CNRS - INPT - UPS - UT1 - UTM

RESEARCH REPORT

IRIT/RT–2010-3–FR

A Logical Account of Social Rationality in Strategic Games

Emiliano LORINI

Equipe LILaC
Emiliano.Lorini@irit.fr

May 1, 2010

UPS-IRIT, 118 route de Narbonne, 31062 Toulouse CEDEX 9
+33 (0) 561 55 67 65 info@irit.fr www.irit.fr

Abstract

We propose a modal logic that enables to reason about different kinds of rationality in strategic games. This logic integrates the concepts of joint action, belief, individual preference and group preference. The first part of the article is focused on the notion of individualistic rationality assumed in classical game theory: an agent decides to perform a certain action only if the agent believes that this action is a best response to what he expects the others will do. The second part of the article explores different kinds of social rationality such as fairness and reciprocity. Differently from individualistically rational agents (*alias* self-interested agents), social rational agents also consider the benefits of their choice for the group. Moreover, their decisions can be affected by their beliefs about other agents' willingness to act for the well-being of the group. In the article we also provide a complete axiomatization of our logic of joint action, belief, individual preference and group preference.

Keywords

Modal logic, game theory

Contents

1 Introduction

In recent times economic theory has started to refute the standard assumption that people are individualistically rational, that is, that they are necessarily motivated by their material self-interest. Several experiments have demonstrated that people are often driven by other-regarding motives, and by social preferences. Consequently, the classical game-theoretic framework has been revised and extended in order to incorporate new important concepts such as fairness, inequity aversion, altruism and reciprocity. Several models of social rationality have been developed in recent years showing that the phenomena observed in experiments with humans can be explained in a rigorous and tractable manner. Following [14], we can distinguish two families of models of social rationality in economics. Some models assume that players are only concerned about the distributional consequences of their acts but they do not care about the intentions that lead their opponents to choose these acts. For example, according to [13], players are inequity averse and may also care about how much material resources are allocated to other players. Other models assume that the behavior of a player during social interaction depends on the player's beliefs and expectations about the intention of the opponent.[1] For example, according to Rabin's model of reciprocity [22], a person is willing to be kind with another person if she believes that the other has been kind to her. On the contrary, a person has a desire to retaliate, if she believes that the other person wanted to hurt her.

Although social preferences and social rationality have been extensively studied in experimental economics, no logical analysis of these concepts has been proposed up to now. In this article, we try to fill this gap by proposing a sound and complete modal logic which enables to model social preferences and to reason about different kinds of social rationality in strategic games. This logic is based on the primitive concepts of joint action, belief, individual preference and group preference.

We think that developing logical models of social interaction integrating social preferences is a promising research avenue especially for the area of multi-agent systems and for the area of social software. Indeed, existing logical models of social interaction proposed in these areas are still anchored to the basic assumption of classical game theory that agents are self-interested. As emphasized above, this assumption is unrealistic for human agents. Therefore, to take the presence of other-regarding preferences into account becomes extremely important when developing formal models of social procedures to be applied to human societies, and when developing logical models of artificial agents which have to interact with human agents (e.g. trading agents, recommender systems, tutoring agents, etc.).

The remainder of the article is organized as follows. In Section 2 we present a modal logic of epistemic strategic games called **L** integrating the concepts of joint action, individual preference and belief. Section 3 is devoted to the analysis in **L** of some basic game-theoretic concepts such as best response and Nash equilibrium, and of the notion of individualistic rationality assumed in classical game theory (i.e. agents are self-interested). In Section 4 the logic **L** is extended by operators of group preference. Section 5 explores two different kinds of social rationality: fairness and reciprocity. Differently from individual-

[1] These models adopt the concept of 'psychological game theory' introduced by [15] and recently elaborated by [3]. In psychological game theory, utilities do not only depend on terminal-node payoffs but also on players' beliefs.

istically rational agents (*alias* self-interested agents), social rational agents also consider the benefits of their choice for the group. Moreover, their decisions can be affected by their beliefs about other agents' willingness to act for the well-being of the group. Section 6 discusses related works in the area of modal logics for game theory.

2 A logic of epistemic strategic games

We present in this section the modal logic **L** integrating the concepts of joint action, belief and individual preference. This logic supports reasoning about epistemic games in strategic form in which an agent might be being uncertain about the current choices of the other agents.

2.1 Syntax

The syntactic primitives of **L** are the finite set of agents Agt, the set of atomic formulas Atm, a nonempty finite set of atomic action names $Act = \{a_1, a_2, \ldots, a_{|Act|}\}$, a non-empty finite set of n integers $I = \{0, \ldots, n\}$. Non-empty sets of agents are called *coalitions* or *groups*, noted $C_1, C_2, \ldots$. We note $2^{Agt*} = 2^{Agt} \setminus \emptyset$ the set of coalitions.

To every agent $i \in Agt$ we associate the set Act_i of all possible ordered pairs $i{:}a$, that is, $Act_i = \{i{:}a \mid a \in Act\}$. Besides, for every coalition C we note Δ_C the set of all joint actions of this coalition, that is, $\Delta_C = \prod_{i \in C} Act_i$. Elements in Δ_C are C-tuples noted α_C, β_C, γ_C, $\delta_C, \ldots$. If $C = Agt$, we write Δ instead of Δ_{Agt}. Elements in Δ are also called strategy profiles. Given $\delta \in \Delta$, we note δ_i the element in δ corresponding to agent i. Moreover, for notational convenience, we write $\delta_{-i} = \delta_{Agt \setminus \{i\}}$.

The language $\mathcal{L}$ of the logic **L** is given by the following BNF:

$$\varphi \ ::= \ p \mid \bot \mid \neg\varphi \mid \varphi \vee \varphi \mid [\delta_C]\varphi \mid \Box\varphi \mid [\mathsf{B}_i]\varphi \mid [\geq_i^k]\varphi$$

where p ranges over Atm, i ranges over Agt, δ_C ranges over $\bigcup_{C \in 2^{Agt*}} \Delta_C$, and k ranges over I. The classical Boolean connectives $\wedge$, $\rightarrow$, $\leftrightarrow$ and $\top$ (tautology) are defined from $\bot$, $\vee$ and $\neg$ in the usual manner.

The formula $[\delta_C]\varphi$ reads "if coalition C chooses the joint action δ_C then φ holds". Therefore, $[\delta_C]\bot$ reads "coalition C does not choose the joint action δ_C ".

The operator $\Box$ is an universal modality, which quantifies over all worlds of the current model. Since in our framework, there is a one-to-one correspondence between models and games (i.e. a model corresponds to a unique strategic game), and every world in a model corresponds to a unique strategy profile of a strategic game, $\Box$ is used to quantify over the strategy profiles of the current game. $\Box\varphi$ reads "φ holds for all strategies profiles of the current game", or simply "φ is necessarily true". The formula $[\mathsf{B}_i]\varphi$ is read as usual "agent i believes that φ", whereas the formula $[\geq_i^k]\varphi$ is read "φ is true in all states which are good for agent i at least to degree k". Operators $[\geq_i^k]$ are used in **L** to define agents' preference orderings over the strategy profiles of the current game.

Moreover, we have the following abbreviations:

$$\langle\delta_C\rangle\varphi \stackrel{\text{def}}{=} \neg[\delta_C]\neg\varphi,$$
$$\Diamond\varphi \stackrel{\text{def}}{=} \neg\Box\neg\varphi,$$
$$\langle\mathsf{B}_i\rangle\varphi \stackrel{\text{def}}{=} \neg[\mathsf{B}_i]\neg\varphi,$$

$$\langle \geq_i^k \rangle \varphi \stackrel{\text{def}}{=} \neg[\geq_i^k]\neg\varphi.$$

$\Diamond\, \varphi$ has to be read "φ is possibly true", and $\langle \delta_C \rangle \varphi$ has to be read "coalition C chooses the joint action δ_C and φ holds". Therefore, $\langle \delta_C \rangle \top$ simply means "coalition C chooses the joint action δ_C". $\langle \geq_i^k \rangle \varphi$ has to read "φ is true in at least one state which is good for agent i at least to degree k". $\langle \mathsf{B}_i \rangle \varphi$ has to be read "agent i thinks that φ is possible".

