From self-regarding to other-regarding agents in strategic games: a logical analysis

Emiliano Lorini

Université de Toulouse, CNRS, IRIT, France

ABSTRACT.
I propose a modal logic that enables to reason about self-regarding and other-regarding motivations in strategic games. This logic integrates the concepts of joint action, belief, individual and group payoff. The first part of the article is focused on self-regarding agents. A self-regarding agent decides to perform a certain action only if he believes that this action maximizes his own personal benefit. The second part of the article explores different kinds of other-regarding motivations such as fairness and reciprocity. Differently from self-regarding agents, other-regarding agents also consider the benefits of their choices 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. The analysis also considers team-directed reasoning, i.e. the mode of reasoning that people use when they take themselves to be acting as members of a group or a team.

KEYWORDS: game theory, modal logic, rationality

DOI:10.3166/JANCL.21.1–33

1. Introduction

In recent times, 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 extended in order to incorporate new important concepts such as fairness, inequity aversion, altruism and reciprocity. Several models of social preferences 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 (Fehr & Schmidt, 2003), we can distinguish two families of models of social preferences 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 (Fehr & Schmidt, 1999), 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. For example, according to Rabin's model of reciprocity (Rabin, 1993), 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 have been extensively studied in the field of behavioral game theory (Camerer, 1997), no logical analysis of social preferences in game-theoretic contexts has been proposed up to now. In this article, I try to fill this gap by proposing a sound and complete modal logic which enables to model different kinds of social preferences in strategic games. 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, as emphasized above, people often take into account of the effects of their actions not only on themselves but on the others as well, that is, people are also concerned about the benefit other agents derive from a given situation. Therefore, to take the presence of other-regarding motivations 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. Section 2 discusses some conceptual issues about the distinction between self-regarding and other-regarding motivations. In Section 3 I present a modal logic integrating the concepts of joint action, belief and individual payoff. Section 4 is devoted to the analysis in this logic of some basic game-theoretic concepts such as best response and Nash equilibrium, and of the notion of self-regarding agent. In Section 6 the base logic is extended with a notion of group payoff based on Rawls's *maximin* criterion of distributive justice. A philosophical introduction to this criterion is provided in Section 5. Section 7 explores two different kinds of other-regarding motivations: fairness and reciprocity. Differently from self-regarding agents, other-regarding 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 Section 8 I compare the logical characterization of other-regarding agents of Section 7 with some economic theories of social preferences and reciprocity which have been proposed in recent times. Section 9 extends the logical analysis to team-directed reasoning. Team-directed reasoning is a mode of reasoning alternative to the best-response reasoning of classical game theory that people use when they take themselves to be acting as members of a group or a team. Finally, Section 10 discusses related works in the area of modal logics for game theory and preference representation.

2. Self-regarding vs. other-regarding agents

Before proceeding into the logical analysis, some conceptual clarifications are in order. Following theories of social preferences in social sciences (Margolis, 1982) and

in economics (Gintis, 2009, Chapter. 3), I distinguish *self-regarding* agents (Section 4.3) from *other-regarding* agents (Section 7.2). A *self-regarding* agent is an agent who acts in order to achieve his *private interests* and to maximize the *personal benefit* he gets in a given situation. On the other hand, an *other-regarding* agent is an agent who is also concerned about the benefit other agents derive from a given situation. In other words, if an agent is self-regarding, the utility of a given state for him coincides with the personal benefit he will get in this state. If the agent is other-regarding, the utility of a given state for him also depends on the benefit the other agents will get in this state. In this sense, his utility might differ from his own personal benefit.

The notion of self-regarding agent studied in this paper should not be confused with the *rationality assumption* of classical game theory: according to classical game theory, individuals are rational in the sense that they maximize their utility. The notions of self-regarding agent and other-regarding agent are not in contradiction with this assumption. We can safely say that an other-regarding agent acts to maximize his utility even though he does not act to maximize his personal benefit (as he also cares about the benefit for the other agents). For example, an agent i may decide to lend his bike to another agent j in order to satisfy agent j's desire to go for a bike ride in the countryside (suppose j does not have a bike). In this situation, the utility of a given action for agent i depends on the extent to which this action promotes the satisfaction of agent j's desires. Therefore, agent i's act is other-regarding even though agent i is still acting to maximize his utility.

It has to be noted there are two possible interpretations (or views) of the concepts of self-regarding and other-regarding agents, depending on how we interpret the preceding notions of 'private interest' and 'personal benefit': (1) according to one view, an agent's private interest coincides with the accumulation of material resources (or material payoffs) such as money, food, etc. and his personal benefit with the quantity of material resources that are available to him; (2) according to the other view, an agent's private interests should be reduced to his own desires, and his personal benefit to the satisfaction of his own desires. The first view is the one commonly taken by economic theories of social preferences. Its main advantage is that it allows to test theories of social preferences experimentally by letting people play games with material incentives. The advantage of the second view is that it provides a higher level of generality, as an agent's desire may be different from the the desire of accumulating material resources. So, according to the first view, a self-regarding agent is an agent who acts in order to maximize his own material payoff, while an other-regarding agent is an agent who is also concerned about the distribution of material payoffs among the other players. According to the second view, a self-regarding agent is an agent who acts in order to maximize the satisfaction of his own desires, while an other-regarding agent is an agent who is also concerned about the satisfaction of other agents' desires. The logic I will present in the rest of the paper is sufficiently general to allow both interpretations. Therefore, I do not need to commit to one of them here.

3. A logic of joint action, beliefs and individual payoffs

I present in this section the modal logic MLEG (*Modal Logic of Epistemic Games*) integrating the concepts of joint action, belief and individual payoff. 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.

3.1. *Syntax*

The syntactic primitives of the logic MLEG are a finite set of agents Agt, a 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 natural integers $I = \{x \in \mathbb{N} \mid 0 \leq x \leq \mathsf{max}\}$ with $\mathsf{max} \in \mathbb{N}$. I stipulate that $|Agt| > 1$ in order to exclude the trivial case $|Agt| = 1$ in which the agent's doxastic relation is the identity relation. Non-empty sets of agents are called *coalitions* or *groups*, noted $C_1, C_2, \ldots$ I note $2^{Agt*} = 2^{Agt} \setminus \emptyset$ the set of coalitions. To every agent $i \in Agt$ I associate the set Act_i of all possible ordered pairs a_i, that is, $Act_i = \{a_i \mid a \in Act\}$. Besides, for every coalition C, I 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 $C = Agt$, I write Δ instead of Δ_{Agt} and α, β, γ, $\delta, \ldots$ instead of α_{Agt}, β_{Agt}, γ_{Agt}, $\delta_{Agt}, \ldots$ for the elements in Δ. Elements in Δ are also called strategy profiles. Given $\delta \in \Delta$, I note δ_i the element in δ corresponding to agent i. Moreover, for notational convenience, I sometimes write δ_{-i} instead of $\delta_{Agt \setminus \{i\}}$. The language $\mathcal{L}$ of the logic MLEG is given by the following BNF:

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

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 δ_C reads "coalition C chooses the joint action δ_C". The operator $\Box$ is the universal modality, which quantifies over all worlds of the current model. Since there are a one-to-one correspondence between models and strategic games (i.e. a model corresponds to a unique strategic game) and a one-to-one correspondence between strategy profiles and worlds in a model (i.e. a world in a model is identified with a strategy profile of the game), $\Box$ is used to quantify over the strategy profiles of the current game. $\Box\varphi$ reads "φ holds for all strategy 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". Operators $[\geq_i^k]$ are used to order strategy profiles according to the benefit agents obtain from them. There are two alternative readings of these operators, depending on how the notion of 'personal benefit' is interpreted (see Section 2 for a discussion). If personal benefit is interpreted as satisfaction of individual desires, then the formula $[\geq_i^k]\varphi$ has to be read "φ is true in all states which are *desirable* for agent i at least to degree k"; if it is interpreted as individual accumulation of material resources, then $[\geq_i^k]\varphi$ has to be read "φ is true in all states which are *profitable* for agent i at least to

degree k". I provide a third more general reading of formula $[\geq_i^k]\varphi$ which does not commit to one of these two interpretations: "φ is true in all states in which agent i gets a payoff of at least k".

Moreover, the following abbreviations are given:

$$\Diamond\varphi \stackrel{\text{def}}{\equiv} \neg\Box\neg\varphi,$$
$$\langle \mathsf{B}_i\rangle\varphi \stackrel{\text{def}}{\equiv} \neg[\mathsf{B}_i]\neg\varphi,$$
$$\langle \geq_i^k\rangle\varphi \stackrel{\text{def}}{\equiv} \neg[\geq_i^k]\neg\varphi.$$

$\Diamond\varphi$ has to be read "φ is possibly true", whereas $\langle \geq_i^k\rangle\varphi$ has to read "φ is true in at least one state in which agent i gets a payoff of least k". $\langle \mathsf{B}_i\rangle\varphi$ has to be read "agent i thinks that φ is possible".

3.2. *Semantics*

DEFINITION 1 (MLEG-MODEL). — MLEG-*models are tuples* $M = \langle W, \mathcal{A}, \mathcal{B}, \mathcal{P}, \pi\rangle$ *where:*

– W *is a nonempty set of possible worlds or states;*

– $\mathcal{A}$ *is a collection of total functions* $\mathcal{A}_C : W \longrightarrow \Delta_C$ *one for every coalition* $C \in 2^{Agt*}$*, mapping every world in* W *to a joint action of the coalition such that:*

C1 $\mathcal{A}_C(w) = \delta_C$ *if and only if, for every* $i \in C$ $\mathcal{A}_i(w) = \delta_i$,[1]

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

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

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

C4 *if* $v \in \mathcal{B}_i(w)$, *then* $\mathcal{A}_i(w) = a_i$ *if and only if* $\mathcal{A}_i(v) = a_i$;

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

C5 $\mathcal{P}_i^0 = W$, *and*

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

– $\pi : Atm \longrightarrow 2^W$ *is a valuation function;*

and $\mathcal{B}_i(w) = \{v \mid (w, v) \in \mathcal{B}_i\}$.

For every coalition C, $\mathcal{A}_C(w) = \delta_C$ 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 this means 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, $(w, v) \in \mathcal{B}_i$ means that at world w, according to agent i, world v is possible. Finally, the function $\mathcal{P}$ is used to specify the personal benefit that an agent gets in a given

1. For notational convenience I write $\mathcal{A}_i$ instead of $\mathcal{A}_{\{i\}}$ for the singletons.

world. In particular, for every $k \in I$ and $i \in Agt$, $\mathcal{P}_i^k$ is the set of worlds in which agent i gets a payoff of at least k. Again payoffs can be interpreted either in terms of desire satisfaction or in terms of accumulation of material resources.

According to Constraint **C1** in Definition 1, 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. I 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**). More abstractly, this means that $\mathcal{A}_{Agt}$ is an injective function. Therefore, every δ behaves as a nominal in the sense of hybrid logic.

