ECOLE POLYTECHNIQUE

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

CHOOSING ONE'S PREFERENCES

Guilhem LECOUTEUX

September 2013

Cahier n° 2013-19

DEPARTEMENT D'ECONOMIE
Route de Saclay
91128 PALAISEAU CEDEX
(33) 1 69333033
http://www.economie.polytechnique.edu/
mailto:chantal.poujouly@polytechnique.edu

Choosing one's preferences

Guilhem Lecouteux*

September 18, 2013

Abstract

A central assumption in economics is that individuals are rational in the sense that they seek to satisfy their preferences, by choosing the action that maximizes a utility function that represents those preferences. However, it appears that in strategic interaction with other rational agents, the individuals can benefit from strategic commitments. We determine the set of games for which strategic commitments can be beneficial to the players, by building an analytical framework in which players can choose their own preferences before playing a game. We show that players make strategic commitments as soon as there exists a Stackelberg equilibrium that is not a Nash equilibrium, but also that there always exists at least one set of preference relations at the equilibrium such that a Nash equilibrium is implemented. We then show that the possibility of making strategic commitments generates cooperative behaviours in the case of supermodular games. *Journal of Economic Literature* Classification Numbers: C72, D01.

Keywords: strategic commitment, choice of preferences, Stackelberg, supermodularity.

1 Introduction

A disturbing issue of rational choice theory is that the theory can be self-defeating. There indeed exists situations in which a person who is perfectly rational, trying to maximize her payoff, can achieve *in fine* a lower outcome that a less rational individual. Sugden (1991) considers for instance two types of games for which rational choice can lead to self-defeating behaviours: coordination games, such as the Hi-Lo game, and games of commitment, such as the Toxin puzzle (Kavka, 1983). The counter-intuitive implication of this observation is that, in those specific games, if an individual wants to achieve her objective, then it is in her interest to adopt an apparently non rational behaviour: being irrational can therefore be rational in those games. The idea that the individuals can benefit from strategic commitments — i.e. voluntary deviations from the rational behaviour — have been suggested by Schelling (1960), and already discussed by Stackelberg (1934), with the introduction of timing in oligopoly. Different approaches have then been developed in order to study specific kinds of commitments, such as strategic delegation

*École Polytechnique, Laboratoire d'économétrie PREG-CECO (CNRS UMR 7176), E-mail: guilhem.lecouteux@polytechnique.edu. Earlier versions of this article were presented to the X-CREST microeconomics seminar at École Polytechnique, the Economics Internal Workshop at University of East Anglia, and the lunch seminar Theory, Organisation and Markets at Paris School of Economics. I thank the participants in those seminars for their comments, and my PhD advisors Francis Bloch and Robert Sugden for numerous and constructive discussions.

(Fershtman and Kalai, 1997, Fershtman and Gneezy, 2001, Sengul et al., 2012), the evolution of preferences (Guth and Yaari, 1992, Samuelson, 2001, Heifetz et al., 2007), but also the role of emotions (Franck, 1987, 1988). In addition, some experimental results suggest that players can learn to make the optimal strategic commitment (Poulsen and Roos, 2012).
We can however notice that those different approaches focus on a specific type of commitment, such as evolutionary stable payoff distortions, contract with third parties or — in the case of strategic delegation — specific mechanisms such as the provision of incentives or the allocation of decision rights (Sengul et al., 2012, p.387). We suggest here adopting a slightly different approach: we will not consider a specific mechanism that could enable the implementation of a strategic commitment, but we will define the conditions under which a strategic commitment can be beneficial. In particular, we develop a notion of equilibrium that is immune to strategic commitments, i.e. a strategy profile such that no players can make a strategic commitment in order to increase her own welfare. The interest of this notion compared to the other approaches is that such an equilibrium is immune against *any* strategic commitment, whereas previous works usually focused on a specific range of strategic commitment, such as payoff distortions.

The main feature of the theories of strategic commitment is the distinction between on the one hand the relation that determines the choice of the individual, and on the other the relation that determines her outcome. In this paper, we keep a similar distinction while distinguishing between preferences and welfare. We refer to the idea of a total subjective comparative evaluation in order to define both notions:

> "To say that Jill prefers x to y is to say that when Jill has thought about everything she takes to bear on how much she values x and y, Jill ranks x above y. [...] Because Jill's total subjective ranking does not leave out anything that she regards as relevant to the evaluation of alternatives, it combines with beliefs to determine her choices." (Hausman, 2012, p.34)