2.2 Semantics

Definition 1 (**L**-model)**.** **L**-*models are tuples* $M = \langle W, R, B, P, \pi \rangle$ *where:*

- W *is a nonempty set of possible worlds or states;*
- R *is a collection of total functions* $R_C : W \longrightarrow \Delta_C$ *one for every coalition* $C \in 2^{Agt*}$*, mapping every world in* W *to a joint action of the coalition such that:*

  C1 $\delta_C = R_C(w)$ *if and only if, for every* $i \in C$ $\delta_i = R_i(w)$*, and*

  C2 *if for every* $i \in Agt$ *there is* $w_i \in W$ *such that* $\delta_i = R_i(w_i)$ *then there is a* $w \in W$ *such that* $\delta = R_{Agt}(w)$*,*

  C3 *if* $\delta = R_{Agt}(w)$ *and* $\delta = R_{Agt}(v)$*, then* $w = v$*;*

- $B : Agt \longrightarrow 2^{W \times W}$ *is a function mapping every agent* i *to a serial, transitive, Euclidean relation* B_i *on* W *such that:*

  C4 *if* $(w, v) \in B_i$*, then* $i{:}a = R_i(w)$ *if and only if* $i{:}a = R_i(v)$*;*

- $P : I \times Agt \longrightarrow 2^W$ *is a function mapping every integer* k *in* I *and agent* i *in* Agt *to a (possibly empty) set of worlds* P_i^k *such that:*

  C5 $P_i^0 = W$*, and*

  C6 *for every* $k \in I \setminus \{0\}$*,* $P_i^k \subseteq P_i^{k-1}$*;*

- $\pi : Atm \longrightarrow 2^W$ *is a valuation function.*

For every coalition C, $\delta_C = R_C(w)$ means that at world w coalition C chooses the joint action δ_C. Furthermore, if there exists $w \in W$ such that C performs δ_C at w then we say that δ_C is *possible* or that δ_C *can be performed*. Thus, δ can be performed if and only if δ is a strategy profile of the current game.

For every agent i, $(w, v) \in B_i$ means that at world w, according to agent i, world v is possible. Finally, the function P is used to specify the degree of *goodness* of a certain world according to a given agent. In particular, for every $k \in I$ and $i \in Agt$, P_i^k is the set of worlds that are good for agent i at least to degree k.

Let us discuss the six semantic constraints **C1**-**C6** in Definition 1.

According to Constraint **C1**, at world w coalition C chooses the joint action δ_C if and only if, every agent i in C chooses the action δ_i at w. In other words, a certain joint action is performed by a coalition if and only if every agent in the coalition does his part of the joint action. According to the Constraint **C2**, if every individual action in a joint action δ is possible, then their simultaneous occurrence is also possible. We moreover suppose determinism for the joint actions of all agents: different worlds in the same model correspond to the occurrences of *different* strategy profiles (Constraint **C3**).

Constraint **C4** just says that agents believe what they are doing. This is a standard assumption in interactive epistemology and epistemic analysis of games (see [7] for instance).

According to Constraint **C5**, every world is good for an agent i at least to minimal degree 0. This implies that an agent can always compare the goodness of two different worlds. According to the Constraint **C6**, for every integer $k \in I \setminus \{0\}$, the set of worlds that are good for agent i at least to degree k is a subset of of the set of worlds that are good for agent i at least to degree $k-1$. This means that worlds in a model are ordered according to their degrees of goodness for an agent and form a "system of spheres" [18] for every agent i. In particular, an *ordinal conditional function* (OCF) $\kappa_i : W \longmapsto I$ in Spohn's sense [26] can be associated with every agent i, where $\kappa_i(w)$ corresponds to the *exact* degree of goodness of world w for agent i.[2]

Definition 2 (κ_i)**.** *For every $i \in Agt$ and for every $w \in W$ we define:*

- $\kappa_i(w) = n$ *if and only if* $w \in P_i^n$*;*
- *for every $k \in I$ such that $k < n$, $\kappa_i(w) = k$ if and only if $w \in P_i^k$ and $w \notin P_i^{k+1}$.*

Example 1. *By way of example, consider a two-player Prisoner Dilemma (PD) game [21]. We suppose $Agt = \{i_1, i_2\}$ and ACT $= \{c, d\}$ where c is the action of cooperating and d is the action of defecting. In order to model PD game we just need four different degrees of goodness $I = \{0, 1, 2, 3\}$. The elements W, R and P of the model M corresponding to the two-player PD game are defined as follows:*

- $W = \{w_1, w_2, w_3, w_4\}$*;*
- $R_{\{i_1,i_2\}}(w_1) = \langle i_1{:}c, i_2{:}c\rangle$, $R_{\{i_1,i_2\}}(w_2) = \langle i_1{:}d, i_2{:}d\rangle$,
  $R_{\{i_1,i_2\}}(w_3) = \langle i_1{:}c, i_2{:}d\rangle$, $R_{\{i_1,i_2\}}(w_4) = \langle i_1{:}d, i_2{:}c\rangle$*;*
- $P_{i_1}^3 = \{w_4\}$, $P_{i_1}^2 = \{w_1, w_4\}$, $P_{i_1}^1 = \{w_1, w_2, w_4\}$, $P_{i_1}^0 = \{w_1, w_2, w_3, w_4\}$*;*
- $P_{i_2}^3 = \{w_3\}$, $P_{i_2}^2 = \{w_1, w_3\}$, $P_{i_2}^1 = \{w_1, w_2, w_3\}$, $P_{i_2}^0 = \{w_1, w_2, w_3, w_4\}$.

World w_3 (resp. w_4) belongs to $P_{i_1}^0$ (resp. $P_{i_2}^0$) and does not belong to $P_{i_1}^1$ (resp. $P_{i_2}^1$), hence it is a world that is good for agent i_1 (resp. agent i_2) exactly *to degree* 0*. That is, $\kappa_{i_1}(w_3) = 0$ (resp. $\kappa_{i_2}(w_4) = 0$).*

World w_2 belongs to $P_{i_1}^1$ (resp. $P_{i_2}^1$) and does not belong to $P_{i_1}^2$ (resp. $P_{i_2}^2$), hence it is a world that is good for agent i_1 (resp. agent i_2) exactly *to degree* 1*. That is, $\kappa_{i_1}(w_2) = 1$ (resp. $\kappa_{i_2}(w_2) = 1$).*

World w_1 belongs to $P_{i_1}^2$ (resp. $P_{i_2}^2$) and does not belong to $P_{i_1}^3$ (resp. $P_{i_2}^3$), hence it is a world that is good for agent i_1 (resp. agent i_2) exactly *to degree* 2*. That is, $\kappa_{i_1}(w_1) = 2$ (resp. $\kappa_{i_2}(w_1) = 2$).*

World w_4 (resp. w_3) belongs to $P_{i_1}^3$ (resp. $P_{i_2}^3$), hence it is a world that is good for agent i_1 (resp. agent i_2) exactly *to degree* 3*. That is, $\kappa_{i_1}(w_4) = 3$ (resp. $\kappa_{i_2}(w_3) = 3$).*

For every OCF function κ_i, we define the corresponding preference order $\preceq_i$ over the worlds in W.

[2] On this issue, see also [17].

Definition 3 ($\preceq_i$)**.** *For every $i \in Agt$ and for every $w, v \in W$ we have $w \preceq_i v$ if and only if $\kappa_i(w) \leq \kappa_i(v)$.*

Truth conditions for atomic formulas and the Boolean operators are entirely standard. The truth conditions for the operators of the logic **L** are:

- $M, w \models [\delta_C]\varphi \iff$ if $\delta_C = R_C(w)$ then $M, w \models \varphi$;
- $M, w \models \Box\varphi \iff M, v \models \varphi$ for all v such that $v \in W$;
- $M, w \models [\mathsf{B}_i]\varphi \iff M, v \models \varphi$ for all v such that $(w, v) \in B_i$;
- $M, w \models [\geq_i^k]\varphi \iff M, v \models \varphi$ for all v such that $v \in P_i^k$.