REMARK 2. — Although the Constraint **C3** excludes pure uncertainty about uncertainty (i.e. states where the agents do the same thing, but have different beliefs), it is imposed here for two reasons. The first reason is that I want to establish a clear correspondence between MLEG-models and normal form games in the classical sense (i.e. a MLEG-model should correspond to a unique game in normal form). The Constraint **C3** is needed to ensure that a unique payoff is assigned to each pure strategy profile, as required by the classical representation of normal form games (Osborne & Rubinstein, 1994). The second reason is that, if we remove the Constraint **C3**, the complexity of the satisfiability problem of the logic MLEG becomes EXPTIME-hard. This is due to the fact that the complexity of the satisfiability problem of the bimodal logic KD45^2 extended with universal modality is EXPTIME-hard, which is is a straightforward corollary of (Hemaspaandra, 1996, Theorem 5.1). On the contrary, the Constraint **C3** ensures that the logic MLEG is NP-complete (see Section 3.3 for more details). □

Constraint **C4** just says that an agent i chooses the action a (or an agent i decides to perform the action a) if and only if he believes this. This is a standard assumption in interactive epistemology and epistemic analysis of games (see (Bonanno, 2008) for instance). According to Constraint **C5**, in every world an agent gets a payoff of at least 0. This implies that an agent can always compare the payoffs of two different worlds. According to the Constraint **C6**, for every integer $k \in I \setminus \{0\}$, the set of worlds in which an agent gets a payoff of at least k is a subset of the set of worlds in which the agent gets a payoff of at least $k-1$. A function $\kappa_i : W \longrightarrow I$ can be associated with every agent i, where $\kappa_i(w)$ corresponds to the *exact* payoff that agent i gets at world w.

DEFINITION 3 (κ_i). — *For every $i \in Agt$ and for every $w \in W$ I define:*

– $\kappa_i(w) = n$ *if and only if* $w \in \mathcal{P}_i^n$*;*

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

The function κ_i is similar to Spohn's *ordinal conditional function* (OCF) (Spohn, 1998). It has to be noted that Spohn's OCFs are functions into the class of ordinals,

while utility functions used in economics are functions into the set of reals. Here, for simplicity it is assumed that payoffs are measured on the finite integer scale I.

EXAMPLE 4. — Consider the well-known two-player Prisoner's Dilemma (PD) game in Fig. 1 (Osborne & Rubinstein, 1994). 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 Prisoner's Dilemma four different payoffs are needed $I = \{0, 1, 2, 3\}$. The elements $W, \mathcal{A}$ and $\mathcal{P}$ of the model M corresponding to the two-player PD game are defined as follows:

– $W = \{w_1, w_2, w_3, w_4\}$;
– $\mathcal{A}_{\{i_1,i_2\}}(w_1) = \langle c_{i_1}, c_{i_2}\rangle$, $\mathcal{A}_{\{i_1,i_2\}}(w_2) = \langle d_{i_1}, d_{i_2}\rangle$, $\mathcal{A}_{\{i_1,i_2\}}(w_3) = \langle c_{i_1}, d_{i_2}\rangle$, $\mathcal{A}_{\{i_1,i_2\}}(w_4) = \langle d_{i_1}, c_{i_2}\rangle$;
– $\mathcal{P}^3_{i_1} = \{w_4\}$, $\mathcal{P}^2_{i_1} = \{w_1, w_4\}$, $\mathcal{P}^1_{i_1} = \{w_1, w_2, w_4\}$, $\mathcal{P}^0_{i_1} = \{w_1, w_2, w_3, w_4\}$;
– $\mathcal{P}^3_{i_2} = \{w_3\}$, $\mathcal{P}^2_{i_2} = \{w_1, w_3\}$, $\mathcal{P}^1_{i_2} = \{w_1, w_2, w_3\}$, $\mathcal{P}^0_{i_2} = \{w_1, w_2, w_3, w_4\}$.

Therefore, $\kappa_{i_1}(w_3) = \kappa_{i_2}(w_4) = 0$, $\kappa_{i_1}(w_2) = \kappa_{i_2}(w_2) = 1$, $\kappa_{i_1}(w_1) = \kappa_{i_2}(w_1) = 2$ and $\kappa_{i_1}(w_4) = \kappa_{i_2}(w_3) = 3$. □

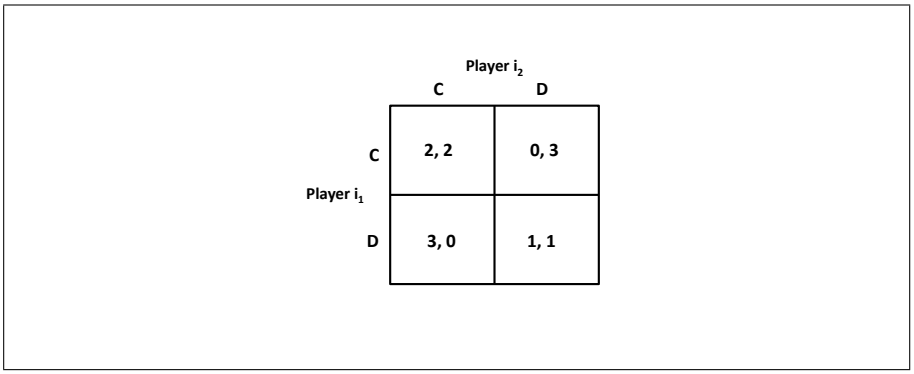

Figure 1. *Prisoner's dilemma*

DEFINITION 5 (TRUTH CONDITIONS). — *Truth of a formula in a model at a given world is defined as follows:*

– $M, w \models p$ *iff* $w \in \pi(p)$;
– $M, w \not\models \bot$;
– $M, w \models \neg\varphi$ *iff* $M, w \not\models \varphi$;
– $M, w \models \varphi \vee \psi$ *iff* $M, w \models \varphi$ *or* $M, w \models \psi$;
– $M, w \models \delta_C$ *iff* $\mathcal{A}_C(w) = \delta_C$;
– $M, w \models \Box\varphi$ *iff* $M, v \models \varphi$ *for all* $v \in W$;
– $M, w \models [\mathsf{B}_i]\varphi$ *iff* $M, v \models \varphi$ *for all* $v \in \mathcal{B}_i(w)$;
– $M, w \models [\geq^k_i]\varphi$ *iff* $M, v \models \varphi$ *for all* $v \in \mathcal{P}^k_i$.

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

3.3. *Axiomatization*

Table 1. *Axiomatization of* MLEG

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

I call MLEG the logic that is axiomatized by the principles given in Table 1, and I write $\vdash_{\mathrm{MLEG}} \varphi$ if φ is a theorem of MLEG. The proof of the following Theorem 6

is based on a straightforward adaptation of the canonical model argument (Blackburn et al., 2001, Chap. 4).[2]

THEOREM 6. — *The logic* MLEG *is sound and complete with respect to the class of* MLEG*-models.*

Clearly the satisfiability problem of MLEG is NP-complete. Indeed, the satisfiability problem of MLEG is NP-hard because MLEG is a conservative extension of classical propositional logic whose satisfiability problem in NP-complete. Moreover, if a formula φ is MLEG-satisfiable, then there exists a MLEG-model $M = \langle W, \mathcal{A}, \mathcal{B}, \mathcal{P}, \pi \rangle$ whose size is bounded by $\mathit{card}(Act)^{\mathit{card}(Agt)}$ which satisfies φ. The following is a non-deterministic algorithm to check if a given formula φ is satisfiable:

– guess non-deterministically a MLEG-model $M = \langle W, \mathcal{A}, \mathcal{B}, \mathcal{P}, \pi \rangle$ whose size is bounded by $\mathit{card}(Act)^{\mathit{card}(Agt)}$ where π only assigns a truth value to propositions occurring in φ;
– guess non-deterministically a world w of M;
– check if $M, w \models \varphi$.

This algorithm non-deterministically runs in polynomial time. So the satisfiability problem of MLEG is in NP.

Before concluding this section, I would like to emphasize an aspect of the present work that could be interesting for a game-theorist. As the logic MLEG is sound and complete, it allows to study strategic interaction both at the semantic level and at the syntactic level. Note that the syntactic derivations of various results concerning the epistemic foundations of game theory are not interesting in itself. Instead, this kind of analysis is useful to identify specific features that are important for the foundations of game theory: for example whether certain assumptions on the players' knowledge and beliefs are indeed necessary to prove results concerning the epistemic conditions of equilibrium notions such as Nash equilibrium. Typical assumptions on players' knowledge and beliefs are for example the assumption that beliefs (and knowledge) are positively and negatively introspective (Axioms 4 and 5 for $[\mathsf{B}_i]$), or the factivity principle for knowledge that knowing that φ implies that φ is true. This point will be discussed in more detail in Section 4.3.

4. Strategic games with self-regarding agents

This section is devoted to the analysis in the logic MLEG of strategic games in which agents are self-regarding. I first define ordering over formulas. Then, I consider the basic game-theoretic concepts of best response and Nash equilibrium, and their relationships with the notion of self-regarding agent.

2. The entire proof can be found in the extended version of this paper available at `http://www.irit.fr/~Emiliano.Lorini/`

4.1. *Ordering of formulas*

For every $i \in Agt$, I define an ordering over formulas:

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

$\psi \leq_i \varphi$ means that "φ is for agent i at least as good as ψ" (i.e. in the states in which φ is true agent i receives a payoff at least as large as the payoff he receives in the states in which ψ is true). As the following proposition highlights the comparative statement $\psi \leq_i \varphi$ might also be read as "for every ψ-state there is a φ-state which is at least as good".

PROPOSITION 7. — *$M, w \models \psi \leq_i \varphi$ if and only if for every $v \in W$, if $M, v \models \psi$ then there is $u \in W$ such that $\kappa_i(v) \leq \kappa_i(u)$ and $M, u \models \varphi$.*

There is no consensus in the literature on how preferential statements between formulas should be defined. In (van Benthem et al., 2009) other kinds of preference comparisons between formulas (i.e. between sets of states) are defined. For instance, the previous $\forall\exists$-reading of preference statements is distinguished from a $\forall\forall$-reading ("for every ψ-state and for every φ-state the φ-state is at least as good as the ψ-state") and a $\exists\exists$-reading ("there are a ψ-state and a φ-state such that the φ-state is at least as good as the ψ-state"). As the following proposition highlights, in the case of comparison between strategy profiles the $\forall\exists$-reading and the $\forall\forall$-reading coincide.

PROPOSITION 8. — *If there is $u \in W$ such that $M, u \models \delta$, then the following two items are equivalent for every $\delta, \delta' \in \Delta$:*

($\forall\exists$) *for every $v \in W$, if $M, v \models \delta'$ then there is $u \in W$ such that $\kappa_i(v) \leq \kappa_i(u)$ and $M, u \models \delta$;*

($\forall\forall$) *for every $v, u \in W$, if $M, v \models \delta'$ and $M, u \models \delta$ then $\kappa_i(v) \leq \kappa_i(u)$.*

The proof of the direction "($\forall\forall$) implies ($\forall\exists$)" is just straightforward. The direction "($\forall\exists$) implies ($\forall\forall$)" follows from Constraint **C3** (i.e. $\mathcal{A}_{Agt}$ is injective). Indeed, Constraint **C3** ensures that, *if* "there is $u \in W$ such that $\kappa_i(v) \leq \kappa_i(u)$ and $M, u \models \delta$" *then* "for all $u \in W$, if $M, u \models \delta$ then $\kappa_i(v) \leq \kappa_i(u)$".