We define a *preference relation* as the relation over the set of actions that determines the choice of the individual, integrating among other things possible strategic commitments. We define a *welfare relation* as the relation over the set of actions that would have determined the choice of the individual if she was not able to make strategic commitments: the welfare relation therefore represents the "true" preferences of the individual, in the sense that it is this last relation that will determine her outcome. We must notice here that phenomena such as altruism or spite — considered for instance in the evolutionary approach as dispositions that create a wedge between the preferences and the payoff —, as well as ethical or religious commitments, are already integrated in the evaluation of the outcomes in terms of welfare. Our point is that a strategic commitment is only a deformation of the preferences that can help achieving a higher level of welfare, whereas altruism is probably more an end in itself, and should be integrated in the welfare of the individual. As argued by Hausman (2012), the logical properties of comparative judgements imply that these welfare and preference relations meet the standard axioms of rationality of completeness and transitivity. It is therefore possible to represent them respectively by a *welfare function* and by a *utility function*.

The central element of our approach is the idea that the players are able to choose their own preferences in a pre-commitment game. We justify this assumption by arguing that it provides a more accurate theory of what should be a *rational* behaviour. The main assumption concerning human behaviour in economics is indeed that individuals are rational in the sense that

they choose the actions that best satisfy their ends: we have therefore an instrumental theory of rationality. This implies that we do not need to question the ends of the individual, but only the choice of the means that are at her disposal[1].
If we consider rational individuals whose end is the maximization of their own welfare, they should choose the procedure that enables them to get the highest level of welfare. In particular, it means that if maximizing one's welfare function leads to a self-defeating behaviour, then an individual who maximizes her own welfare can achieve a lower level of welfare than one who does not. If we consider that the individuals are rational and want to choose the strategy that give them the highest level of welfare, then it can be rational to commit oneself to maximize a different function than one's own welfare function. An equilibrium notion in game theory that enables players to choose preferences different from their welfare can therefore offer a better model of rational behaviour than Nash equilibrium, which presupposes that rational individuals are committed to maximize their own welfare function. Indeed, when the strategy of the other players is given, maximizing one's welfare is probably the best procedure a player can choose in order to get the highest possible level of welfare (it is the definition of a Nash equilibrium). We should however notice that in a game, it is not the chosen strategies, but the welfare functions (i.e. the functions the players want to maximize *in fine*) that are common knowledge: the players try therefore to maximize their welfare, given the *objective* of the other players, and not given the *strategy* of the other players. As rational individuals, it can therefore be in their interest to commit themselves to the maximization of a utility function which is not their welfare function.

We will therefore model the choice of preferences as follows: individuals are characterized by their welfare relation over the space of strategies, and are able to commit themselves to act according to a different relation than their welfare relation. Before choosing their strategies, the players choose their preference relations in a pre-commitment game; they then choose their strategy in order to satisfy those preferences in a second stage: we will therefore represent any game by a two-stage game. A strategy profile in the second stage game and a set of preference relations will be defined as a subgame perfect $\mathcal{P}$-equilibrium if and only if this strategy profile simultaneously maximizes the utility functions of the players, and the set of preference relations which are represented by those utility functions constitute a Nash equilibrium of the pre-commitment game.
Under the assumption of common knowledge of rationality, the players know that each player can have an incentive in choosing a different preference relation than her welfare relation, and therefore that a rational player can make a strategic commitment. If the structure of the game — i.e. the set of players, the available actions and the welfare relations — is common knowledge, then each player knows that the other players can choose a preference relation different from their welfare relation. The two-stage game structure is therefore common knowledge. If it is not the case, then it would be necessary to implement a mechanism such that each player can publicly announce her preferences at the end of the precommitment game.

We firstly show that the players make strategic commitments as soon as one of them can get a first mover advantage: this means that for a very large class of games, rational choice is self-defeating, and it is in the interest of the players to commit themselves to adopt an apparently non

[1]We can find the origins of this tradition on the one hand in the work of Hume (1739), who suggested a fundamental difference between passions (which cannot be subject to rational scrutiny) and reason, considering that we cannot discuss the ends and motives of an individual, which are simply psychological phenomena; and on the other hand in the definition of political economy of Mill (1843, Book VI, Chap.IX, par. 3), and the science of logical actions of Pareto (Pareto (1909, chap.3 , par. 1) and Pareto (1916, par. 2146)), which restrict by definition the scope of economics to the study of human actions guided by the pursuit of self-interest.

rational behaviour. Our second result is that, despite the fact that the players will generally make strategic commitments, any Nash equilibrium can be implemented as a second stage equilibrium. This second result implies in particular that the existence of a Nash equilibrium is a sufficient condition for the existence of a subgame perfect $\mathcal{P}$-equilibrium. These results are quite general, since the only assumption we make is that the welfare relation of the individuals is complete and transitive, and can therefore be represented by an ordinal welfare function. We then show in a more restricted framework that, in supermodular two-players games, both players have an interest in choosing a cooperative utility function in the pre-commitment game. Conversely, at least one player will choose competitive preferences in the case of a submodular two-players game.