A formula φ is *true in an* **L**-*model* M iff $M, w \models \varphi$ for every world w in M. φ is **L**-*valid* (noted $\models \varphi$) iff φ is true in all **L**-models. φ is **L**-*satisfiable* iff $\neg\varphi$ is not **L**-valid.

2.3 Axiomatization

(**CPL**)	All principles of classical propositional logic
(**S5**$_\Box$)	All principles of modal logic S5 for $\Box$
(**S5**$_{[\mathsf{B}_i]}$)	All principles of modal logic KD45 for every $[\mathsf{B}_i]$
(**K**$_{[\geq_i^k]}$)	All principles of modal logic K for every $[\geq_i^k]$
(**Def**$_{[\delta_C]}$)	$[\delta_C]\varphi \leftrightarrow (\langle\delta_C\rangle\top \to \varphi)$
(**JointAct**)	$\langle\delta_C\rangle\top \leftrightarrow \bigwedge_{i\in C} \langle\delta_i\rangle\top$
(**Active**)	$\bigvee_{\delta_C\in\Delta_C} \langle\delta_C\rangle\top$
(**Single**)	$\langle\delta_C\rangle\top \to [\delta'_C]\bot$ if $\delta_C \neq \delta'_C$
(**Indep**)	$\left(\bigwedge_{i\in Agt} \Diamond\langle\delta_i\rangle\top\right) \to \Diamond\langle\delta\rangle\top$
(**JointDet**)	$(\langle\delta\rangle\top \wedge \varphi) \to \Box(\langle\delta\rangle\top \to \varphi)$
(**Aware**)	$\langle i{:}a\rangle\top \leftrightarrow [\mathsf{B}_i]\langle i{:}a\rangle\top$
(**Incl**$_{\Box,[\mathsf{B}_i]}$)	$\Box\varphi \to [\mathsf{B}_i]\varphi$
(**Incl**$_{\Box,[\geq_i^k]}$)	$\Box\varphi \to [\geq_i^k]\varphi$
(**Def**$_{[\geq_i^0]}$)	$[\geq_i^0]\varphi \leftrightarrow \Box\varphi$
(**Incl**$_{[\geq_i^{k-1}],[\geq_i^k]}$)	$[\geq_i^{k-1}]\varphi \to [\geq_i^k]\varphi$
(**PNec**$_{[\geq_i^k]}$)	$[\geq_i^k]\varphi \to \Box[\geq_i^k]\varphi$
(**NNec**$_{[\geq_i^k]}$)	$\neg[\geq_i^k]\varphi \to \Box\neg[\geq_i^k]\varphi$

Table 1: Axiomatization of **L**

We call **L** the logic that is axiomatized by the principles given in Table 1, and we write $\vdash_{\mathbf{L}} \varphi$ if φ is a theorem of **L**.

Note that Axiom **Indep** is the **L** counterpart of the so-called *axiom of independence of agents* of STIT logic [4]. This axiom enables to express the basic game theoretic assumption that the set of strategy profiles of a game in strategic form is the cartesian product of the sets of individual actions for the agents in Agt. Note also that Axiom $\mathbf{Def}_{[\delta_C]}$ could be replaced by the simpler Axiom $\varphi \rightarrow [\delta_C]\varphi$. We have chosen $\mathbf{Def}_{[\delta_C]}$ because we think that it better expresses the meaning of the operators $[\delta_C]$. Furthermore, by the Axiom $\mathbf{Def}_{[\delta_C]}$, we can prove that every operator $[\delta_C]$ satisfies all principles of the basic modal logic K and, in particular, Axiom K:

$$\vdash_{\mathbf{L}} ([\delta_C]\varphi \wedge [\delta_C](\varphi \rightarrow \psi)) \rightarrow [\delta_C]\psi, \tag{1}$$

and the rule of necessitation:

$$\text{if } \vdash_{\mathbf{L}} \varphi \text{ then } \vdash_{\mathbf{L}} [\delta_C]\varphi. \tag{2}$$

The following property is also derivable by Axiom T for $\Box$ and Axiom $\mathbf{Def}_{[\delta_C]}$:

$$\vdash_{\mathbf{L}} \Box\varphi \rightarrow [\delta_C]\varphi.$$

Theorem 1. *The logic* **L** *is sound and complete with respect to the class of* **L***-models.*

Proof. We here provide a sketch of the proof.

It is straightforward to show that all our axioms are valid and that the rules of inference preserve validity in the class of **L**-models. The other part of the proof is shown using two major steps.

Step 1. We provide an alternative semantics for **L** in terms of standard Kripke models whose semantic conditions correspond one-to-one to the axioms in Table 1. The definition of Kripke **L**-models is the following one.

Definition 4 (Kripke **L**-model)**.** *Kripke* **L***-models are tuples* $M = \langle W, \sim, R, B, P, \pi\rangle$ *where:*

- W *is a nonempty set of possible worlds or states;*
- $\sim$ *is an equivalence relation on* W*;*
- $R : \bigcup_{C \in 2^{Agt*}} \Delta_C \longrightarrow 2^{W\times W}$ *maps every joint action* δ_C *to a transition relation* $R_{\delta_C} \subseteq W \times W$ *between possible worlds such that:*

  S1 $R_{\delta_C}(w) \neq \emptyset$ *if and only if, for every* $i \in C$ $R_{\delta_i}(w) \neq \emptyset$*,*

  S2 if $R_{\delta_C}(w) \neq \emptyset$ *then* $R_{\delta_C}(w) = \{w\}$*,*

  S3 $\bigcup_{\delta_C \in \Delta_C} R_{\delta_C}(w) \neq \emptyset$*,*

  S4 if $\delta_C \neq \delta'_C$ *then* $R_{\delta_C}(w) = \emptyset$ *or* $R_{\delta'_C}(w) = \emptyset$*,*

  S5 if for every $i \in Agt$ *there is* v_i *such that* $w \sim v_i$ *and* $R_{\delta_i}(v_i) \neq \emptyset$ *then there is a* v *such that* $w \sim v$ *and* $R_\delta(v) \neq \emptyset$*,*

  S6 if $w \sim v$ *and* $R_\delta(w) \neq \emptyset$ *and* $R_\delta(v) \neq \emptyset$*, then* $w = v$*;*

- $B : Agt \longrightarrow 2^{W\times W}$ *is a function mapping every agent* i *to a serial, transitive and Euclidean relation* B_i *on* W *such that:*

S7 *if* $(w, v) \in B_i$*, then* $i{:}a = R_i(w)$ *if and only if* $i{:}a = R_i(v)$*,*

S8 *if* $(w, v) \in B_i$ *then* $w \sim v$*;*

- $P : I \times Agt \longrightarrow 2^{W \times W}$ *maps every integer* k *in* I *and agent* i *in* Agt *to a binary relation* P_i^k *on* W *such that:*

  S9 $P_i^0(w) = \sim(w)$*,*

  S10 $P_i^k(w) \subseteq \sim(w)$*,*

  S11 *if* $k \in I \setminus \{0\}$ *then* $P_i^k(w) \subseteq P_i^{k-1}(w)$*,*

  S12 *if* $w \sim v$ *then* $P_i^k(w) \subseteq P_i^k(v)$*,*

  S13 *if* $w \sim v$ *then* $P_i^k(v) \subseteq P_i^k(w)$*;*

- $\pi : Atm \longrightarrow 2^W$ *is a valuation function.*

As in the previous Constraint C2, accessibility relations on W can be viewed as functions from W to 2^W. Therefore, we write $\sim(w) = \{v \mid w \sim v\}$, $P_i^k(w) = \{v \mid (w, v) \in P_i^k\}$, etc. Truth conditions of **L** formulas in Kripke **L**-models are again standard for atomic formulas and the Boolean operators. The truth conditions for the other operators are:

- $M, w \models [\delta_C]\varphi \iff M, v \models \varphi$ for all $v \in R_{\delta_C}(w)$;
- $M, w \models \Box\varphi \iff M, v \models \varphi$ for all $v \in \sim(w)$;
- $M, w \models [\mathsf{B}_i]\varphi \iff M, v \models \varphi$ for all $v \in B_i(w)$;
- $M, w \models [\geq_i^k]\varphi \iff M, v \models \varphi$ for all $v \in P_i^k(w)$.