The following are some central properties of the operator $\leq_i$:

$$\vdash_{\text{MLEG}} \psi \leq_i \psi \tag{1}$$

$$\vdash_{\text{MLEG}} ((\varphi_1 \leq_i \varphi_2) \wedge (\varphi_2 \leq_i \varphi_3)) \rightarrow (\varphi_1 \leq_i \varphi_3) \tag{2}$$

$$\vdash_{\text{MLEG}} (\varphi_1 \leq_i \varphi_2) \vee (\varphi_2 \leq_i \varphi_1) \tag{3}$$

$$\vdash_{\text{MLEG}} \bot \leq_i \top \tag{4}$$

$$\text{if } \vdash_{\text{MLEG}} \varphi \rightarrow (\psi_1 \vee ... \vee \psi_s) \text{ then } \vdash_{\text{MLEG}} (\varphi \leq_i \psi_1) \vee ... \vee (\varphi \leq_i \psi_s) \tag{5}$$

The MLEG-theorems 1-3 highlight that $\leq_i$ is a total preorder. The MLEG-theorems 2-4 are the three fundamental principles of Lewis's conditional logic (Lewis, 1973).

4.2. *Best response and Nash equilibrium*

Some basic concepts of game theory can be expressed in MLEG. I 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 $\mathsf{BR}(a_i,\delta_{-i})$, if and only if i cannot improve his payoff by deciding to do something different from a while the others choose the joint action δ_{-i}, that is:

$$\mathsf{BR}(a_i,\delta_{-i}) \stackrel{\text{def}}{\equiv} \bigwedge_{b\in Act} ((b_i \wedge \delta_{-i}) \leq_i (a_i \wedge \delta_{-i})).$$

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}:

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

EXAMPLE 9. — It is well-known that in the Prisoner's Dilemma the only Nash equilibrium is mutual defection. Hence, in the model illustrated in the Example 4 the formula $\mathsf{Nash}(\langle d_{i_1}, d_{i_2}\rangle)$ is true at each world w_1, w_2, w_3, w_4 of the model M. □

4.3. *Self-regarding agents*

The following MLEG formula characterizes a choice criterion for self-regarding agents. It corresponds to the choice criterion commonly assumed in epistemic analysis of games (Battigalli & Bonanno, 1999; van Benthem, 2007):

$$\mathsf{Self}_i \stackrel{\text{def}}{\equiv} \bigwedge_{a,b\in Act} (a_i \rightarrow \bigvee_{\delta\in\Delta} (\langle \mathsf{B}_i\rangle\delta_{-i} \wedge (\langle \delta_{-i}, b_i\rangle \leq_i \langle \delta_{-i}, a_i\rangle))).$$

This means that an agent i is self-regarding, noted Self_i, 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 others 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}. This is the same thing as saying that: if agent i chooses a particular action a then there is no other action b such that, for every joint action δ_{-i} of the others that i considers possible, playing b while the others play δ_{-i} is for i strictly better than playing a while the others play δ_{-i}. More concisely, if agent i chooses a particular action a, then action a is not strictly dominated within his set of epistemic alternatives. For every $i \in Agt$ we have:

$$\vdash_{\mathrm{MLEG}} \mathsf{Self}_i \leftrightarrow [\mathsf{B}_i]\mathsf{Self}_i \tag{6}$$

$$\vdash_{\mathrm{MLEG}} \neg\mathsf{Self}_i \leftrightarrow [\mathsf{B}_i]\neg\mathsf{Self}_i \tag{7}$$

MLEG-theorems 6 and 7 highlight that the property of being a self-regarding agent is positively and negatively introspective. The following theorem specifies 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:

$$\vdash_{\mathrm{MLEG}} (\bigwedge_{i \in Agt} (\mathsf{Self}_i \wedge [\mathsf{B}_i]\delta_{-i}) \wedge \delta) \rightarrow \mathsf{Nash}(\delta) \tag{8}$$

A similar theorem has been stated by Aumann & Brandeburger (Aumann & Brandenburger, 1995). The only difference is that Aumann & Brandeburger used knowledge instead of belief.[3] The following is the syntactic proof of the MLEG-theorem 8:

1. $\vdash (\mathsf{Self}_i \wedge [\mathsf{B}_i]\delta_{-i} \wedge \delta_i) \rightarrow$
$([\mathsf{B}_i]\delta_{-i} \wedge \bigwedge_{b \in Act} \bigvee_{\delta \in \Delta} (\langle \mathsf{B}_i \rangle \delta_{-i} \wedge (\langle \delta_{-i}, b_i \rangle \leq_i \langle \delta_{-i}, a_i \rangle)))$
by definition of Self_i and CPL;
2. $\vdash \delta_{-i} \rightarrow \neg\beta_{-i}$ if $\beta \neq \delta$ by Axiom **Single**;
3. $\vdash [\mathsf{B}_i](\delta_{-i} \rightarrow \neg\beta_{-i})$ if $\beta \neq \delta$ from 2 by necessitation for $[\mathsf{B}_i]$;
4. $\vdash [\mathsf{B}_i]\delta_{-i} \rightarrow [\mathsf{B}_i]\neg\beta_{-i}$ if $\beta \neq \delta$ from 3 by Axiom K for $[\mathsf{B}_i]$;
5. $\vdash ([\mathsf{B}_i]\delta_{-i} \wedge \bigwedge_{b \in Act} \bigvee_{\delta \in \Delta} (\langle \mathsf{B}_i \rangle \delta_{-i} \wedge (\langle \delta_{-i}, b_i \rangle \leq_i \langle \delta_{-i}, a_i \rangle))) \rightarrow$
$(\bigwedge_{\beta \neq \delta} \neg\langle \mathsf{B}_i \rangle \beta_{-i} \wedge \bigwedge_{b \in Act} \bigvee_{\delta \in \Delta} (\langle \mathsf{B}_i \rangle \delta_{-i} \wedge (\langle \delta_{-i}, b_i \rangle \leq_i \langle \delta_{-i}, a_i \rangle)))$
by 4;
6. $\vdash (\bigwedge_{\beta \neq \delta} \neg\langle \mathsf{B}_i \rangle \beta_{-i} \wedge \bigwedge_{b \in Act} \bigvee_{\delta \in \Delta} (\langle \mathsf{B}_i \rangle \delta_{-i} \wedge (\langle \delta_{-i}, b_i \rangle \leq_i \langle \delta_{-i}, a_i \rangle))) \rightarrow$
$(\langle \mathsf{B}_i \rangle \delta_{-i} \wedge \bigwedge_{b \in Act} (\langle \delta_{-i}, b_i \rangle \leq_i \langle \delta_{-i}, a_i \rangle))$
by CPL;
7. $\vdash (\langle \mathsf{B}_i \rangle \delta_{-i} \wedge \bigwedge_{b \in Act} (\langle \delta_{-i}, b_i \rangle \leq_i \langle \delta_{-i}, a_i \rangle)) \rightarrow \bigwedge_{b \in Act} (\langle \delta_{-i}, b_i \rangle \leq_i \langle \delta_{-i}, a_i \rangle)$
by CPL;
8. $\vdash \bigwedge_{b \in Act} (\langle \delta_{-i}, b_i \rangle \leq_i \langle \delta_{-i}, a_i \rangle) \leftrightarrow \mathsf{BR}(a_i, \delta_{-i})$ by definition of $\mathsf{BR}(a_i, \delta_{-i})$;
9. $\vdash (\bigwedge_{i \in Agt} (\mathsf{Self}_i \wedge [\mathsf{B}_i]\delta_{-i}) \wedge \delta) \rightarrow \bigwedge_{i \in Agt} \mathsf{BR}(a_i, \delta_{-i})$ from 1 and 5-8;
10. $\vdash (\bigwedge_{i \in Agt} (\mathsf{Self}_i \wedge [\mathsf{B}_i]\delta_{-i}) \wedge \delta) \rightarrow \mathsf{Nash}(\delta)$ by 9 and definition of $\mathsf{Nash}(\delta)$.

The syntactic proof shows that MLEG-theorem 8 can be proved just by means of Axioms K and the rule of necessitation for the belief operators. In other words, the weakest normal modal logic for belief is sufficient to prove it. Note that, differently from knowledge, beliefs do not satisfy the truth axiom T, i.e. believing that φ does not necessarily imply that φ is true. However, we are still able to conclude from a formula involving beliefs a formula (viz. $\mathsf{Nash}(\delta)$) that does not include any belief operators. This is quite surprising, as one would not expect to be able to do this in a doxastic logic without the truth axiom T. Note also that not even Axiom **JointDet** is required. One can prove MLEG-theorem 8 without assuming that states where the agents do the same thing, have the same beliefs.

3. An earlier version of this theorem can be found in (Spohn, 1982). The difference is that Spohn used probabilities for modelling beliefs, while I here use the normal modal logic KD45.

5. Rawls's *maximin* criterion as a basis for the notion of group payoff

In this second part of the paper I move from individual payoffs to group payoffs. This is the first step for the analysis of other-regarding agents who are motivated not only by their own personal interest, but also by the interest of the group they belong to. The notion of group payoff I formalize is based on Rawls's *maximin* criterion of distributive justice (Rawls, 1971). Some background and clarifications of this criterion are needed in order to ground the logical analysis on a solid conceptual foundation.