The paper is organized as follows. In section 2, we precisely present our framework of the rational choice of preferences, before showing our two main results in section 3. We present in section 4 an analysis of games with continuous strategy sets, and show that supermodularity generates cooperative behaviours. Section 5 concludes.

2 Rational choice of one's preferences

In this first section, we present our framework of the choice of one's own preferences. We firstly describe the two-stage game, and then define equilibrium notions for each stage of the two-stage game.

2.1 Description of a game

Let $N = \{1, \ldots, n\}$ denote the set of players. Suppose that the set of actions A_i of each player i is a non empty subset of a topological space E_i. Let X_i be the set of probability distributions over A_i, i.e. the set of mixed strategies of player i. Denote by $X = \prod_{i \in N} X_i$ the set of strategy profiles in mixed strategies, and by $X_S = \Pi_{i \in S} X_i$ the cartesian product of all strategies of players in a coalition $S \subset N$. For all coalition $S \subset N$, denote by $-S$ the set $-S = \{j \in N \mid j \notin S\}$.
Let $\mathcal{P}_i \subseteq X_i \times X_i$ denote a preference relation on the set X_i, and $\mathcal{P} = \{\mathcal{P}_i\}_{i \in N}$ a set of preference relations on the sets X_i, $\forall i \in N$. $\mathfrak{P}_i$ denotes the set of all the complete pre-orders $\mathcal{P}_i \subseteq X_i \times X_i$. Let $\mathcal{W}_i \in \mathfrak{P}_i$ denote the welfare relation of the player i, i.e. the ranking of the set X_i that corresponds to her total subjective comparative evaluation. We define a utility function $u_i : X \mapsto \mathbb{R}$ as follows[2]:

$$\forall x, x' \in X, \qquad x \mathcal{P}_i x' \quad \Longleftrightarrow \quad u_i(x|\mathcal{P}_i) \geq u_i(x'|\mathcal{P}_i). \tag{1}$$

In particular, we can define a welfare function $w_i : X \mapsto \mathbb{R}$:

$$w_i(x) = u_i(x|\mathcal{W}_i). \tag{2}$$

We assume that the welfare relations $\mathcal{W}_i$, $\forall i \in N$, as well as the space of strategies X are common knowledge. Γ denotes the game in normal form $\Gamma = \langle N; \{X_i\}_{i \in N}; \{w_i\}_{i \in N} \rangle$.

[2]We consider here only an ordinal notion of utility: the utility function u_i is therefore defined up to a monotonic transformation.

We now assume that — before playing the game Γ — the individuals choose their preference relation $\mathcal{P}_i$. This means that for any game Γ in normal form, we can associate a two-stage game G^*, in which the players choose in the first stage game G_0 a relation $\mathcal{P}_i \in \mathfrak{P}_i$, before playing in a second stage the initial game Γ with the utility function $u_i(x|\mathcal{P}_i)$. G denotes the second stage game $G = \langle N; \{X_i\}_{i\in N}; \{u_i(x|\mathcal{P}_i)\}_{i\in N}\rangle$.

2.2 Second stage game

In the second stage G of the two-stage game G^*, each player has chosen a preference relation $\mathcal{P}_i \in \mathfrak{P}_i$, and wants to satisfy those preferences. Each player therefore seeks to maximize one's utility, knowing that the others players want also to maximize their utility. We define a second stage equilibrium as a strategy profile for which no player can unilaterally increase her own utility by deviating from the equilibrium, and name it a Nash equilibrium for $\mathcal{P}$:

Definition 2.1. *Consider the game G defined in section 2.1. Let $\mathcal{P} \in \mathfrak{P} = \{\mathfrak{P}_i\}_{i\in N}$ be a set of preference relations. A strategy profile $\bar{x} \in X$ is a Nash equilibrium for $\mathcal{P}$ if and only if, $\forall i \in N$:*

$$u_i(\bar{x}|\mathcal{P}_i) \geq u_i(x_i, \bar{x}_{-i}|\mathcal{P}_i), \qquad \forall x_i \in X_i. \tag{3}$$