It is a routine task to prove that the axiomatic system of the logic **L** given in Table 1 is sound and complete with respect to this class of Kripke **L**-models via the Sahlqvist theorem, cf. [6, Th. 2.42]. Indeed all axioms in Table 1 are in the so-called Sahlqvist class [25]. Thus, they are all expressible as first-order conditions on Kripke models and are complete with respect to the defined model classes.

Step 2. The second step shows that the semantics in terms of **L**-models of Definition 1 and the semantics in terms of Kripke **L**-models of Definition 4 are equivalent. As the logic **L** is sound and complete for the class of Kripke **L**-models and is sound for the class of **L**-models, we have that for every **L** formula φ, if φ is valid in the class of Kripke **L**-models then φ is valid in the class of **L**-models. Consequently, for every **L** formula φ, if φ is satisfiable in the class of **L**-models then φ is satisfiable in the class of Kripke **L**-models. Therefore, in this second step we just need to show that for every **L** formula φ, if φ is satisfiable in the class of Kripke **L**-models then φ is satisfiable in the class of **L**-models.

Suppose φ is satisfiable in the class of Kripke **L**-models. This means that there is a Kripke **L**-model $M = \langle W, \sim, R, B, P, \pi\rangle$ and world w such that $M, w \models \varphi$. We can now build a **L**-model $M' = \langle W', R', B', P', \pi'\rangle$ which satisfies φ. The construction of the model M' is made in two steps. We first consider the submodel $M_w = \langle W_w, \sim_w, R_w, B_w, P_w, \pi_w\rangle$ generated from M and w: by the generated submodel property [6] we have $M_w, w \models \varphi$. M_w is also a Kripke **L**-model, and $\sim_w = W_w \times W_w$. The latter means that $\Box$ is interpreted as a universal modal operator. Finally, we can define $M' = \langle W', R', B', P', \pi'\rangle$ as follows:

- $W' = W_w$;
- for every $C \in 2^{Agt*}$ and $v \in W'$, $R'_C(v) = \delta_C$ if and only if $R_{\delta_C,w}(v) \neq \emptyset$;
- for every $i \in Agt$, $B'_i = B_{i,w}$;
- for every $i \in Agt$ and $k \in I$, $P'^k_i = P^k_{i,w}(w)$;
- $\pi' = \pi_w$.

It is a routine task to check that M' is indeed a **L**-model and, by induction on the structure of φ, that we have $M', w \models \varphi$.

□

3 Strategic games with individual rationality

This section is devoted to the analysis in the logic **L** of strategic games in which agents are self-interested. We first define preference operators. Then, we consider the basic game-theoretic concepts of best response and Nash equilibrium, and their relationships with the notion of individualistic rationality assumed in classical game theory.

3.1 Preferences over formulas

For every $i \in Agt$, we define a preference relation over formulas:

$$\psi \leq_i \varphi \stackrel{\text{def}}{=} \bigwedge_{k \in I} (\langle \geq_i^k \rangle \psi \rightarrow \langle \geq_i^k \rangle \varphi).$$

According to this definition "φ is for agent i at least as good as ψ" (noted $\psi \leq_i \varphi$) if and only if for every integer $k \in I$ if ψ is true in at least one state which is good for agent i at least to degree k then φ is also true in at least one state which is good for agent i at least to degree k. The following are central properties of the operator $\leq_i$. For every $i \in Agt$ we have:

(3) $\vdash_{\mathbf{L}} \psi \leq_i \psi$

(4) $\vdash_{\mathbf{L}} ((\varphi_1 \leq_i \varphi_2) \wedge (\varphi_2 \leq_i \varphi_3)) \rightarrow (\varphi_1 \leq_i \varphi_3)$

(5) $\vdash_{\mathbf{L}} (\varphi_1 \leq_i \varphi_2) \vee (\varphi_2 \leq_i \varphi_1)$

(6) $\vdash_{\mathbf{L}} \bot \leq_i \top$

(7) if $\vdash_{\mathbf{L}} \varphi \rightarrow (\psi_1 \vee ... \vee \psi_s)$ then $\vdash_{\mathbf{L}} (\varphi \leq_i \psi_1) \vee ... \vee (\varphi \leq_i \psi_s)$

The **L**-theorems 3-5 highlight that $\leq_i$ is a total preorder. The **L**-theorems 4-6 are exactly the three fundamental principles of Lewis's conditional logic (called $\mathcal{VN}$) [18]. The language of Lewis's logic $\mathcal{VN}$ is that of classical logic augmented with a dyadic connective $\leq$. In this sense, the operators $\langle \geq_i^k \rangle$ enables to reconstruct Lewis' dyadic connective. A similar result has been proved by del Cerro & Herzig [11] in the 90ies. The two authors have developed a modal logic of possibility called $\mathcal{PL}$ in which a normal operator of the form $[x]$ for every parameter x in a set of parameters P is introduced (the dual of $[x]$ is noted $\langle x \rangle$). In a way similar to the **L** formula $[\geq_i^k]\varphi$, the $\mathcal{PL}$ formula $[x]\,\varphi$ is read "the necessity of formula φ is at least degree x". $\mathcal{PL}$ have the standard axioms and rules of inference of system K for every operator $[x]$ plus the following two axioms:

PL1 $\neg [1] \bot$

PL2 $([x]\,\varphi \to [y]\,\varphi) \vee ([x]\,\varphi \to [y]\,\varphi)$ for every $x, y \in P$

where 1 is a particular element in the set of parameters P. The two authors reconstruct Lewis's preferential operators by means of the following translation from $\mathcal{PL}$ to Lewis's conditional logic: $\psi \leq \varphi \stackrel{\text{def}}{=} \bigwedge_{x \in P}(\langle x \rangle\, \psi \to \langle x \rangle\, \varphi)$. [3] We can easily prove that our logic **L** satisfies the previous two properties of del Cerro & Herzig's logic. Indeed, the following **L**-theorem is provable by Axiom $\mathbf{Def}_{[\geq_i^0]}$ and by Axiom T for $\Box$:

$$\vdash_{\mathbf{L}} \neg[\geq_i^0]\bot. \tag{8}$$

This **L**-theorem corresponds to Axiom PL1 of $\mathcal{PL}$. Moreover, by the **L** Axiom $\mathbf{Incl}_{[\geq_i^{k-1}],[\geq_i^k]}$, the following **L**-theorem is also provable for every $k, h \in I$:

$$\vdash_{\mathbf{L}} ([\geq_i^k]\varphi \to [\geq_i^h]\varphi) \vee ([\geq_i^k]\varphi \to [\geq_i^h]\varphi) \tag{9}$$

This **L**-theorem corresponds to Axiom PL2 of $\mathcal{PL}$. On the contrary in $\mathcal{PL}$ we cannot infer a theorem which corresponds to the **L** Axiom $\mathbf{Incl}_{[\geq_i^{k-1}],[\geq_i^k]}$, that is, in $\mathcal{PL}$ we cannot infer as a theorem something like $[x]\,\varphi \to [y]\,\varphi$ if $x < y$. In this sense, the fragment of our logic **L** with preferential operators of type $[\geq_i^k]$ can be seen as a generalization of $\mathcal{PL}$.