RAWLS'S *maximin* CRITERION VS. UTILITARIANISM. Rawls's theory of distributive justice is commonly opposed to the 'utilitarian' theory advanced by (Harsanyi, 1955). In contrast with Rawls, the classical utilitarian program regards individual agents as production units producing individual welfare and its aim consists of maximizing the sum of the individual agents' payoffs. In particular, according to utilitarianism, a given state of affairs w is better than another state of affairs v for a certain group if and only if the sum of the individual payoffs in the state w is higher than the the sum of the individual payoffs in the state v. In response to classical utilitarianism, Rawls proposed the *maximin* criterion of making the least happy agent as happy as possible: for all alternative states w and v, if the level of well-being in the worst-off position is strictly higher in w than in v, then w is better than v. According to this well-known criterion of distributive justice, a fair society should be organized so as to admit economic inequalities to the extent that they are beneficial to the less advantaged agents.[4] For example, suppose that there are only three agents i_1, i_2 and i_3. The state w and the state v ensure the following individual payoffs for the three agents: $\kappa_{i_1}(w) = 1$, $\kappa_{i_2}(w) = 3$, $\kappa_{i_3}(w) = 5$, $\kappa_{i_1}(v) = 2$, $\kappa_{i_2}(v) = 3$ and $\kappa_{i_3}(v) = 3$. Rawls's criterion would select the state v over the state w, because $\min_{i \in \{i_1,i_2,i_3\}} \kappa_i(v) > \min_{i \in \{i_1,i_2,i_3\}} \kappa_i(w)$. A refinement of Rawls's *maximin* principle is the so-called *leximin* principle. The *leximin* principle works by comparing first the payoffs of the less advantaged agents in the state and, if they coincide, by comparing those of the next less advantaged agents, and so on. Once the individual payoffs differ, the *leximin* principle selects the state which maximizes the well-being in the worst-off position. In this sense, the *leximin* principle refines the *maximin* principle when, according to the *maximin* principle, all states ensure the same value of social welfare (i.e. the level of well-being in the worst-off position is equal for all social states). For example, suppose that the state w ensures the following individual payoffs for the three agents i_1, i_2 and i_3: $\kappa_{i_1}(w) = 1$, $\kappa_{i_2}(w) = 2$ and $\kappa_{i_3}(w) = 4$; while the state v ensures the following individual payoffs for the three agents: $\kappa_{i_1}(v) = 3$, $\kappa_{i_2}(v) = 1$ and $\kappa_{i_3}(v) = 4$. According to the *leximin* principle, the state w and the state v are equally good from the point of view of

4. It has to be noted that Rawls's theory of justice is specified in terms of justice over primary goods. Rawls's list of primary goods includes for instance basic liberties and rights, freedom of movement and free choice of occupation, income and wealth, the social bases of self-respect. This difference is however beyond the scope of the present article. See (Sen, 1970) for a discussion on this issue.

the group, because $\min_{i\in\{i_1,i_2,i_3\}} \kappa_i(v) = \min_{i\in\{i_1,i_2,i_3\}} \kappa_i(w)$. On the contrary, according to the *leximin* principle, w is better than v because the payoff of the second 'poorest' agent in w is higher than the payoff of second 'poorest' agent in v.

REQUIREMENTS AND LIMITATIONS OF THE TWO CRITERIA. Rawls's *maximin* criterion and the utilitarian criterion impose different requirements on the representation of preferences. For instance, the application of Rawls's *maximin* criterion requires interpersonal comparison of payoff levels, so-called "level comparability" (see (Sen, 1977) for more details). That is, in order to apply Rawls's criterion, we should be able to say whether an agent i in the state w is better off or worse off than (or the same as) another agent j in the state v. On the contrary, classical utilitarianism (i.e. summation of individual payoffs) requires interpersonal comparison of payoff differences. That is, in order to apply the utilitarian's criterion, we should be able to say whether the difference in payoff between two states w and v for an agent i is higher or lower than (or equal to) the difference in payoff between the same two states or some other states for another agent j. It follows that simple orderings over states of the form $\leq_i$, just saying whether a state w is at least as good as another state v for a given agent i (i.e. $v \leq_i w$), are insufficient for the application of Rawls's *maximin* criterion and of the utilitarian criterion (see Section 10.2).
Futhermore, both criteria have some limitations that is important to be clear on. Rawls's criterion, for instance, prefers one being very well off and all other quite bad to one being still worse off and all others quite well. This is the reason why "[...] Our values about inequality cannot be adequately reflected in the *maximin* rule" (Sen, 1970, p. 139). For example, suppose that the state w and the state v ensure the following individual payoffs for the three agents i_1, i_2 and i_3: $\kappa_{i_1}(w) = 10$, $\kappa_{i_2}(w) = 8$, $\kappa_{i_3}(w) = 4$, $\kappa_{i_1}(v) = 10$, $\kappa_{i_2}(v) = 5$ and $\kappa_{i_3}(v) = 5$. The *maximin* rule indicates that v is better than w. The problem is that, while the gap between i_2 and i_3 is reduced, that between i_1 and i_2 is accentuated in v. On the other hand, as pointed out by Sen (Sen, 1970), classical utilitarianism à la Harsanyi may sacrifice the welfare of few agents for the sake of improving total welfare. The paradigmatic example is that of a slave society in which *few* slaves serve *many* free men in such a way that the individual welfare of a slave decreases from 2 to 0, while the individual welfare of a free man increases from 2 to 4. According to utilitarianism, this society - that most people would consider morally unacceptable - is preferred to a society in which all men are free and the individual welfare of everyone is equal to 2.[5]

RAWLS'S CRITERION AND PRIORITARIANISM. Rawls's theory of justice has been the historical starting point for the conception of distributive justice called *prioritarianism* (see, e.g., (Parfit, 1997; Nagel, 1991)). The main concern of prioritarianism is to correct classical utilitarianism by holding that a benefit has greater moral value

5. As emphasized by Sen, another problem with classical utilitarianism is that it is indifferent between an unequal distribution and an equal distribution when the total welfare is the same for the two distributions. For example, suppose that: $\kappa_{i_1}(w) = 1$, $\kappa_{i_2}(w) = 1$, $\kappa_{i_1}(v) = 2$ and $\kappa_{i_2}(v) = 0$. Utilitarianism is indifferent between the state w and the state v.

the worse the situation of the individual to whom it accrues. That is, according to prioritarianism, "[...] Benefiting people matter more the worse off these people are" (Parfit, 1997, p. 213).[6] Rawls's *maximin* criterion can be seen as the extreme version of prioritarianism which gives infinitely greater weight to the benefits of a worse-off person: according to Rawls's criterion giving any benefit - no matter how small - to a worse-off person is better than giving a benefit - no matter how large - to a better-off person. According to prioritarianism the priority to the worse off is not absolute, since "[...] benefits to the worse off could be morally outweighed by sufficiently great benefits to the better off" (Parfit, 1997, p. 213). This is the reason why *strict prioritarianism* à la Rawls has been distinguished by several authors from *finitely weighted prioritarianism* which gives only finitely weight to benefits for those who are worse off (Vallentyne, 2007; Arneson, 2007). It has to be noted that Sen's criticism to Rawls's *finitely weighted prioritarianism* discussed above does not apply to *finitely weighted prioritarianism*.

RAWLS'S CRITERION AND OTHER-REGARDING MOTIVATIONS. Before concluding this section, I want to explain why I have decided to base the concept of group payoff on Rawls's *maximin* criterion. There are two main reasons: (1) Rawls's *maximin* criterion is representative of prioritarianism and (2) prioritarianism is a good basis for modelling other-regarding motivations in strategic situations. The first point has been clarified in the preceding paragraph. Let me now clarify the second point. When we are driven by other-regarding motivations, the elevation of other people to a minimally decent standard of existence seems overwhelmingly important for us, consequently, we give more weight to improving the well-being of those who are worse off against increasing the benefit of those better off. In this sense, the prioritarian conception of distributive justice is a basic constituent of other-regarding motivations. Nagel (Nagel, 1991) presents a similar argument in justifying prioritarianism as the basis of the impersonal standpoint, in contrast with the individual standpoint. According to Nagel, when assuming the individual standpoint, an agent simply acts in the pursuit of his own interests (i.e. the agent is self-regarding). On the contrary, from the impersonal standpoint, one regards the social world as a benevolent outsider whose aim is to promote impartiality among individuals (i.e. the agent is other-regarding). As pointed out above, I am aware of the limitations of Rawls's criterion. However, the present work should be seen as a first attempt to provide a logical characterization of group payoffs in strategic situations based on a prioritarian view of social justice. A logical analysis of group payoffs based on a less strict prioritarian criterion such

6. Prioritarianism is commonly opposed to *strict egalitarianism* (Vallentyne, 2007; Parfit, 1997) whose aim is to diminish (or eliminate) interpersonal differences in individual payoffs. The most striking criticism to strict egalitarianism is the so-called *levelling down objection*: since strict egalitarianism values the reduction of the gap between the better and the worse off for its own sake, it regards a worsening of the condition of the better off as valuable even though this is does not benefit anyone. Many find this counter-intuitive. For example, suppose that: $\kappa_{i_1}(w) = 4$, $\kappa_{i_2}(w) = 1$, $\kappa_{i_1}(v) = 2$ and $\kappa_{i_2}(v) = 1$. According to strict egalitarianism, v is better than w even though v is better for no one and worse for some.

as finitely weighted prioritarianism goes beyond the objectives of this work and will pursued in future research.

6. A logic of joint action, beliefs, individual and group payoffs

In the sequel, I will provide a MLEG-extension with group payoffs based on Rawls's *maximin* criterion of distributive justice. I will first present its syntax (Section 6.1). I will then specify its semantics (Section 6.2) and provide an axiomatization (Section 6.3).

6.1. *Syntax*

I extend the logic MLEG with operators of group payoffs of the form $[\geq^k_C]$ with $C \in 2^{Agt*}$ and I call MLEG$^+$ the resulting logic. More precisely, the language of the logic MLEG$^+$ is defined by the following BNF:

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

where p ranges over Atm, i ranges over Agt, δ_C ranges over $\bigcup_{C\in 2^{Agt*}} \Delta_C$, 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 in which group C gets a payoff of at least k". I define $\langle\geq^k_C\rangle\varphi \stackrel{\text{def}}{\equiv} \neg[\geq^k_C]\neg\varphi$.

6.2. *Semantics*

DEFINITION 10 (MLEG$^+$-MODEL). — *MLEG$^+$-models are tuples $M = \langle W, \mathcal{A}, \mathcal{B}, \mathcal{P}, \mathcal{GP}, \pi\rangle$ where:*

– *$M = \langle W, \mathcal{A}, \mathcal{B}, \mathcal{P}, \pi\rangle$ is a* MLEG*-model;*

– *$\mathcal{GP} : I \times 2^{Agt*} \longrightarrow 2^W$ is a function mapping every integer k in I and coalition C to a (possibly empty) set of worlds $\mathcal{GP}^k_C$ such that:*

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

The function $\mathcal{GP}$ is used to specify the payoff that a group of agents gets in a given world. In particular, for every $k \in I$ and $C \in 2^{Agt*}$, $\mathcal{GP}^k_C$ is the set of worlds in which group C gets a payoff of at least k. According to the Constraint **C7** the payoff of a group C is determined by the intersection of the payoffs of the agents in the group.

DEFINITION 11 (κ_C). — *For every $C \in 2^{Agt*}$ and for every $w \in W$ I define:*

– *$\kappa_C(w) = n$ if and only if $w \in \mathcal{GP}^n_C$;*

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

DEFINITION 12 (Max_C). — *For every* $C \in 2^{Agt*}$ *I define* $\mathsf{Max}_C = \underset{w \in W}{\operatorname{argmax}}\, \kappa_C(w)$.

Max_C is the set of best worlds for the group C. The following propositions highlight some interesting aspects of the present notion of group payoff.

PROPOSITION 13. — *For every* $C \in 2^{Agt*}$ *and for every* $w \in W$ *we have* $\kappa_C(w) = \min_{i \in C} \kappa_i(w)$.

According to Proposition 13, the payoff of world w for the group of agents C is equal to the lowest payoff of w for an agent in C.

PROPOSITION 14. — *For every* $C \in 2^{Agt*}$ *we have* $\mathsf{Max}_C = \underset{w \in W}{\operatorname{argmax}}\, (\min_{i \in C} \kappa_i(w))$.

Proposition 14, which follows from Definition 12 and Proposition 13, highlights that the present notion of group payoff is based on Rawls's *maximin* criterion of distributive justice discussed above. It is worth noting that the *maximin* principle ensures *weak* Pareto optimality.