We can notice that a Nash equilibrium for $\mathcal{W}$, with $\mathcal{W} = \{\mathcal{W}_i\}_{i\in N}$ the set of welfare relations, is simply a Nash equilibrium. Since the set of preferences $\mathfrak{P}_i$ is not restricted, we can have either one, several or no second stage equilibrium. It is therefore possible to associate to each $\mathcal{P} \in \mathfrak{P}$ the set of Nash equilibria for $\mathcal{P}$, and therefore a set of vectors of welfare for each Nash equilibrium for $\mathcal{P}$.
Let $\mathcal{X}(\mathcal{P})$ denote the set of Nash equilibria for $\mathcal{P}$. Let $v(\mathcal{P}) : \mathfrak{P} \mapsto 2^{\mathbb{R}^n}$, with $2^{\mathbb{R}^n}$ the power set of $\mathbb{R}^n$, denote a relation that associates to any set of preferences $\mathcal{P} \in \mathfrak{P}$ the set of vectors of outcomes of the Nash equilibria for $\mathcal{P}$:

$$v(\mathcal{P}) = \{(w_i(x))_{i\in N} \mid x \in \mathcal{X}(\mathcal{P})\}. \tag{4}$$

2.3 First stage game

In the first stage game G_0, each player chooses her own preference relation $\mathcal{P}_i \in \mathfrak{P}_i$ in order to maximize *in fine* her welfare function $w_i(x)$. It has to be noticed that, since there exists a difference between the function that determines the choice in the second stage game — the utility function — and the one that determines the outcome that will be used in the first stage game — the welfare function —, we cannot represent the two stage game as a one-stage game in normal form[3]. We will therefore define an equilibrium notion based on the same idea than a subgame perfect Nash equilibrium, but extended to Nash equilibria for $\mathcal{P}$. This requires to solve the issue of the possible non-uniqueness of second stage equilibrium.

[3]We can for instance notice that we cannot simply refer to the notion of subgame perfect Nash equilibrium, since the second stage equilibrium is not a standard Nash equilibrium but a Nash equilibrium for $\mathcal{P}$.

The main idea we will use in order to tackle the possibility of none or several second stage equilibria will be to define the beliefs about the rule of selection of the second stage equilibrium. If there exists several equilibria in the second stage game, then each player can have a belief about the equilibrium that will effectively occur; if there exists no second stage equilibrium, then each player can have a belief about the outcome she will effectively get. Since the non-existence of a second stage equilibrium[4] can be quite disturbing for a rational player — in the sense that she does not know what she will play — we will assume that, when playing the first stage game, the players expect their worst possible outcome in the second stage game if there is no second stage equilibrium for a given $\mathcal{P} \in \mathfrak{P}$. If there exists some strategy profiles in the first stage game such that there exists several second stage equilibria, then we will represent as many first stage games as there exists possible beliefs about the set of second stage equilibria that could effectively occur. We will therefore study a multiplicity of well-defined first stage games — each one corresponding to a possible first stage game if the rule of selection of the second stage equilibrium was known — and not study a unique first stage game which could present none or several second stage equilibria.

We define now a set of first stage games $\{G_{0,k}\}_{k\in K}$ (it is possible to define this set of first stage games for all games Γ, and therefore to treat any game Γ as a two-stage game G^*):

$$G_{0,k} = \langle N, \mathfrak{P}, \{v_{i,k}\}_{i\in N}\rangle, \tag{5}$$

with $\forall i \in N$, $v_{i,k} : \mathfrak{P} \mapsto \mathbb{R}$ an indirect welfare function such that:

$$\{v_{i,k}(\mathcal{P})\}_{i\in N} \in v(\mathcal{P}), \qquad \text{if} \qquad \mathcal{X}(\mathcal{P}) \neq \emptyset, \tag{6}$$

$$v_{i,k}(\mathcal{P}) = \min_{x\in X} w_i(x), \qquad \text{if} \qquad \mathcal{X}(\mathcal{P}) = \emptyset, \tag{7}$$

$$\bigcup_{k\in K} \{v_{i,k}(\mathcal{P})\}_{i\in N} = v(\mathcal{P}), \qquad \forall \mathcal{P} \in \mathfrak{P}. \tag{8}$$

The game $G_{0,k}$ is therefore a possible first stage game of the game G^*, for which a unique vector of welfare $\{w_i(x)\}_{i\in N}$ has been selected for each strategy profile $\mathcal{P} \in \mathfrak{P}$: either the vector of a Nash equilibrium for $\mathcal{P}$ (condition (6)), or the worst possible outcome for every player if there is no second stage equilibrium (condition (7)). The last condition (8) means that the set of possible first stage games $\{G_{0,k}\}_{k\in K}$ covers every possible combinations of second stage equilibria, i.e. that whatever the rules of selection of the second stage equilibrium are, there exists a first stage game $G_{0,k}$ that represents those rules. We are now able to properly define the two-stage game equilibrium:

Definition 2.2. *Let Γ be a game in normal form as defined in section 2.1, and $\{G_{0,k}\}_{k\in K}$ its set of first stage games. A strategy profile $(\bar{x}; \bar{\mathcal{P}}) \in X \times \mathfrak{P}$ is a subgame perfect $\mathcal{P}$-equilibrium of the two stage game G^* if and only if:*

- *$\bar{x} \in X$ is a Nash equilibrium for $\bar{\mathcal{P}}$ of the game G,*

[4] We consider here a case where there does not exist any equilibrium, even in mixed strategies.