Before concluding let us define $\psi <_i \varphi \stackrel{\text{def}}{=} (\psi \leq_i \varphi) \wedge \neg(\varphi \leq_i \psi)$, $\delta \leq_i \delta' \stackrel{\text{def}}{=} \langle \delta \rangle \top \leq_i \langle \delta' \rangle \top$ and $\delta <_i \delta' \stackrel{\text{def}}{=} (\delta \leq_i \delta') \wedge \neg(\delta' \leq_i \delta)$.

3.2 Best response and Nash equilibrium

Some basic concepts of game theory can be expressed in **L** in terms of comparative goodness. We first consider *best response*. Agent i's action a is said to be a best response to the other agents' joint action δ_{-i}, noted $\mathsf{BR}(i{:}a{,}\delta_{-i})$, if and only if i cannot improve his utility by deciding to do something different from a while the others choose the joint action δ_{-i}, that is:

$$\mathsf{BR}(i{:}a{,}\delta_{-i}) \stackrel{\text{def}}{=} \bigwedge_{b \in Act} ((\langle i{:}b \rangle \top \wedge \langle \delta_{-i} \rangle \top) \leq_i (\langle i{:}a \rangle \top \wedge \langle \delta_{-i} \rangle \top)).$$

Given a certain strategic game, the strategy profile (or joint action) δ is said to be a *Nash equilibrium* if and only if for every agent $i \in Agt$, i's action δ_i is a best response to the other agents' joint action δ_{-i}:

$$\mathsf{Nash}(\delta) \stackrel{\text{def}}{=} \bigwedge_{i \in Agt} \mathsf{BR}(\delta_i{,}\delta_{-i}).$$

Example 2. *It is well-known that in the Prisoner Dilemma the only Nash equilibrium is mutual defection. Hence, in the model illustrated in the Example 1 of Section 2.2 we have that formula* $\mathsf{Nash}(\langle i_1{:}d, i_2{:}d \rangle)$ *is true at each world* w_1, w_2, w_3, w_4 *of the model* M.

In order to show that the previous formalization of Nash equilibrium is correct, we need to introduce the definition of strategic game in the game-theoretic sense, and then we need a notion of correspondence between strategic games and **L**-models.

[3] An additional result proved by the two authors is that Lewis's conditional logic is nothing more that the logic of qualitative possibility developed by Dubois & Prade [12]. On this see also [8].

Definition 5 (Strategic game)**.** *A strategic game is a tuple* $\Gamma = \langle N, A, U\rangle$ *where* N *is a finite set of agents and:*

- A *is a collection of sets of actions* A_i *one for every agent* $i \in N$*;*
- U *is a collection of total preorders* $\leq_i^U$ *over* $\prod_{i\in Agt} A_i$*, one for every agent* $i \in Agt$*.*

The strategy profile δ is a Nash equilibrium in the game Γ if and only if, for every $i \in N$ and for every $\delta'_i \in A_i$, $(\delta'_i, \delta_{-i}) \leq_i^U (\delta_i, \delta_{-i})$.

Definition 6 ($\Gamma \approx M$)**.** *We say that a strategic game* Γ corresponds *to a* **L**-*model* $M = \langle W, R, B, P, \pi\rangle$*, noted* $\Gamma \approx M$*, if and only if the following conditions are met:*

- $Agt = N$*;*
- *for every* $i \in Agt$*,*
  $A_i = \{i{:}a \mid a \in Act$ *and there is* $w \in W$ *such that* $R_i(w) = i{:}a\}$*;*
- *for every* $i \in Agt$*,*
  and for every $\delta, \delta' \in \prod_{i\in Agt} A_i$*,* $\delta' \leq_i^U \delta$ *if and only if* $R_{Agt}^{-1}(\delta') \preceq_i R_{Agt}^{-1}(\delta)$*;*

where R_{Agt}^{-1} *is the inverse function of* R_{Agt}*.* [4]

Proposition 1. *Suppose that* $\Gamma \approx M$ *and* $\delta \in \prod_{i\in Agt} A_i$*. Then, we have* $M, w \models \mathsf{Nash}(\delta)$ *for every* $w \in W$ *if and only if,* δ *is a Nash equilibrium in the game* Γ*.*

3.3 Individualistic rationality

The following **L** formula characterizes a notion of individualistic rationality which is commonly supposed in the epistemic analysis of games (see, e.g., [2, 5]):

$$\bigwedge_{a,b\in Act}(\langle i{:}a\rangle\top \to \bigvee_{\delta\in\Delta}(\langle \mathsf{B}_i\rangle\langle\delta_{-i}\rangle\top \wedge (\langle\delta_{-i}, i{:}b\rangle \leq_i \langle\delta_{-i}, i{:}a\rangle))).$$

This means that an agent i is (individualistically) rational if and only if, if he chooses a particular action a then for every alternative action b, there exists a joint action δ_{-i} of the other agents that he considers possible such that, playing a while the others play δ_{-i} is for i at least as good as playing b while the others play δ_{-i}. As in **L** $\delta \leq_i \delta'$ and $[\mathsf{B}_i](\delta \leq_i \delta')$ are equivalent, the previous definition of rationality can be rewritten in the following equivalent form:

$$\mathsf{IRat}_i \stackrel{\text{def}}{=} \bigwedge_{a,b\in Act}\left(\langle i{:}a\rangle\top \to \bigvee_{\delta\in\Delta}\left(\langle \mathsf{B}_i\rangle\langle\delta_{-i}\rangle\top \wedge [\mathsf{B}_i]\left(\langle\delta_{-i}, i{:}b\rangle \leq_i \langle\delta_{-i}, i{:}a\rangle\right)\right)\right).$$

For every $i \in Agt$ we have:

(10) $\vdash_{\mathbf{L}} \mathsf{IRat}_i \leftrightarrow [\mathsf{B}_i]\mathsf{IRat}_i$

(11) $\vdash_{\mathbf{L}} \neg\mathsf{IRat}_i \leftrightarrow [\mathsf{B}_i]\neg\mathsf{IRat}_i$

L-theorems 10 and 11 highlight that the concepts of individualistic rationality and irrationality are positively and negatively introspective. The following theorem specifies

[4] Note that this function exists, as Constraint **C3** on **L**-models ensures that R_{Agt} is a bijection.

some sufficient conditions for guaranteeing that the chosen strategy profile is a Nash equilibrium: if all agents are individualistically rational (*alias* self-interested) and every agent knows the choices of the other agents, then the selected strategy profile is a Nash equilibrium. For every $\delta \in \Delta$ we have:

(12) $$\vdash_{\mathbf{L}} \Big(\bigwedge_{i \in Agt} (\mathsf{IRat}_i \wedge [\mathsf{B}_i]\langle \delta_{-i} \rangle \top) \wedge \langle \delta \rangle \top \Big) \rightarrow \mathsf{Nash}(\delta)$$

A similar theorem has been stated for the first time by Aumann & Brandeburger [1, 9]. The only difference is that Aumann & Brandeburger used knowledge instead of belief.

4 An extension with group preferences

In this section we extend the logic **L** by operators of group preferences of the form $[\geq^k_C]$. We call $\mathbf{L}^+$ the extended logic. Operators $[\geq^k_C]$ are just a generalization of the operators of individual preference of the form $[\geq^k_i]$. In the present approach a preference for a group C is given by the intersection of the preferences of the agents in the group.

4.1 Syntax

The language of the logic $\mathbf{L}^+$ is obtained by extending the language of the logic **L** by formulas $[\geq^k_C]\varphi$ where C ranges over 2^{Agt*} and k ranges over I. The formula $[\geq^k_C]\varphi$ has to be read "φ is true in all states which are good for group C at least to degree k". We define $\langle \geq^k_C \rangle \varphi \stackrel{\text{def}}{=} \neg[\geq^k_C]\neg\varphi$. $\langle \geq^k_C \rangle \varphi$ has to be read "φ is true in at least one state which is good for group C at least to degree k".