PROPOSITION 15. — *If* $w \in \mathsf{Max}_C$ *then* $w \in \mathsf{WPareto}_C$, *with* $\mathsf{WPareto}_C = \{v \in W | \forall\, u \in W, \exists i \in C\colon \kappa_i(u) \leq \kappa_i(v)\}$.

EXAMPLE 16. — Consider again the two-player Prisoner's Dilemma (PD) game introduced in Section 4. As to group payoffs, we have: $\mathcal{GP}^2_{\{i_1,i_2\}} = \{w_1\}$, $\mathcal{GP}^1_{\{i_1,i_2\}} = \{w_1, w_2\}$ and $\mathcal{GP}^0_{\{i_1,i_2\}} = \{w_1, w_2, w_3, w_4\}$. Therefore, $\mathsf{Max}_{\{i_1,i_2\}} = \{w_1\}$. Indeed, w_1 is the unique state whose lowest payoff for an agent in $\{i_1, i_2\}$ is maximal. □

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

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

6.3. *Axiomatization*

As the following Proposition 17 highlights the expressive power of MLEG^+ is no higher than the expressive power of MLEG, as formulas containing group payoff operators can be reduced to Boolean combinations of individual payoff operators.

PROPOSITION 17. — *The following schemata is valid:*

$$\langle \geq^k_C \rangle \varphi \leftrightarrow \Big(\bigvee_{\delta \in \Delta} \bigwedge_{i \in C} \langle \geq^k_i \rangle (\delta \wedge \varphi) \Big) \qquad \textbf{(GroupPref)}$$

Given a MLEG^+ formula φ, let $red(\varphi)$ be the formula obtained by iterating the application of the reduction axiom **GroupPref** from the left to the right. Thanks to the rule of replacement of equivalents (REP) it is clear that $\varphi \leftrightarrow red(\varphi)$ is valid. Furthermore, the length of $red(\varphi)$ is exponential in the maximum number of nestings of group payoff operators within φ. Therefore, the logic MLEG^+ allows to represent the

notion of group payoff in a more compact way than the logic MLEG. The following Theorem 18 provides an axiomatization result for the logic MLEG^{+}.

THEOREM 18. — *The logic* MLEG^{+} *is completely axiomatized by the axioms and inference rules of* MLEG *together with the schemata* ***GroupPref*** *in Proposition 17.*

It has to be noted that the rule of replacement of equivalents does not need to be added to the axiomatization, as it can be derived from the other principles. Indeed, one can show that the rule of necessitation for the group payoff operators is derivable. Here it is the proof:

1. $\vdash \varphi$ — by hypothesis;
2. $\vdash [\geq_i^k]\varphi$ — by rule of necessitation for $[\geq_i^k]$;
3. $\vdash [\geq_i^k](\delta \to \varphi)$ — from 2 by rule of necessitation and Axiom K for $[\geq_i^k]$;
4. $\vdash \bigwedge_{\delta\in\Delta}\bigvee_{i\in C}[\geq_i^k](\delta \to \varphi)$ — from 3 by CPL;
5. $\vdash [\geq_C^k]\varphi$ — from 4 by Axiom **GroupPref** and CPL.

Then, by the previous rule of necessitation and Axiom **GroupPref**, one can prove the following rule of monotony (RM): $\frac{\varphi\to\psi}{[\geq_C^k]\varphi\to[\geq_C^k]\psi}$. Finally, the rule of replacement of equivalents can be easily proved by means of the previous rule of monotony for the group payoff operators, together with the rules of monotony for all modal operators of the base logic MLEG. We just need to adapt the inductive proof given in (Chellas, 1980, Theorem 4.7).

In the rest of the paper, I write $\vdash_{\mathrm{MLEG}^{+}} \varphi$ if formula φ is a theorem of MLEG^{+}. The following are some properties of the group payoff operators. For every $B, C \in 2^{Agt*}$ such that $B \subseteq C$ and for every $k \in I$ we have:

$$\vdash_{\mathrm{MLEG}^{+}} [\geq_B^k]\varphi \to [\geq_C^k]\varphi \tag{9}$$

$$\vdash_{\mathrm{MLEG}^{+}} [\geq_C^0]\varphi \leftrightarrow \Box\varphi \tag{10}$$

$$\vdash_{\mathrm{MLEG}^{+}} [\geq_C^{k-1}]\varphi \to [\geq_C^k]\varphi \tag{11}$$

$$\vdash_{\mathrm{MLEG}^{+}} [\geq_i^k]\varphi \leftrightarrow [\geq_{\{i\}}^k]\varphi \tag{12}$$

For instance, according to the MLEG^{+}-theorem 12, individual payoff operators are just group payoff operators for singleton groups.

7. Strategic games with other-regarding agents

In this section I explore strategic games in which agents are driven by other-regarding motivations. I consider two different kinds of other-regarding motivations: fairness and reciprocity.

7.1. *Social-welfare equilibrium*

I first generalize the ordering over formulas for an agent to an ordering over formulas for a group. For every $C \in 2^{Agt*}$, I define $\psi \leq_C \varphi$ ("φ is for group C at least as good as ψ") as follows:

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

As the following MLEG^+-theorems highlight the dyadic operators $\leq_C$ too are total preorders:

$$\vdash_{\mathrm{MLEG}^+} \psi \leq_C \psi \tag{13}$$

$$\vdash_{\mathrm{MLEG}^+} ((\varphi_1 \leq_C \varphi_2) \wedge (\varphi_2 \leq_C \varphi_3)) \rightarrow (\varphi_1 \leq_C \varphi_3) \tag{14}$$

$$\vdash_{\mathrm{MLEG}^+} (\varphi_1 \leq_C \varphi_2) \vee (\varphi_2 \leq_C \varphi_1) \tag{15}$$

For instance, MLEG^+-theorem 14 can be proved in three steps as follows:

1. $\vdash ((\varphi_1 \leq_C \varphi_2) \wedge (\varphi_2 \leq_C \varphi_3)) \leftrightarrow$
   $\bigwedge_{k \in I}((\langle \geq^k_C \rangle \varphi_1 \rightarrow \langle \geq^k_C \rangle \varphi_2) \wedge (\langle \geq^k_C \rangle \varphi_2 \rightarrow \langle \geq^k_C \rangle \varphi_3))$
   by definition of $\psi \leq_C \varphi$ and CPL;
2. $\vdash ((\varphi_1 \leq_C \varphi_2) \wedge (\varphi_2 \leq_C \varphi_3)) \rightarrow \bigwedge_{k \in I}(\langle \geq^k_C \rangle \varphi_1 \rightarrow \langle \geq^k_C \rangle \varphi_3)$
   by 1 and CPL;
3. $\vdash ((\varphi_1 \leq_C \varphi_2) \wedge (\varphi_2 \leq_C \varphi_3)) \rightarrow (\varphi_1 \leq_C \varphi_3)$ from 1,2 by definition of $\psi \leq_C \varphi$.

The notion of best response of Section 4.2 can now be generalized to groups. I say that agent i's action a is for the group C a best response to the other agents' joint action δ_{-i}, noted $\mathsf{BR}_C(a_i,\delta_{-i})$, if and only if i cannot improve the payoff of group C by deciding to do something different from a while the others choose the joint action δ_{-i}, that is:

$$\mathsf{BR}_C(a_i,\delta_{-i}) \stackrel{\text{def}}{\equiv} \bigwedge_{b \in Act} ((b_i \wedge \delta_{-i}) \leq_C (a_i \wedge \delta_{-i})).$$

Note that the notion of best response of Section 4.2 is just a special case of the previous definition, when $C = \{i\}$.

Before concluding this section, I define a notion of social-welfare equilibrium as the collective counterpart of the notion of Nash equilibrium. A similar notion has been studied by Charness & Rabin in the context of economic theory of social preferences (Charness & Rabin, 2002, p. 852). Given a certain strategic game, the strategy profile (or joint action) δ is said to be a *social-welfare 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}:

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

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

EXAMPLE 19. — In the Prisoner's Dilemma there are two social-welfare equilibria: mutual defection and mutual cooperation. Hence, in the model illustrated in the Example 4 of Section 3.2 the formulas $\mathsf{SW}(\langle d_{i_1}, d_{i_2}\rangle)$ and $\mathsf{SW}(\langle c_{i_1}, c_{i_2}\rangle)$ are both true at each world w_1, w_2, w_3, w_4 of the model M. The former social-welfare equilibrium coincides with the unique Nash equilibrium of this game. □

EXAMPLE 20. — Consider the game "Chicken" in Fig. 2. Each player chooses either to attack the other player (the action "dare") or to back down (the action "chicken"). The best situation for a player is when he attacks while the other backs down, while the worst is when he backs down while the other attacks. A player prefers the situation in which both players back down to the situation in which they both attack. This game has two Nash equilibria: (D,C) and (C,D). Moreover, it has a unique social-welfare equilibrium (C,C). An interesting aspect of this game is that its set of Nash equilibria and its set of social-welfare equilibria are disjoint. This is different from the Prisoner's Dilemma in which the two sets are different but not disjoint. □

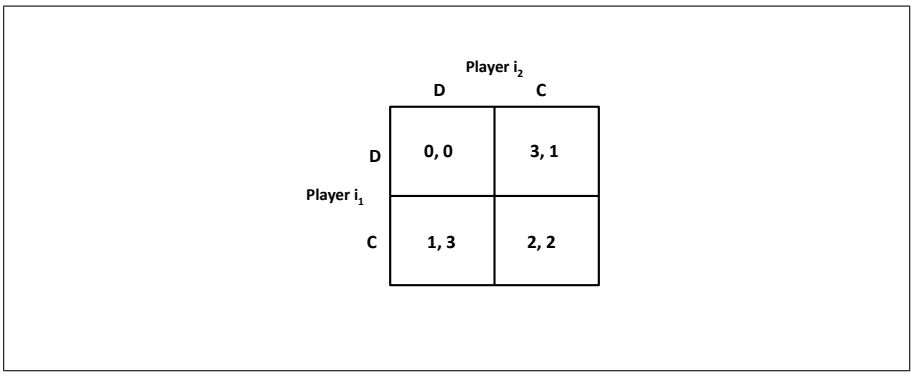

Figure 2. *Chicken*

7.2. *Other-regarding agents*

It is time to look at other-regarding motivations. Differently from self-regarding agents defined in Section 4.3, other-regarding 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. I first define fairness. For every $i \in Agt$ and $C \in 2^{Agt*}$ such that $i \in C$:

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

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 is a *conditional* form of choice criterion which can be defined as follows. For every $i \in Agt$ and $C \in 2^{Agt*}$ such that $i \in C$:

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

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 egoistically. For notational convenience I 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 types of other-regarding motivations are also positively and negatively introspective:

$$\vdash_{\mathrm{MLEG}^+} \mathsf{Fair}_{i,C} \leftrightarrow [\mathsf{B}_i]\mathsf{Fair}_{i,C} \tag{16}$$

$$\vdash_{\mathrm{MLEG}^+} \neg\mathsf{Fair}_{i,C} \leftrightarrow [\mathsf{B}_i]\neg\mathsf{Fair}_{i,C} \tag{17}$$

$$\vdash_{\mathrm{MLEG}^+} \mathsf{Rec}_{i,C} \leftrightarrow [\mathsf{B}_i]\mathsf{Rec}_{i,C} \tag{18}$$

$$\vdash_{\mathrm{MLEG}^+} \neg\mathsf{Rec}_{i,C} \leftrightarrow [\mathsf{B}_i]\neg\mathsf{Rec}_{i,C} \tag{19}$$

The following MLEG^+-theorems specify some sufficient conditions for social-welfare equilibrium and Nash equilibrium under fairness and reciprocity:

$$\vdash_{\mathrm{MLEG}^+} (\bigwedge_{i \in Agt} (\mathsf{Fair}_i \wedge [\mathsf{B}_i]\delta_{-i}) \wedge \delta) \rightarrow \mathsf{SW}(\delta) \tag{20}$$

$$\vdash_{\mathrm{MLEG}^+} (\bigwedge_{i \in Agt} (\mathsf{Rec}_i \wedge [\mathsf{B}_i](\bigwedge_{j \neq i} \mathsf{Fair}_j) \wedge [\mathsf{B}_i]\delta_{-i}) \wedge \delta) \rightarrow \mathsf{SW}(\delta) \tag{21}$$

$$\vdash_{\mathrm{MLEG}^+} (\bigwedge_{i \in Agt} (\mathsf{Rec}_i \wedge \neg[\mathsf{B}_i](\bigwedge_{j \neq i} \mathsf{Fair}_j) \wedge [\mathsf{B}_i]\delta_{-i}) \wedge \delta) \rightarrow \mathsf{Nash}(\delta) \tag{22}$$

$$\vdash_{\mathrm{MLEG}^+} (\bigwedge_{i \in Agt} (\mathsf{Rec}_i \wedge [\mathsf{B}_i]\delta_{-i} \wedge [\mathsf{B}_i](\bigwedge_{j \neq i} [\mathsf{B}_j]\delta_{-j})) \wedge \delta) \rightarrow (\mathsf{Nash}(\delta) \vee \mathsf{SW}(\delta)) \tag{23}$$

According to MLEG^+-theorem 20 if all agents are fair and every agent knows the choices of the others, then the selected strategy profile is a social-welfare equilibrium. According to MLEG^+-theorem 21 if every agent is a reciprocator, every agent believes that all others are fair, and every agent knows the choices of the others, then the selected strategy profile is a social-welfare equilibrium. According to MLEG^+-theorem 22 if all agents are reciprocators, every agent does not believe that all other agents are fair, and every agent knows the choices of the others, then the selected

strategy profile is a Nash equilibrium. According to MLEG^{+}-theorem 23 if all agents are reciprocators, every agent knows the choices of the others, every agent knows that every agent knows the choices of the others, then the selected strategy profile is either a social-welfare equilibrium or a Nash equilibrium. It is worth noting that, in the cases of self-regarding and fair agents, equilibrium just requires first-order beliefs over other players' choices (MLEG-theorem 8 and MLEG^{+}-theorem 20). On the contrary, in the case of reciprocity, equilibrium requires *first-order* and *second-order* beliefs over other players' choices (on this point see also (Rabin, 1993)). Indeed, the formula $(\bigwedge_{i\in Agt}(\mathsf{Rec}_i \wedge [\mathsf{B}_i]\delta_{-i}) \wedge \delta) \rightarrow (\mathsf{Nash}(\delta) \vee \mathsf{SW}(\delta))$ is not valid in MLEG^{+}.

8. A comparison with economic theories of social preferences

The logical analysis of other-regarding motivations presented above is supported by some experimental results obtained in behavioral game theory. For instance, there are experiments with people playing the Prisoner's Dilemma for money, in which the proportion of participants choosing cooperate is typically around the 50 percent (Sally, 1995). In some studies of "Chicken" a high proportion of participants chose the social-welfare equilibrium ("chicken", "chicken") (Rutström et al., 1994; Camerer, 1997). Several economic theories have been proposed in recent times which extend classical game theory with the concept of social preference in order to explain the phenomena observed in experiments with humans. I here compare the model of other-regarding motivations presented in Section 7 with some of these theories.

THEORIES OF INEQUITY AVERSION AND FAIRNESS As pointed out in Section 2, economic models of social preferences start from the idea that a player's choice is determined not only by self-interested motivations but also by social motivations (Margolis, 1982). In this sense, the utility of a given state (or strategy profile) for a player is determined not only by the payoff that the player gets in this state but also by the payoffs of the other players. Technically, this amounts to defining a utility function for an agent in which different weights are assigned to the agent's individual payoff and to the other agents' payoffs (the higher is the weight of the agent's individual payoff and the more self-regarding the agent is, the higher is the weight of the other agents' payoffs and the more other-regarding the agent is).
For instance, in the models proposed by (Fehr & Schmidt, 1999) and (Bolton & Ockenfels, 2000), players are assumed to be intrinsically motivated to distribute payoffs in an equitable way: a player dislikes being either better off or worse off than another player. In other terms, utilities are calculated in such a way that equitable allocations of payoffs are preferred by all players.
(Charness & Rabin, 2002) consider a specific form of social preferences they call *quasi-maximin preferences*. In Charness & Rabin model, group payoff is computed by means of a social welfare function which is a *weighted* combination of Rawls' *maximin* and of the utilitarian welfare function (i.e. summation of individual payoffs) discussed in Section 5 (see (Charness & Rabin, 2002, p. 851)). The weight of the 'Rawlsian' component of the social welfare function is assumed to be ϵ with $\epsilon \in [0, 1]$

and the weight of the 'utilitarian' component is assumed to be $\epsilon - 1$. According to Charness & Rabin, a fair player is a player who computes the utility of a given strategy profile by means of this social-welfare function. The notion of fair player I presented in Section 7.2 can be seen as a special case of Charness & Rabin's notion: the case in which social welfare is measured solely according to Rawls' *maximin* criterion (i.e. the 'Rawlsian' component of the social welfare function has weight 1 and the 'utilitarian' component has weight 0). Charness & Rabin also introduce a notion of social-welfare equilibrium which corresponds to the notion of social-welfare equilibrium formalized in Section (see 7.1 (Charness & Rabin, 2002, p. 852)). In Charness & Rabin's model a given strategy profile δ is said to be a social-welfare equilibrium if and only if, it is a Nash equilibrium of the game in which the utility of a given strategy profile for a player is computed by means of the social-welfare function.

RABIN'S MODEL OF RECIPROCITY In his model of reciprocity (Rabin, 1993), Rabin adopts the concept of 'psychological game theory' introduced by (Geanakoplos et al., 1989) in which it is assumed that utilities do not only depend on terminal-node payoffs but also on players' beliefs. Rabin assumes that a player's utility is determined not only by the payoff he gets in a given situation, but also by his beliefs about the dispositions of other players. In particular, if a player i believes that another player j will be kind with him then i will be more inclined to be kind with j and to sacrifice his own personal interest for the interest of j. On the contrary, if i believes that j will be unkind with him then i will have a desire to retaliate. The most important similarity between Rabin's concept of reciprocity and the concept I defined in Section 7.2 is that they both highlight the *conditional* nature of reciprocity: reciprocity should be seen as a player's disposition to behave generously *if* the others are behaving generously and to behave egoistically *if* the others are behaving egoistically (and eventually to punish them *if* they are mean to the player). This latter aspect of punishment is explicit in Rabin's model, but it is not addressed in my model. I postpone a logical analysis of punishment to future work.

9. Towards team reasoning

The game in Fig. 3, called the Hi-Lo Matching game, has received quite a lot of attention in recent times. If both players choose the option H (the action "high"), each gains two units; if both choose L (the action "low"), each gains one unit; otherwise neither gains anything. There are two Nash equilibria in this game: the situation in which both players choose H (H,H), and the situation in which they both choose L (L,L). Hi-Lo is similar to a pure coordination game: the interests of the players are perfectly aligned and there are two Nash equilibria. But, differently from a pure coordination game, in Hi-Lo one of the two Nash equilibria is strictly better than the other, as it yields a higher payoff to both players. In this case, we say that (H,H) payoff-dominates (L,L). This is the reason why the coordination problem in the Hi-Lo seems trivial. It is clear that the players should coordinate on the preferred equilibrium (H,H). These intuitions have been confirmed by experiments with humans playing the

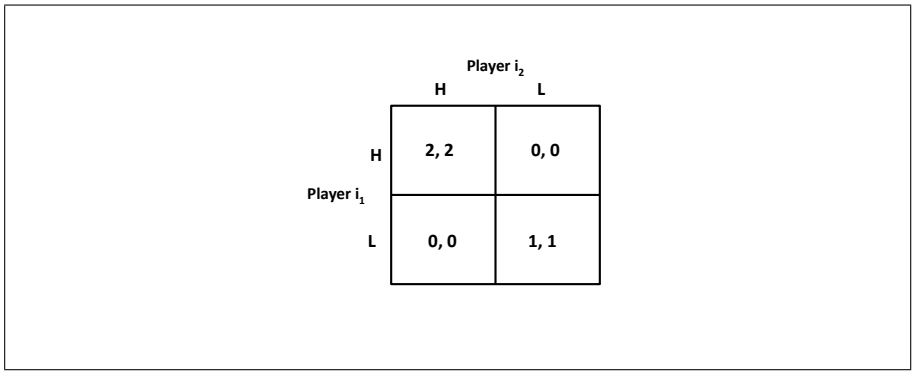

Figure 3. *Hi-Lo*

Hi-Lo game with material payoffs: for instance it is shown in (Bacharach, 2006) that a very large majority of people choose the option H and coordinate on the preferred equilibrium (H,H).

The Hi-Lo game presents a fundamental problem for the best-response reasoning assumed in classical game theory (i.e. agents choose their best response to what they expect the others will do) and, consequently, for the notion of self-regarding agent I have presented in Section 4. In fact, from the assumptions that the players are self-regarding and that they know the other players' choices, we cannot deduce that each player will choose H.[7] All we can say is that if a player expects that the other player will choose H then it is rational for him to choose H. But exactly the same can be said about L: if a player expects that the other player will choose L then it is rational for him to choose L. Theories of social preferences too fail to explain why people coordinate on the preferred equilibrium in the Hi-Lo game. Consider for instance the notion of social-welfare equilibrium introduced in Section 7.1. In the Hi-Lo game both (H,H) and (L,L) are social-welfare equilibria. Indeed, in (H,H) and in (L,L) every player chooses an action that, given what the other player chooses, maximizes the group payoff of the two players. Therefore, the notion of other-regarding agent studied in Section 7.2 does not help to solve the problem presented by the Hi-Lo game.