- *there exists a first stage game $G_{0,k}$ such that:*

  (i) $v_k(\bar{\mathcal{P}}) = \{u_i(\bar{x}|\bar{\mathcal{P}})\}_{i\in N}$,

  (ii) $\bar{\mathcal{P}} \in \mathfrak{P}$ *is a Nash equilibrium of the game* $G_{0,k}$.

A subgame perfect $\mathcal{P}$-equilibrium is therefore a strategy profile $\{\bar{x};\bar{\mathcal{P}}\} \in X \times \mathfrak{P}$ of the two-stage game G^* such that (i) the players maximize simultaneously their utility functions $u_i(x|\bar{\mathcal{P}})$, and (ii) there exists a first game $G_{0,k}$ (in which $\bar{x} \in X$ occurs among the possible Nash equilibria for $\bar{\mathcal{P}}$) such that the strategy profile $\bar{\mathcal{P}} \in \mathfrak{P}$ is a Nash equilibrium, i.e. when, $\forall i \in N$:

$$v_{i,k}(\bar{\mathcal{P}}) \geq v_{i,k}(\mathcal{P}_i;\bar{\mathcal{P}}_{-i}), \qquad \forall \mathcal{P}_i \in \mathfrak{P}_i. \tag{9}$$

3 Stackelberg leadership and choice of one's preferences

We show the two following results:

Proposition 1. *Let Γ be a game in normal form as defined in section 2.1 and G^* its associated two-stages game. $(\bar{x};\mathcal{W}) \in X \times \mathfrak{P}$ is a subgame perfect $\mathcal{P}$-equilibrium of G^* if and only if the Nash equilibrium $\bar{x} \in X$ is also a Stackelberg equilibrium when i is the leader, $\forall i \in N$.*

Proposition 2. *Let Γ be a game in normal form as defined in section 2.1, G^* its associated two-stages game, and $\bar{x} \in X$ a Nash equilibrium of Γ. There always exists a set of preference relations $\bar{\mathcal{P}} \in \mathfrak{P}$ such that $\{\bar{x};\bar{\mathcal{P}}\} \in X \times \mathfrak{P}$ is a subgame perfect $\mathcal{P}$-equilibrium.*

The first result implies that, as soon as one player can improve her welfare by becoming a Stackelberg leader, then there exists at least one player who will choose a preference relation $\bar{\mathcal{P}}_i$ different from her welfare relation $\mathcal{W}_i$, i.e. who will make a strategic commitment. The second result means that, although the players will generally decide to maximize a different function than their welfare function, they can always choose a set of preference relations $\bar{\mathcal{P}}$ such that they play *in fine* a Nash equilibrium: this means in particular that the existence of a Nash equilibrium in Γ ensures the existence of a subgame perfect $\mathcal{P}$-equilibrium in the two-stage game G^*.

3.1 Stackelberg function

In order to show the proposition 1, we need to define a *Stackelberg function*. The Stackelberg function of player i is her welfare function if she were a Stackelberg leader with $(n-1)$ followers, i.e. taking into account the best reply functions[5] of the other players:

[5]It has to be noticed that our definition of a "best reply function" is more restrictive than the usual one, since it already integrates the best reply functions of the other players but one.

Definition 3.1. *The function $f_j : X_i \mapsto X_j$ is the best reply function of player j for $\mathcal{P} \in \mathfrak{P}$ if and only if, $\forall x_i \in X_i$:*

$$u_j(f_1(x_i); \ldots; f_j(x_i); \ldots; f_n(x_i)|\mathcal{P}_j) \geq u_j(f_1(x_i); \ldots; x_j; \ldots; f_n(x_i)|\mathcal{P}_j), \qquad \forall x_j \in X_j, \quad (10)$$

with $f_k : X_i \mapsto X_k$ the best reply function of player k for $\mathcal{P} \in \mathfrak{P}$, $\forall k \neq i, j$.