4.2 Semantics

Definition 7 ($\mathbf{L}^+$-model)**.** *$\mathbf{L}^+$-models are tuples $M = \langle W, R, B, P, GP, \pi \rangle$ where:*

- *$M = \langle W, R, B, P, \pi \rangle$ is a **L**-model;*
- *$GP : I \times 2^{Agt*} \longrightarrow 2^W$ is a function mapping every integer k in I and coalition C to a (possibly empty) set of worlds GP^k_C such that:*

  C7 $GP^k_C = \bigcap_{i \in C} P^k_i$.

The function GP is used to specify the degree of *goodness* of a certain world according to a given group of agents. In particular, for every $k \in I$ and $C \in 2^{Agt*}$, GP^k_C is the set of worlds that are good for group C at least to degree k.

Let us generalize the functions κ_i à la Spohn and the preference orders $\preceq_i$ of Section 2.2 to groups.

Definition 8 (κ_C)**.** *For every $C \in 2^{Agt*}$ and for every $w \in W$ we define:*

- *$\kappa_C(w) = n$ if and only if $w \in GP^n_C$;*
- *for every $k \in I$ such that $k < n$, $\kappa_C(w) = k$ if and only if $w \in GP^k_C$ and $w \notin GP^{k+1}_C$.*

Definition 9 ($\preceq_C$)**.** *For every $C \in 2^{Agt*}$ and for every $w, v \in W$ we have $w \preceq_C v$ if and only if $\kappa_C(w) \leq \kappa_C(v)$.*

Definition 10 (Max_C)**.** *For every $C \in 2^{Agt*}$ we define $\mathsf{Max}_C = \max_{w \in W} \kappa_C(w)$.*

The following propositions highlight some interesting aspect of the present notion of group preference.

Proposition 2. $\kappa_C(w) \leq \kappa_C(v)$ *if and only if* $\min_C \kappa_i(w) \leq \min_{i \in C} \kappa_i(v)$.

According to Proposition 2, for the group of agents C the degree of goodness of world v is higher than the degree of goodness of world w if and only if, the lowest degree of goodness of world v for an agent in the group C is higher than the lowest degree of goodness of world w for an agent in the group C.

Proposition 3. $w \in \mathsf{Max}_C$ *if and only if* $w \in \max_{v \in W} \min_{i \in C} \kappa_i(v)$.

Proposition 3 is about a *maximin* principle for group preferences: the degree of goodness of world w for group C is maximal if and only if, w belongs to the set of worlds whose lowest degree of goodness for an agent in the group C is maximal. Note that since [23], social procedures based on a *maximin* principle have been considered appropriate for ensuring fairness and justice. Consequently, we think that *maximin* is also a good method for defining the concept of group preference.

REMARK. Note that the *maximin* principle also ensures Pareto optimality. In particular, if $w \in \mathsf{Max}_C$ then $w \in \mathsf{Pareto}_C$, where $\mathsf{Pareto}_C = \{v | \forall\, u \in W, \exists i \in C\colon u \preceq_i v\}$.

Example 3. *Consider again the two-player Prisoner Dilemma (PD) game illustrated in the Example 1 of Section 2.2 with $Agt = \{i_1, i_2\}$ and $ACT = \{c, d\}$ where c is the action of cooperating and d is the action of defecting. The PD game just needs four different degrees of goodness $I = \{0, 1, 2, 3\}$ and corresponds to the model M whose elements W, R, P are defined as follows:*

- $W = \{w_1, w_2, w_3, w_4\}$;
- $R_{\{i_1,i_2\}}(w_1) = \langle i_1{:}c, i_2{:}c\rangle$, $R_{\{i_1,i_2\}}(w_2) = \langle i_1{:}d, i_2{:}d\rangle$,
  $R_{\{i_1,i_2\}}(w_3) = \langle i_1{:}c, i_2{:}d\rangle$, $R_{\{i_1,i_2\}}(w_4) = \langle i_1{:}d, i_2{:}c\rangle$;
- $P^3_{i_1} = \{w_4\}$, $P^2_{i_1} = \{w_1, w_4\}$, $P^1_{i_1} = \{w_1, w_2, w_4\}$, $P^0_{i_1} = \{w_1, w_2, w_3, w_4\}$;
- $P^3_{i_2} = \{w_3\}$, $P^2_{i_2} = \{w_1, w_3\}$, $P^1_{i_2} = \{w_1, w_2, w_3\}$, $P^0_{i_2} = \{w_1, w_2, w_3, w_4\}$.

Consider now group preferences. We have:

- $GP^2_{\{i_1,i_2\}} = \{w_1\}$, $GP^1_{\{i_1,i_2\}} = \{w_1, w_2\}$, $GP^0_{\{i_1,i_2\}} = \{w_1, w_2, w_3, w_4\}$.

Therefore, $w_1 \in \mathsf{Max}_{\{i_1,i_2\}}$. In other words, the state in which both i_1 and i_2 cooperate is the best state for the group $\{i_1, i_2\}$. Indeed, this is the state whose lowest degree of goodness for an agent in $\{i_1, i_2\}$ is maximal.

Before concluding we have to provide the truth condition for $[\geq^k_C]\varphi$:

$$M, w \models [\geq^k_C]\varphi \iff M, v \models \varphi \text{ for all } v \text{ such that } v \in GP^k_C.$$

4.3 Axiomatization

Every operator $[\geq^k_C]$ satisfies the principles of the basic modal logic K *plus* the following two additional principles:

(**Def**$_{[\geq^k_i]}$) $\quad [\geq^k_i]\varphi \leftrightarrow [\geq^k_{\{i\}}]\varphi$

(**GroupPref**) $\quad \langle\geq^k_C\rangle\langle\delta\rangle\top \leftrightarrow (\bigwedge_{i\in C}\langle\geq^k_i\rangle\langle\delta\rangle\top)$

Note that Axiom **GroupPref** is the fundamental principle relating group preferences with individual preferences.

We call $\mathbf{L}^+$ the logic axiomatized by the principles of the logic $\mathbf{L}$ given in Section 2.3 (Table 1) plus the axioms and the rules of inference of the basic system K for every operator $[\geq^k_C]$ and the previous two Axioms **Def**$_{[\geq^k_i]}$ and **GroupPref**. Definitions of validity and theoremhood for the logic $\mathbf{L}^+$ are just standard. We write $\models_{\mathbf{L}^+}\varphi$ if formula φ is valid in all $\mathbf{L}^+$-models. We write $\vdash_{\mathbf{L}^+}\varphi$ if formula φ is a theorem of $\mathbf{L}^+$.

The following Theorem 2 can be proved in a way similar to Theorem 1.

Theorem 2. *The logic* $\mathbf{L}^+$ *is sound and complete with respect to the class of* $\mathbf{L}^+$*-models.*

The following are some properties of the operators of collective preference. For every $B, C \in 2^{Agt*}$ such that $B \subseteq C$ and for every $k \in I$ we have:

(13) $\quad \vdash_{\mathbf{L}^+} [\geq^k_B]\varphi \to [\geq^k_C]\varphi$

(14) $\quad \vdash_{\mathbf{L}^+} [\geq^0_C]\varphi \leftrightarrow \Box\varphi$

(15) $\quad \vdash_{\mathbf{L}^+} [\geq^{k-1}_C]\varphi \to [\geq^k_C]\varphi$

5 Strategic games with social rationality

In this section we explore strategic games in which agents are driven by other-regarding motives and social preferences. We consider two different kinds of social rationality: fairness and reciprocity.