In order to solve this problem some theorists have proposed to incorporate new modes of reasoning into game theory by which players can arrive at the conclusion that they ought to choose the option H when facing a Hi-Lo game. For instance, starting from the work of Gilbert (Gilbert, 1989) and Regan (Regan, 1980), some economists have studied team reasoning as a mode of reasoning alternative to the best-response reasoning assumed in classical game theory (Sugden, 2003; Sugden, 2000; Bacharach, 1999; Colman et al., 2008). Team-directed reasoning is the kind of reasoning that people use when they take themselves to be acting as members of a group or team (Sugden, 2000). That is, when an agent i engages in team reasoning, he identifies

7. The same deduction cannot be made even assuming that the players are self-regarding and that they have common knowledge of this.

himself as a member of a group of agents C and conceives C as a unit of agency acting as a single entity in pursuit of some collective objective. A team reasoning player acts for the interest of his group by identifying a strategy profile that maximizes the collective payoff of the group, and then, if the maximizing strategy profile is unique, by choosing the action that forms a component of this strategy profile. As pointed out by Sugden (Sugden, 2003), a player has reason to act as a team member and to choose the action that forms a component of the strategy profile maximizing collective payoff, conditional on assurance that the other players also act as team members. That is, to act as a member of a team, one must be confident that the other players act as members too. More fundamentally, "[...] team reasoning does not generate reasons for choice unless each member of a team has reason to believe that there is *common reason* to believe that each member of the team endorses and acts on team reasoning [...]. This is a condition of assurance" (Sugden, 2003, p. 176-177).

I want to show here how the logical framework presented in the previous sections can be easily extended in order to formalize this form of team reasoning. To this aim, I use $[\mathsf{EB}_C]\varphi$ as an abbreviation of $\bigwedge_{i\in C}[\mathsf{B}_i]\varphi$, i.e. every agent in C believes that φ (if $C = \emptyset$ then $[\mathsf{EB}_C]\varphi$ is equivalent to $\top$). Then, I define by induction $[\mathsf{EB}^k_C]\varphi$ for every natural number $k \in \mathbb{N}$:

$$[\mathsf{EB}^0_C]\varphi \stackrel{\text{def}}{\equiv} \varphi$$

and for all $k \geq 1$,

$$[\mathsf{EB}^k_C]\varphi \stackrel{\text{def}}{\equiv} [\mathsf{EB}_C]([\mathsf{EB}^{k-1}_C]\varphi).$$

I define for all natural numbers $n \in \mathbb{N}$, $[\mathsf{CB}^n_C]\varphi$ as an abbreviation of $\bigwedge_{1\leq k\leq n}[\mathsf{EB}^k_C]\varphi$. $[\mathsf{CB}^n_C]\varphi$ represents C's mutual belief that φ up to n iterations, i.e. everyone in C believes that φ, everyone in C believes that everyone in C believes that φ, and so on until level n. In order to capture team reasoning in the logic MLEG^+ some new constructions are needed. I add special constants of the form $member(i,C)$, with $i \in C$. $member(i,C)$ has to be read "agent i identifies himself as a member of the group C" or "agent i acts as a member of the group C". In the semantics the constructions $member(i,C)$ are interpreted by means of a collection of total functions

$$\mathcal{T}_C : W \longrightarrow 2^C$$

one for every coalition $C \in 2^{Agt*}$, mapping every world in W to a (possibly empty) subset of the coalition. For any $C \in 2^{Agt*}$ and $w \in W$, $\mathcal{T}_C(w)$ is the set of agents in C who identify themselves as members of C at the world w.[8] Truth condition of $member(i,C)$ is then defined as follows:

$$M, w \models member(i,C) \text{ iff } i \in \mathcal{T}_C(w).$$

8. An interesting alternative is to define a notion of a group membership based on a social network structure (e.g. Facebook), that is to say, the agents in the set C are members of the same group if and only if, every agent in C is *connected* to the other agents in C in the social network. See (Seligman et al., 2011) for a recent work on a modal logic of social networks.

Following (Sugden, 2003) I assume that an agent i acts as a member of the group C only if he believes that all agents in C act as members of C:

$$member(i,C) \rightarrow [\mathsf{B}_i](\bigwedge_{j \in C} member(j,C)) \qquad \textbf{(TeamAct)}$$

The following is the semantic constraint corresponding to the preceding Axiom **TeamAct**. For every $C \in 2^{Agt*}$, $i \in C$ and $w \in W$:

C8 if $i \in \mathcal{T}_C(w)$ then for all $v \in \mathcal{B}_i(w)$ and for all $j \in C$, $j \in \mathcal{T}_C(v)$.

In order to provide a logical formalization of team reasoning, the following additional definition is needed. I say that a given strategy profile δ *uniquely maximizes* the collective payoff of the group C, noted $\mathsf{UniqueBest}(\delta,C)$, if and only if, for all strategy profiles δ' different from δ, δ is for C strictly better than δ':

$$\mathsf{UniqueBest}(\delta,C) \stackrel{\text{def}}{\equiv} \bigwedge_{\delta' \in \Delta, \delta' \neq \delta} (\delta' <_C \delta)$$

with $\delta' <_C \delta \stackrel{\text{def}}{\equiv} (\delta' \leq_C \delta) \wedge \neg(\delta \leq_C \delta')$. Team reasoning is captured by the following logical axiom:

$$(member(i,Agt) \wedge \mathsf{UniqueBest}(\delta,Agt)) \rightarrow \delta_i \qquad \textbf{(TeamReason)}$$

According to Axiom **TeamReason**, if agent i identifies himself as a member of the group Agt and the strategy profile δ uniquely maximizes the collective payoff of Agt, then i chooses the action that forms a component of the strategy profile δ. The following is the semantic constraint corresponding to the preceding Axiom **TeamReason**. For every $\delta \in \Delta$, $i \in C$ and $w \in W$:

C9 if $i \in \mathcal{T}_{Agt}(w)$ and there is w such that $\mathcal{A}_{Agt}(w) = \delta$ and $\kappa_{Agt}(w) > \kappa_{Agt}(v)$ for all $v \neq w$, then $\mathcal{A}_i(w) = \delta_i$.

I call MLEG^{++} the logic axiomatized by the principles of the logic MLEG^{+} given in Section 6.3 plus the Axiom **TeamAct** and the Axiom **TeamReason**. The definition of MLEG^{++} model is a straightforward adaptation of the definition MLEG^{+} model (Definition 10) to which the functions $\mathcal{T}_C$ and the Constraints **C8** and **C9** are added.

By means of Axiom **TeamAct**, we can prove that if an agent i acts as a member of a group C then i believes that there is common belief in C that every agent in C acts as a member of C too. That is, for every $n \in \mathbb{N}$ we have:

$$\vdash_{\mathrm{MLEG}^{++}} member(i,C) \rightarrow [\mathsf{B}_i][\mathsf{CB}^n_C](\bigwedge_{j \in C} member(j,C)) \qquad (24)$$

The previous MLEG^{++}-theorem provides a formal characterization of Sugden's observation that acting as a member of a group entails having reason to believe that there

is common reason to believe that each member of the team acts as a member of the group.[9]

EXAMPLE 21. — Assume $Agt = \{i_1, i_2\}$ and $ACT = \{h, l\}$ where h and l are respectively the actions "high" and "low" in the Hi-Lo game in Fig. 3. A MLEG^{++} model for the Hi-Lo game is a MLEG^{++} model whose elements $W, \mathcal{A}, \mathcal{P}$ and $\mathcal{GP}$ are:

– $W = \{w_1, w_2, w_3, w_4\}$;
– $\mathcal{A}_{\{i_1,i_2\}}(w_1) = \langle h_{i_1}, h_{i_2}\rangle$, $\mathcal{A}_{\{i_1,i_2\}}(w_2) = \langle l_{i_1}, l_{i_2}\rangle$, $\mathcal{A}_{\{i_1,i_2\}}(w_3) = \langle h_{i_1}, l_{i_2}\rangle$, $\mathcal{A}_{\{i_1,i_2\}}(w_4) = \langle l_{i_1}, h_{i_2}\rangle$;
– $\mathcal{P}^2_{i_1} = \mathcal{P}^2_{i_2} = \{w_1\}$, $\mathcal{P}^1_{i_1} = \mathcal{P}^1_{i_2} = \{w_1, w_2\}$, $\mathcal{P}^0_{i_1} = \mathcal{P}^0_{i_2} = \{w_1, w_2, w_3, w_4\}$;
– $\mathcal{GP}^2_{\{i_1,i_2\}} = \{w_1\}$, $\mathcal{GP}^1_{\{i_1,i_2\}} = \{w_1, w_2\}$, $\mathcal{GP}^0_{\{i_1,i_2\}} = \{w_1, w_2, w_3, w_4\}$.

By the Constraint **C9**, it is easy to check that there is no MLEG^{++} model for the Hi-Lo game such that either $\mathcal{T}_{\{i_1,i_2\}}(w_2) = \{i_1, i_2\}$ or $\mathcal{T}_{\{i_1,i_2\}}(w_3) = \{i_1, i_2\}$ or $\mathcal{T}_{\{i_1,i_2\}}(w_4) = \{i_1, i_2\}$. On the contrary there exists a MLEG^{++} model for the Hi-Lo game such that $\mathcal{T}_{\{i_1,i_2\}}(w_1) = \{i_1, i_2\}$. [10] This means that, if i_1 and i_2 identify themselves as members of the group $\{i_1, i_2\}$ in the Hi-Lo game, then they must coordinate on the preferred Nash equilibrium $\langle h_{i_1}, h_{i_2}\rangle$. □

10. Related works on logics for game theory and preference logics

As emphasized in the introduction and in Section 8, although social preferences have been extensively studied in behavioral game theory, no logical analysis of social preferences in game-theoretic contexts has been proposed up to now. However, several logical systems exist which support reasoning about strategic games (van der Hoek et al., 2005; Lorini, 2010), knowledge and beliefs in strategic games (de Bruin, 2004; Roy, 2008; Bonanno, 2008; Lorini & Schwarzentruber, 2010), and agents' preferences. I here discuss some of these logics. I first consider modal logic analysis of the epistemic aspects of strategic games. Then, I compare my logical approach to preferences with some alternative approaches.

10.1. *Knowledge and beliefs in strategic games*

De Bruin (de Bruin, 2004) has developed a very rich logical framework which enables to reason about the epistemic aspects of strategic games and of extensive games.

9. The concept of common reason for belief used by Sugden is inspired by (Lewis, 1969). For simplicity, this concept is here formalized by means of the standard iterative definition of common belief. See (Cubitt & Sugden, 2003) for a more elaborated logical analysis of Lewis' concept of common reason for belief.
10. Note that, by the Constraint **C8**, such a model must satisfy $\mathcal{B}_{i_1}(w_1) = \mathcal{B}_{i_2}(w_1) = \{w_1\}$.