As soon as there exists a Nash equilibrium for $\mathcal{P}$, we know that the best reply functions f_j for $\mathcal{P}$ are defined on a non empty subset of X_i. Indeed, if it was not the case, then a second stage equilibrium could not exist, since $\bar{x} \in X$ is a Nash equilibrium for $\mathcal{P}$ if and only if:

$$f_j(\bar{x}_i) = \bar{x}_j, \qquad \forall i, j \in N. \quad (11)$$

We can now define the Stackelberg function:

Definition 3.2. *Consider the game G as defined in section 2.1. Let $f_j(x_i)$ denotes the best reply function of player j for $\mathcal{P} \in \mathfrak{P}$. The function $\Psi_i : X_i \mapsto \mathbb{R}$ is the Stackelberg function of player i if and only if:*

$$\Psi_i(x_i|\mathcal{P}) = w_i(f_1(x_i); \ldots; f_n(x_i)). \quad (12)$$

A Stackelberg function is therefore the welfare function of player i which integrates the best reply function of the players $j \neq i$ as a function of the strategy x_i.

Consider now a first stage game $G_{0,k}$. We look for the conditions under which $\mathcal{W}$ can be a first stage equilibrium, i.e. the players decide to directly maximize their welfare function, and not another utility function. A set of preference relations $\bar{\mathcal{P}} \in \mathfrak{P}$ is a Nash equilibrium of $G_{0,k}$ if and only if, $\forall i \in N$, $\forall \mathcal{P}_i \in \mathfrak{P}_i$:

$$v_{i,k}(\bar{\mathcal{P}}) \geq v_{i,k}(\mathcal{P}_i; \bar{\mathcal{P}}_{-i}), \quad (13)$$

$$\iff \quad \exists \bar{x} \in \mathcal{X}(\bar{\mathcal{P}}),\ \exists \tilde{x} \in \mathcal{X}(\mathcal{P}_i; \bar{\mathcal{P}}_i),\ w_i(\bar{x}) \geq w_i(\tilde{x}). \quad (14)$$

Since $\bar{x}(\mathcal{P})$ is a Nash equilibrium for $\mathcal{P}$, we have $\forall i \in N$:

$$f_j(\bar{x}_i) = \bar{x}_j, \qquad \forall j \in N. \quad (15)$$

We can therefore rewrite the condition (14) as follows, $\forall i \in N$, $\forall \mathcal{P}_i \in \mathfrak{P}_i$:

$$w_i(f_1(\bar{x}_i);\ldots;f_n(\bar{x}_i)|\bar{x}\in\mathcal{X}(\bar{\mathcal{P}}))\geq w_i(f_1(\tilde{x}_i);\ldots;f_n(\tilde{x}_i)|\tilde{x}\in\mathcal{X}(\mathcal{P}_i;\bar{\mathcal{P}}_i)), \tag{16}$$

$$\iff \quad \Psi_i(\bar{x}_i|\bar{x}\in\mathcal{X}(\bar{\mathcal{P}}))\geq\Psi_i(\tilde{x}_i|\tilde{x}\in\mathcal{X}(\mathcal{P}_i;\bar{\mathcal{P}}_i)). \tag{17}$$

The condition (17) means that choosing one's preference relation $\mathcal{P}_i$ implies choosing the strategy profile $\bar{x}_i \in X_i$ that maximizes the Stackelberg function $\Psi_i(x_i)$, knowing the preference relations of the other players.
A direct corollary of this result is that, if a player can improve her welfare by becoming a Stackelberg leader — knowing that the other players are maximizing their welfare functions — then $\mathcal{W}$ cannot be a first stage equilibrium, and therefore at least one player will not maximize her own welfare in the second stage game G. Furthermore, if there exists a Nash equilibrium which is also a Stackelberg equilibrium when i is the leader, $\forall i \in N$, then no player have an interest in changing unilaterally her preference relation, and the players can decide to play the second stage game according to their welfare relation $\mathcal{W}_i$. There can however exist in this configuration other set of preference relations $\bar{\mathcal{P}}$ which are first stage equilibrium, and therefore subgame perfect $\mathcal{P}$-equilibria which are not Nash equilibria (this is the case of the prisoner's dilemma we will present in section 3.3). We have therefore shown the proposition 1.