5.1 Group preferences over formulas and fairness equilibrium

We can now generalize the notion of preference over formulas to groups. For every $C \in 2^{Agt*}$, we define $\psi \leq_C \varphi$ ("φ is for group C at least as good as ψ") as follows:

$$\psi \leq_C \varphi \stackrel{\text{def}}{=} \bigwedge_{k\in I}(\langle\geq^k_C\rangle\psi \to \langle\geq^k_C\rangle\varphi).$$

As the following $\mathbf{L}^+$-theorems highlight the operators $\leq_C$ are also total preorders. For every $C \in 2^{Agt*}$ we have:

(16) $\quad \vdash_{\mathbf{L}^+} \psi \leq_C \psi$

(17) $\quad \vdash_{\mathbf{L}^+} ((\varphi_1 \leq_C \varphi_2) \wedge (\varphi_2 \leq_C \varphi_3)) \to (\varphi_1 \leq_C \varphi_3)$

(18) $\quad \vdash_{\mathbf{L}^+} (\varphi_1 \leq_C \varphi_2) \vee (\varphi_2 \leq_C \varphi_1)$

We define $\psi <_C \varphi \stackrel{\text{def}}{=} (\psi \leq_C \varphi) \wedge \neg(\varphi \leq_C \psi)$, $\delta \leq_C \delta' \stackrel{\text{def}}{=} \langle\delta\rangle\top \leq_C \langle\delta'\rangle\top$ and $\delta <_C \delta' \stackrel{\text{def}}{=} (\delta \leq_C \delta') \wedge \neg(\delta' \leq_C \delta)$.

The notion of best response of Section 3.2 can also be generalized to groups. We say that agent i's action a is for the group C a best response to the other agents' joint action δ_{-i}, noted $\mathsf{BR}_C(i{:}a,\delta_{-i})$, if and only if i cannot improve the utility of group C by deciding to do something different from a while the others choose the joint action δ_{-i}, that is:

$$\mathsf{BR}_C(i{:}a,\delta_{-i}) \stackrel{\text{def}}{=} \bigwedge_{b\in Act} ((\langle i{:}b\rangle\top \wedge \langle\delta_{-i}\rangle\top) \leq_C (\langle i{:}a\rangle\top \wedge \langle\delta_{-i}\rangle\top)).$$

Note that the notion of best response of Section 3.2 is just a special case of the previous definition, when $C = \{i\}$. In fact, $\mathsf{BR}_{\{i\}}(i{:}a,\delta_{-i})$ is logically equivalent to $\mathsf{BR}(i{:}a,\delta_{-i})$.

Before concluding this section, we provide a notion of fairness equilibrium. Basically, fairness equilibrium is the collective counterpart of the individualistic notion of Nash equilibrium. Given a certain strategic game, the strategy profile (or joint action) δ is said to be a *fairness equilibrium* if and only if for every agent $i \in Agt$, i's action δ_i is for the entire group Agt a best response to the other agents' joint action δ_{-i}:

$$\mathsf{FairEq}(\delta) \stackrel{\text{def}}{=} \bigwedge_{i\in Agt} \mathsf{BR}_{Agt}(\delta_i,\delta_{-i}).$$

In other words, in a fairness equilibrium every agent chooses an action that, given what the others choose, maximizes the utility of the entire group of agents Agt.

Example 4. *In the Prisoner Dilemma there are two fairness equilibria: mutual cooperation and mutual defection. Hence, in the model illustrated in the Example 1 of Section 2.2 we have that formulas* $\mathsf{FairEq}(\langle i_1{:}c, i_2{:}c\rangle)$ *and* $\mathsf{FairEq}(\langle i_1{:}d, i_2{:}d\rangle)$ *are both true at each world* w_1, w_2, w_3, w_4 *of the model* M.

5.2 Social rationality

It is time to look at social rationality. Differently from individualistically rational agents defined in Section 3.3, social rational agents also consider the benefits of their choice for the group. Moreover, their decisions can be affected by their beliefs about other agents' willingness to act for the well-being of the group.

We first define fairness. For every $i \in Agt$ and $C \in 2^{Agt*}$ such that $i \in C$:

$$\mathsf{Fair}_{i,C} \stackrel{\text{def}}{=} \bigwedge_{a,b\in Act} (\langle i{:}a\rangle\top \rightarrow \bigvee_{\delta\in\Delta} (\langle \mathsf{B}_i\rangle\langle\delta_{-i}\rangle\top \wedge (\langle\delta_{-i}, i{:}b\rangle \leq_C \langle\delta_{-i}, i{:}a\rangle)))).$$

According to the previous definition, an agent i is fair with respect to his group C (noted $\mathsf{Fair}_{i,C}$) if and only if, if he chooses action a then for every alternative action b, there exists a joint action δ_{-i} of the other agents that he considers possible such that, playing a while the others play δ_{-i} is for group C at least as good as playing b while the others play δ_{-i}. Reciprocity can be defined from individualistic rationality and fairness. For every $i \in Agt$ and $C \in 2^{Agt*}$ such that $i \in C$:

$$\mathsf{Rec}_{i,C} \stackrel{\text{def}}{=} ([\mathsf{B}_i](\bigwedge_{j\in C\setminus\{i\}} \mathsf{Fair}_{j,C}) \rightarrow \mathsf{Fair}_{i,C}) \wedge (\neg[\mathsf{B}_i](\bigwedge_{j\in C} \mathsf{Fair}_{j,C}) \rightarrow \mathsf{IRat}_i).$$

That is, an agent i is a reciprocator with respect to his group C (noted $\mathsf{Rec}_{i,C}$) if and only if, if he believes that the other agents in $C \setminus \{i\}$ play fair then he plays fair, otherwise he plays egoistic. For notational convenience we write Fair_i instead of $\mathsf{Fair}_{i,Agt}$ and Rec_i instead of $\mathsf{Rec}_{i,Agt}$.

It has to be noted that the previous two notions of social rationality are also positively and negatively introspective. For every $i \in Agt$ and $C \in 2^{Agt*}$ we have:

(19) $\vdash_{\mathbf{L}^+} \mathsf{Fair}_{i,C} \leftrightarrow [\mathsf{B}_i]\mathsf{Fair}_{i,C}$

(20) $\vdash_{\mathbf{L}^+} \neg\mathsf{Fair}_{i,C} \leftrightarrow [\mathsf{B}_i]\neg\mathsf{Fair}_{i,C}$

(21) $\vdash_{\mathbf{L}^+} \mathsf{Rec}_{i,C} \leftrightarrow [\mathsf{B}_i]\mathsf{Rec}_{i,C}$

(22) $\vdash_{\mathbf{L}^+} \neg\mathsf{Rec}_{i,C} \leftrightarrow [\mathsf{B}_i]\neg\mathsf{Rec}_{i,C}$

The following $\mathbf{L}^+$-theorems specify some sufficient conditions for fairness equilibrium. For every $\delta \in \Delta$ we have:

(23) $\vdash_{\mathbf{L}^+} (\bigwedge_{i\in Agt} (\mathsf{Fair}_i \wedge [\mathsf{B}_i]\langle\delta_{-i}\rangle\top) \wedge \langle\delta\rangle\top) \rightarrow \mathsf{FairEq}(\delta)$

(24) $\vdash_{\mathbf{L}^+} (\bigwedge_{i\in Agt} (\mathsf{Rec}_i \wedge [\mathsf{B}_i](\bigwedge_{j\neq i} \mathsf{Fair}_j) \wedge [\mathsf{B}_i]\langle\delta_{-i}\rangle\top) \wedge \langle\delta\rangle\top) \rightarrow \mathsf{FairEq}(\delta)$

(25) $\vdash_{\mathbf{L}^+} (\bigwedge_{i\in Agt} (\mathsf{Rec}_i \wedge \neg[\mathsf{B}_i](\bigwedge_{j\neq i} \mathsf{Fair}_j) \wedge [\mathsf{B}_i]\langle\delta_{-i}\rangle\top) \wedge \langle\delta\rangle\top) \rightarrow \mathsf{Nash}(\delta)$

According to $\mathbf{L}^+$-theorem 23 if all agents are fair and every agent knows the choices of the other agents, then the selected strategy profile is a fairness equilibrium. According to $\mathbf{L}^+$-theorem 24 if all agents are reciprocator, every agent believes that all other agents are fair, and every agent knows the choices of the other agents, then the selected strategy profile is a fairness equilibrium. According to $\mathbf{L}^+$-theorem 25 if all agents are reciprocator, every agent does not believe that all other agents are fair, and every agent knows the choices of the other agents, then the selected strategy profile is a Nash equilibrium.

