028
3. Bloch F, van den Nouweland A (2020) Farsighted stability with heterogeneous expectations. Games Econom Behav 121:32–54
4. Chwe MS (1994) Farsighted coalitional stability. J Econ Theory 63:299–325
5. Diamantoudi E, Xue L (2003) Farsighted stability in hedonic games. Soc Choice Welf 21:39–61
6. Dutta B, Ghosal S, Ray D (2005) Farsighted network formation. J Econ Theory 122:143–164
7. Dutta B, Vartiainen H (2020) Coalition formation and history dependence. Theor Econ 15:159–197
8. Dutta B, Vohra R (2017) Rational expectations and farsighted stability. Theor Econ 12:1191–1227
9. Herings PJJ, Mauleon A, Vannetelbosch V (2000) Social rationalizability. CentER Discussion Paper 2000-81, Tilburg University, The Netherlands
10. Herings PJJ, Mauleon A, Vannetelbosch V (2004) Rationalizability for social environments. Games Econom Behav 49:135–156
11. Herings PJJ, Mauleon A, Vannetelbosch V (2009) Farsightedly stable networks. Games Econom Behav 67:526–541
12. Herings PJJ, Mauleon A, Vannetelbosch V (2010) Coalition formation among farsighted agents. Games 1:286–298
13. Herings PJJ, Mauleon A, Vannetelbosch V (2019) Stability of networks under horizon-K farsightedness. Econ Theor 68:177–200
14. Herings PJJ, Mauleon A, Vannetelbosch V (2020) Matching with myopic and farsighted players. J Econ Theory 190:105125
15. Karos D, Robles L (2021) Full farsighted rationality. Games Econom Behav 130:409–424
16. Kimya M (2020) Equilibrium coalitional behavior. Theor Econ 15:669–714
17. Luo C, Mauleon A, Vannetelbosch V (2021) Network formation with myopic and farsighted players. Econ Theor 71:1283–1317
18. Mauleon A, Vannetelbosch V (2004) Farsightedness and cautiousness in coalition formation games with positive spillovers. Theor Decis 56:291–324
19. Mauleon A, Vannetelbosch V, Vergote W (2011) von Neumann Morgenstern farsightedly stable sets in two-sided matching. Theor Econ 6:499–521
20. Page FH Jr, Wooders M (2009) Strategic basins of attraction, the path dominance core, and network formation games. Games Econom Behav 66:462–487

21. Page FH Jr, Wooders M, Kamat S (2005) Networks and farsighted stability. J Econ Theory 120:257–269
22. Pearce DG (1984) Rationalizable strategic behavior and the problem of perfection. Econometrica 52:1029–1050
23. Ray D, Vohra R (2015) The farsighted stable set. Econometrica 83:977–1011
24. Ray D, Vohra R (2019) Maximality in the farsighted stable set. Econometrica 87:1763–1779
25. Shimoji M, Watson J (1998) Conditional dominance, rationalizability, and game forms. J Econ Theory 83:161–195
26. Vannetelbosch V (1999) Rationalizability and equilibrium in N-person sequential bargaining. Econ Theor 14:353–371
27. Xue L (1998) Coalitional stability under perfect foresight. Econ Theor 11:603–627