3.2 Nash equilibrium and subgame perfect $\mathcal{P}$-equilibrium

Although the players generally decide to maximize a different function than their own welfare function, we show in this section that there always exists a first stage equilibrium $\bar{\mathcal{P}} \in \mathfrak{P}$ such that the players play a Nash equilibrium in the second stage game G: if there exists a Nash equilibrium $\bar{x} \in X$ for the game Γ, then there exists $\bar{\mathcal{P}} \in \mathfrak{P}$ such that $(\bar{x};\bar{\mathcal{P}}) \in X \times \mathfrak{P}$ is a subgame perfect $\mathcal{P}$-equilibrium for the game G^*.
Let $\bar{x} \in X$ be a Nash equilibrium of the game Γ. For all players but i, consider the following preference relations $\mathcal{P}_j$, $\forall j \in -i$:

$$(\bar{x}_j, x_{-j})\mathcal{P}_j(x_j, x_{-j}) \qquad \forall x_j \in X_j,\ x_{-j} \in X_{-j}. \tag{18}$$

This means that, whatever the strategies played by the other players are, player j will always rank $\bar{x}_j$ as her preferred strategy. In this situation, since having a first mover advantage cannot help player i to get a higher level of welfare (the other players will not change their strategy), she chooses a preference relation $\mathcal{P}_i \in \mathfrak{P}_i$ such that the second stage equilibrium maximizes her welfare function, i.e. such that the second stage equilibrium is the Nash equilibrium $\bar{x}$. In particular, if she chooses $\bar{\mathcal{P}}_i$ such that:

$$(\bar{x}_i, x_{-i})\mathcal{P}_i(x_i, x_{-i}) \qquad \forall x_i \in X_i,\ x_{-i} \in X_{-i}, \tag{19}$$

then we can see that no other player will have an interest in changing unilaterally her preference relation. This means that, if there exists a Nash equilibrium $\bar{x} \in X$ in the game G, then $(\bar{x};\bar{\mathcal{P}}) \in X \times \mathfrak{P}$ is a subgame perfect $\mathcal{P}$-equilibrium if:

$$(\bar{x}_i, x_{-i})\mathcal{P}_i(x_i, x_{-i}) \qquad \forall x_i \in X_i,\ x_{-i} \in X_{-i},\ \forall i \in N. \tag{20}$$

We have therefore shown the proposition 2.

3.3 Illustrations

We now consider the case of several standard games in order to highlight the relevance of our framework. We firstly present an analysis of symmetric 2×2 games, before providing a more complete analysis of games with continuous strategy sets in the next section.

Consider a symmetric 2×2 game:

	A_2	B_2
A_1	(R;R)	(S;T)
B_1	(T;S)	(P;P)

For convenience, we will consider that the welfare relation is an ordinal one (we cannot therefore aggregate the welfare of the players in a single function), and that mixed strategies are not allowed. In this situation, for each player i, there exists 4 different preference relations in $\mathfrak{P}_i$, i.e.:

- $\mathcal{P}_{i,AA} = \{(A_i; A_{-i}); (A_i; B_{-i})\}$: player i always prefers the strategy A_i;
- $\mathcal{P}_{i,BB} = \{(B_i; A_{-i}); (B_i; B_{-i})\}$: player i always prefers the strategy B_i;
- $\mathcal{P}_{i,AB} = \{(A_i; A_{-i}); (B_i; B_{-i})\}$: player i prefers the strategy A_i if and only if player $-i$ plays A_{-i};
- $\mathcal{P}_{i,BA} = \{(B_i; A_{-i}); (A_i; B_{-i})\}$: player i prefers the strategy A_i if and only if player $-i$ plays B_{-i}.

We can therefore notice that, as long as one of the player chooses either $\mathcal{P}_{i,AA}$ or $\mathcal{P}_{i,BB}$, there necessarily exists a unique second stage equilibrium. If both players choose either $\mathcal{P}_{i,AB}$ or $\mathcal{P}_{i,BA}$, then there are two second stage equilibria. And if one player chooses $\mathcal{P}_{i,AB}$ and the other $\mathcal{P}_{i,BA}$, then there is no second stage equilibrium. In those latter cases, both players therefore anticipate their worst payoff $M = \min\{R; T; P; S\}$.
For any symmetric 2×2 game, we can therefore define 4 possible first stage games, according to the second stage equilibrium selected when both players want either to play the same strategy or not:

	$\mathcal{P}_{2,AA}$	$\mathcal{P}_{2,BB}$	$\mathcal{P}_{2,AB}$	$\mathcal{P}_{2,BA}$
$\mathcal{P}_{1,AA}$	(R;R)	(S;T)	(R;R)	(S;T)
$\mathcal{P}_{1,BB}$	(T;S)	(P;P)	(P;P)	(T;S)
$\mathcal{P}_{1,AB}$	(R;R)	(P;P)	(P;P)	(M;M)
$\mathcal{P}_{1,BA}$	(T;S)	(S;T)	(M;M)	(T;S)