6 Related works

As emphasized in the introduction, although social preferences and social rationality have been extensively studied in experimental economics, no logical analysis of these concepts has been proposed up to now. However, several logical systems exist which support reasoning about strategic games (see, e.g., [16, 27, 19]) and epistemic strategic games (see, e.g., [10, 24, 7, 20]). Let us briefly discuss the latter.

De Bruin [10] has developed a very rich logical framework which enables to reason about the epistemic aspects of strategic games and of extensive games. His system deals with several game-theoretic concepts like the concepts of knowledge, rationality, Nash equilibrium, iterated strict dominance, backward induction. Nevertheless, de Bruin's approach differs from our logical approach to epistemic strategic games in several respects. First of all, our approach is *minimalistic* since it relies on few primitive concepts: joint action, belief, (individual and group) preference. All other notions such Nash equilibrium and (individual and social) rationality are defined by means of these three primitive concepts. On the contrary, in de Bruin's logic all those notions are atomic propositions

managed by a *ad hoc* axiomatization (see, e.g., [10, pp. 61,65] where special propositions for rationality are introduced). Secondly, we provide a semantics and a complete axiomatics for our logic of epistemic games. De Bruin's approach is purely syntactic: no model-theoretic analysis of games is proposed nor completeness result for the proposed logic is given.

Roy [24] has recently proposed a modal logic integrating preference, knowledge and intention. In his approach every world in a model is associated to a nominal which directly refers to a strategy profile in a strategic game. This approach is however limited in expressing formally the structure of a strategic game. In particular, in Roy's logic there is no principle like the **L** Axiom **Indep** explaining how possible actions δ_i of individual agents are combined to form a strategy profile δ of the current game. Another limitation of Roy's approach is that it does not enable to express the concept of individualistic rationality that we have been able to define in Section 3.3 (see [24, pp. 101]).

Bonanno [7] integrates modal operators for belief, common belief with constructions expressing agents' preferences over individual actions and strategy profiles, and applies them to the semantic characterization of solution concepts like Iterated Deletion of Strictly Dominated Strategies (IDSDS). Although this logic enables to express the concept of individualistic rationality assumed in classical game theory, it is not sufficiently general to enable to express in the object language solution concepts like Nash equilibrium and IDSDS (note that the latter is defined by Bonanno only in the metalanguage).

Lorini et al. [20] have recently developed a modal logic of epistemic strategic games which overcomes the previous limitations. It enables to express in the object language solution concepts such as Nash equilibrium and IDSDS, and the concept of individualistic rationality. A complete axiomatization of this logic is given and its complexity is studied. However this logic does not have operators for group preference which support reasoning about different kinds of social rationality such as fairness and reciprocity.

References

[1] R. J. Aumann and A. Brandenburger. Epistemic conditions for Nash equilibrium. *Econometrica*, 63:1161–1180, 2008.

[2] P. Battigalli and G. Bonanno. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53:149–225, 1999.

[3] P. Battigalli and M. Dufwenberg. Dynamic psychological games. *Journal of Economic Theory*, 144:1–35, 2009.

[4] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York, 2001.

[5] J. van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):13–45, 2007.

[6] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, Cambridge, 2001.

[7] G. Bonanno. A syntactic approach to rationality in games with ordinal payoffs. In *8th Conference on Logic and the Foundations of Game and Decision Theory (LOFT 2008)*, Texts in Logic and Games Series, pages 59–86. Amsterdam University Press, 2008.

[8] C. Boutilier. Modal logics for qualitative possibility and beliefs. In *Proceedings of the Eight Annual Conference on Uncertainty in Artificial Intelligence (UAI'92)*, pages 17–24. Morgan Kaufmann, 1992.

[9] A. Brandenburger. Knowledge and equilibrium in games. *Journal of Economic Perspectives*, 6:83–101, 1992.

[10] B. de Bruin. *Explaining games: on the logic of game theoretic explanations*. PhD thesis, University of Amsterdam, The Netherlands, 2004.

[11] F. del Cerro and A. Herzig. A modal analysis of possibility theory. In *Proceedings of Fundamentals of Artificial Intelligence Research (FAIR'91)*, pages 11–18, Springer-Verlag, 1991.

[12] D. Dubois and H. Prade. *Possibility theory: an approach to computerized processing of uncertainty*. Plenum Press, New York, 1988.

[13] E. Fehr and K. M. Schmidt. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114:817–868, 1999.

[14] E. Fehr and K. M. Schmidt. Theories of fairness and reciprocity: Evidence and economic applications. In *Advances in Economics and Econometrics*. Cambridge University Press, 2003.

[15] J. Geanakoplos, D. Pearce, and E. Stacchetti. Psychological games and sequential rationality. *Games and Economic Behavior*, 1:60–79, 1989.

[16] W. van der Hoek, W. Jamroga, and M. Wooldridge. A logic for strategic reasoning. In *Proceedings of the Fourth International Joint Conference on Autonomous Agens and Multiagent Systems (AAMAS 2005)*, pages 157–164, ACM Press, 2005. New York.

[17] N. Laverny and J. Lang. From knowledge-based programs to graded belief-based programs, part II: off-line reasoning. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'05)*, pages 497–502. Professional Book Center, 2005.

[18] D. Lewis. *Counterfactuals*. Harvard University Press, 1973.

[19] E. Lorini. A dynamic logic of agency II: deterministic DLA, Coalition Logic, and game theory. *Journal of Logic, Language and Information*. forthcoming.

[20] E. Lorini, F. Schwarzentruber, and A. Herzig. Epistemic Games in Modal Logic: Joint Actions, Knowledge and Preferences all together. In X. He, J. F. Horty, and E. Pacuit, editors, *LORI-II Workshop on Logic, Rationality and Interaction*, pages 212–226. Springer-Verlag, 2009.

[21] M. J. Osborne and A. Rubinstein. *A course in game theory*. MIT Press, 1994.

[22] M. Rabin. Incorporating fairness into game theory and economics. *American Econonic Review*, 83(5):1281–1302, 1993.

[23] J. Rawls. *A theory of Justice*. Harvard University Press, Cambridge, 1971.

[24] O. Roy. *Thinking before acting: intentions, logic, rational choice*. PhD thesis, University of Amsterdam, The Netherlands, 2008.

[25] H. Sahlqvist. Completeness and correspondence in the first and second order semantics for modal logics. In *Proceedings 3rd Scandinavian Logic Symposium 1973*, volume number 82 in Studies in Logic, 1975.

[26] W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In W. L. Harper and B. Skyrms, editors, *Causation in decision, belief change and statistics*, pages 105–134. Kluwer, 1998.

[27] N. Troquard, W. van der Hoek, and M. Wooldridge. A logic of games and propositional control. In *Proceedings of the 8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'09)*. ACM Press, 2009.

CNRS - INPT - UPS - UT1 - UTM

Abstract

We propose a modal logic that enables to reason about different kinds of rationality in strategic games. This logic integrates the concepts of joint action, belief, individual preference and group preference. The first part of the article is focused on the notion of individualistic rationality assumed in classical game theory: an agent decides to perform a certain action only if the agent believes that this action is a best response to what he expects the others will do. The second part of the article explores different kinds of social rationality such as fairness and reciprocity. Differently from individualistically rational agents (*alias* self-interested agents), social rational agents also consider the benefits of their choice for the group. Moreover, their decisions can be affected by their beliefs about other agents' willingness to act for the well-being of the group. In the article we also provide a complete axiomatization of our logic of joint action, belief, individual preference and group preference.

Keywords

Modal logic, game theory

UPS-IRIT, 118 route de Narbonne, 31062 Toulouse CEDEX 9
+33 (0) 561 55 67 65 info@irit.fr www.irit.fr