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 my logical approach to epistemic strategic games in several respects. First of all, my approach is *minimalistic* since it relies on few primitive concepts: joint action, belief, preference. All other notions such as Nash equilibrium, self-regarding and other-regarding agents 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., (de Bruin, 2004, pp. 61,65) where special propositions for rationality are introduced). Secondly, I provide a semantics and a complete axiomatics for the logic MLEG and its extension MLEG^{+}. De Bruin's approach is purely syntactic: no model-theoretic analysis of games is proposed nor completeness result for the proposed logic is given.

van der Hoek & Pauly (van der Hoek & Pauly, 2006) investigate how modal logics can be used to describe and reason about games. They show how epistemic logic can be combined with constructions expressing agents' preferences over strategy profiles in order to study the epistemic aspects of strategic games. Although van der Hoek & Pauly discuss the combination of action, preference and epistemics for the analysis of epistemic games they do not provide a unified modal logic framework combining operators for knowledge, for preference and for action with a complete axiomatization.

Bonanno (Bonanno, 2008) 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 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 & Schwarzentruber (Lorini & Schwarzentruber, 2010) 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 rationality. A complete axiomatization of this logic is given and its complexity is studied.

All previous works have nothing to say about group preferences and about different kinds of other-regarding motivations such as fairness and reciprocity. On the contrary, this is one the main contribution of the present work.

10.2. *Modal logics of preferences*

van Benthem & Liu (van Benthem & Liu, 2007) have proposed a modal logic of preference and preference change. Their base logic consists of the standard Boolean constructions, the universal modal operator $\Box$ which quantifies over all worlds of the current model and preference operators $[pref]_i$ parameterized with agents. Every operator $[pref]_i$ of van Benthem & Liu's logic is a S4 normal modal operator interpreted by means of a preference ordering (a preorder) $\preceq_i$ over the worlds in a Kripke model M: $w \preceq_i v$ means that the world v is for agent i at least as good as the world w. The

formula $[pref]_i\varphi$ is true at a given world w of a model M (i.e. $M, w \models [pref]_i\varphi$) if and only if φ is true in all worlds which are for agent i at least as good as the current one (i.e. $M, v \models \varphi$ for all v such that $w \preceq_i v$). van Benthem & Liu's approach has been recently extended by (van Benthem et al., 2009) in order to express different readings of *ceteris paribus* preferences. In this logic one can express the "all other things being equal" reading of *ceteris paribus* preferences and characterize the notion of Nash equilibrium as a preference for a given strategy profile of a game, given that *others* keep the same strategy. Differently from the present approach, van Benthem and coll. only consider individual preferences and do not study their relationships with group preferences.

(Girard & Seligman, 2009) have recently introduced a logic for characterizing the class of aggregation procedures, known as lexicographic re-orderings. Given a particular hierarchy over a set of agents, lexicographic re-ordering is computed by giving priority to the agents further up the hierarchy in a compensating way: the group follows group preferences, if the agents in the group can agree; it follows the most influential agents, in case of disagreement. Girard & Selignam use a variant of Liu & van Benthem's preference logic with modal operators of weak and strict preference extended with nominals. They define lexicographic re-ordering by composition of two basic operations on individual preference orderings over states: intersection and subordination (which is defined in terms of intersection and union). However, lexicographic re-ordering should not be confused with the *leximin* criterion discussed in Section 5. As pointed out in Section 5, Rawls's *maximin* criterion and the *leximin* criterion require interpersonal comparison of payoff levels. Therefore, simple preference orderings over states as in van Benthem & Liu's approach and in Girard & Selignam's approach are not sufficient for their logical characterization.

Recently, Ågotnes and coll. (Ågotnes et al., 2009) have proposed a modal logic for reasoning about coalitional games which includes modal operators of group preference. These operators are of the form $\langle C\rangle$ where C is a set of agents. The formula $\langle C\rangle\varphi$ stands for "the group C prefers φ". In more detail, $\langle C\rangle\varphi$ is intended to mean that there exists a state (possibly different from the current one) which is weakly preferred over the current one by each agent in the group C and in which φ is true. In other words, as in the logic defined in Section 5, the preference of a group is defined by the intersection of the preferences of the agents in the group. The main limitation of Ågotnes and coll.'s logic is that it does not allow to compare payoff levels of different agents. As emphasized above, this is a necessary prerequisite for the logical characterization of Rawls's *maximin* criterion.

11. Conclusion

I have presented a logical analysis of both self-regarding and other-regarding motivations in strategic games. I have first considered self-regarding agents who choose their best response to what they expect the others will do. Then, I have introduced a notion of group payoff and studied different forms of other-regarding motivations such as fairness and reciprocity. At the end of the article, I have focused on the notion

of team reasoning which provides an explanation of why people choose the preferred equilibrium (H,H) in the Hi-Lo game. Directions for future research are manifold. First of all, I postpone to future work a logical characterization of the *leximin* criterion discussed in Section 5. Furthermore, I plan to enrich my framework with a notion of *ceteris paribus* preference as the one studied by (van Benthem et al., 2009). Indeed, *ceteris paribus* preference operators are well-suited to express notions such as best response and Nash equilibrium. It would also be interesting to introduce a new notion of *ceteris paribus* group preferences, as van Benthem et al. only studied individual preferences. Another interesting direction of future research is a generalization of the present approach to mixed strategies. Indeed, at the current stage the logics MLEG and MLEG^{+} only enable to reason about pure strategies. Finally, I plan to develop a logical analysis of punishment which is explicit in Rabin's model of reciprocity (Rabin, 1993) and which has not been taken into account in this paper (see Section 8 for a discussion).

Acknowledgements

The author would like to thank the anonymous reviewers of this paper for their helpful comments and suggestions. This work is supported by the French ANR project TIES "Social ties in economics: experiments and theory".

12. References

Ågotnes, T., van der Hoek, W., & Wooldridge, M. (2009). Reasoning about coalitional games. *Artificial Intelligence*, 173, 45–79.

Arneson, R. J. (2007). Equality. In R. E. Goodin, P. Pettit, & T. Pogge (Eds.), *A Companion to Contemporary Political Philosophy*, volume 2 (pp. 593–611). Wiley-Blackwell.

Aumann, R. J. & Brandenburger, A. (1995). Epistemic conditions for Nash equilibrium. *Econometrica*, 63, 1161–1180.

Bacharach, M. (1999). Interactive team reasoning: a contribution to the theory of cooperation. *Research in economics*, 23, 117–147.

Bacharach, M. (2006). *Beyond individual choice: teams and frames in game theory*. Oxford: Princeton University Press.

Battigalli, P. & Bonanno, G. (1999). Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53, 149–225.

Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal Logic*. Cambridge: Cambridge University Press.

Bolton, G. E. & Ockenfels, A. (2000). A theory of equity, reciprocity and competition. *America Economic Review*, 100, 166–193.

Bonanno, G. (2008). A syntactic approach to rationality in games with ordinal payoffs. In *Proc. of LOFT 2008*, Texts in Logic and Games (pp. 59–86).: Amsterdam Univ. Press.

Camerer, C. (1997). Progress in behavioral game theory. *Journal of economic perspectives*, 11(4), 167–188.

Charness, G. B. & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117, 817–869.

Chellas, B. F. (1980). *Modal logic: an introduction*. Cambridge: Cambridge University Press.

Colman, A. M., Pulford, B. N., & Rose, J. (2008). Collective rationality in interactive decisions: evidence for team reasoning. *Acta Psychologica*, 128, 387–397.

Cubitt, R. P. & Sugden, R. (2003). Common knowledge, salience and convention: a reconstruction of david lewis' game theory. *Economics and Philosophy*, 19, 175–210.

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

Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114, 817–868.

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

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, 1, 60–79.

Gilbert, M. (1989). *On social facts*. London: Routledge.

Gintis, H. (2009). *The bounds of reason: game theory and the unification of the behavioral sciences*. Cambridge: Princeton University Press.

Girard, P. & Seligman, J. (2009). An analytic logic of aggregation. In *Proc. of the Third Indian Conference on Logic and Applications (ICLA'2009)*, volume 5378 of *LNAI* (pp. 148–163).: Springer-Verlag.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63, 309–321.

Hemaspaandra, E. (1996). The price of universality. *Notre Dame Journal of Formal Logic*, 37(2), 174–203.

Lewis, D. (1973). *Counterfactuals*. Harvard University Press.

Lewis, D. K. (1969). *Convention: a philosophical study*. Cambridge: Harvard University Press.

Lorini, E. (2010). A dynamic logic of agency II: deterministic DLA, Coalition Logic, and game theory. *Journal of Logic, Language and Information*, 19(3), 327–351.

Lorini, E. & Schwarzentruber, F. (2010). A modal logic of epistemic games. *Games*, 1(4), 478–526.

Margolis, H. (1982). *Selfishness, Altruism, and Rationality: A Theory of Social Choice*. Chicago: University of Chicago Press.

Nagel, T. (1991). *Equality and Partiality*. Cambridge: Oxford University Press.

Osborne, M. J. & Rubinstein, A. (1994). *A course in game theory*. MIT Press.

Parfit, D. (1997). Equality and priority. *Ratio*, 10, 202–221.

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

Rawls, J. (1971). *A theory of justice*. Cambridge: Harvard University Press.

Regan, D. (1980). *Utilitarianism and cooperation*. Oxford: Clarendon Press.

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

Rutström, E., McDaniel, T., & Williams, M. (1994). Incorporating fairness into game theory and economics: an experimental test with incentive compatible belief elicitation. University of Central Florida, Department of Economics.

Sally, D. (1995). Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, 7, 58–92.

Seligman, J., Liu, F., & Girard, P. (2011). Logic in the community. In *Proc. of the Fourth Indian Conference on Logic and Applications (ICLA'2011)*, volume 6521 of *LNCS* (pp. 178–188).: Springer-Verlag.

Sen, A. (1970). *Collective choice and social welfare*. San Francisco: Holden-Day.

Sen, A. (1977). On weights and measures: informational constraints in social welfare analysis. *Econometrica*, 45(7), 1539–1572.

Spohn, W. (1982). How to make sense of game theory. In W. Balzer, W. Spohn, & W. Stegmüller (Eds.), *Philosophy of Economics*, volume 2 (pp. 239–270). Berlin: Springer.

Spohn, W. (1998). Ordinal conditional functions: a dynamic theory of epistemic states. In *Causation in decision, belief change and statistics* (pp. 105–134). Kluwer.

Sugden, R. (2000). Team preferences. *Economics and Philosophy*, 16, 175–204.

Sugden, R. (2003). The logic of team reasoning. *Philosophical Explorations*, 6(3), 165–181.

Vallentyne, P. (2007). Distributive justice. In R. E. Goodin, P. Pettit, & T. Pogge (Eds.), *A Companion to Contemporary Political Philosophy*, volume 2 (pp. 548–562). Wiley-Blackwell.

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

van Benthem, J., Girard, P., & Roy, O. (2009). Everything else being equal: a modal logic for *ceteris paribus* preferences. *Journal of Philosophical Logic*, 38(1), 83–125.

van Benthem, J. & Liu, F. (2007). Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 17(2), 157–182.

van der Hoek, W., Jamroga, W., & Wooldridge, M. (2005). A logic for strategic reasoning. In *Proc. of AAMAS 2005* (pp. 157–164). ACM Press: New York.

van der Hoek, W. & Pauly, M. (2006). Modal logic for games and information. In P. Blackburn, J. Van Benthem, & F. Wolter (Eds.), *Handbook of Modal Logic (3)*. Amsterdam: Elsevier.