	$\mathcal{P}_{2,AA}$	$\mathcal{P}_{2,BB}$	$\mathcal{P}_{2,AB}$	$\mathcal{P}_{2,BA}$
$\mathcal{P}_{1,AA}$	(R;R)	(S;T)	(R;R)	(S;T)
$\mathcal{P}_{1,BB}$	(T;S)	(P;P)	(P;P)	(T;S)
$\mathcal{P}_{1,AB}$	(R;R)	(P;P)	(P;P)	(M;M)
$\mathcal{P}_{1,BA}$	(T;S)	(S;T)	(M;M)	(S;T)

	$\mathcal{P}_{2,AA}$	$\mathcal{P}_{2,BB}$	$\mathcal{P}_{2,AB}$	$\mathcal{P}_{2,BA}$
$\mathcal{P}_{1,AA}$	(R;R)	(S;T)	(R;R)	(S;T)
$\mathcal{P}_{1,BB}$	(T;S)	(P;P)	(P;P)	(T;S)
$\mathcal{P}_{1,AB}$	(R;R)	(P;P)	(R;R)	(M;M)
$\mathcal{P}_{1,BA}$	(T;S)	(S;T)	(M;M)	(T;S)

	$\mathcal{P}_{2,AA}$	$\mathcal{P}_{2,BB}$	$\mathcal{P}_{2,AB}$	$\mathcal{P}_{2,BA}$
$\mathcal{P}_{1,AA}$	(R;R)	(S;T)	(R;R)	(S;T)
$\mathcal{P}_{1,BB}$	(T;S)	(P;P)	(P;P)	(T;S)
$\mathcal{P}_{1,AB}$	(R;R)	(P;P)	(R;R)	(M;M)
$\mathcal{P}_{1,BA}$	(T;S)	(S;T)	(M;M)	(S;T)

We now identify the subgame perfect $\mathcal{P}$-equilibria of cooperation and coordination games:

Prisoner's dilemma: $T > R > P > S$; for both players, we have $\mathcal{W}_i = \mathcal{P}_{i,BB}$. The strategy profile $(\mathcal{P}_{1,BB}; \mathcal{P}_{2,BB})$ is always a Nash equilibrium in the first stage game; $(\mathcal{P}_{1,AB}; \mathcal{P}_{2,AB})$ is also a Nash equilibrium in the last two games (i.e. when the second stage equilibrium is joint cooperation). It means that whatever the beliefs of the players about the possible second stage equilibria are, joint defection can be played in the second stage game; and that since there exists a rule of selection[6] of the second stage equilibrium such that conditional cooperation is a Nash equilibrium in the first stage game, then joint cooperation can also be played in the second stage game.
This theoretical prediction seems to be quite more realistic than the systematic defection predicted by the standard theory: our framework predicts that, in order to maximize one's own welfare, we can either unilaterally defect in every circumstances, or choose to become a conditional cooperator, i.e. to cooperate if and only if the other player cooperates too. An interesting point here is that the mechanism of reciprocity — which is quite natural in a repeated prisoner's dilemma — seems also to be effective even without the repetition of the game. There exists therefore two subgame perfect $\mathcal{P}$-equilibria: $((B_1; B_2); (\mathcal{P}_{1,BB}; \mathcal{P}_{2,BB}))$ and $((A_1; A_2); (\mathcal{P}_{1,AB}; \mathcal{P}_{2,AB}))$.

Hi-Lo: $R > P > T = S$; for both players, we have $\mathcal{W}_i = \mathcal{P}_{i,AB}$. The strategy profiles $(\mathcal{P}_{1,AA}; \mathcal{P}_{2,AA})$ and $(\mathcal{P}_{1,BB}; \mathcal{P}_{2,BB})$ are Nash equilibria in each first stage games, and $(\mathcal{P}_{1,AB}; \mathcal{P}_{2,AB})$ is also a Nash equilibrium in the last two games (i.e. when the second stage equilibrium is joint cooperation). It implies that the two Nash equilibria in pure strategies can be played in the second stage game (this confirms our proposition 2). However, we can notice that the preference relation $\mathcal{P}_{i,BB}$ is weakly dominated by $\mathcal{P}_{i,AB}$, $\forall i \in N$: this means that if we assume that the players do not play weakly dominated strategies, then they will never commit themselves to play B_i in any circumstances, and will therefore be able to select the pareto dominant Nash equilibrium. Our framework of the choice of one's preferences can therefore help to solve some coordination issues, since it appears that we do not need strong assumptions about individuals' rationality (such as a rule of selection of equilibrium) in order to explain the selection of the Pareto-dominant equilibrium: we only need to assume that rational players will not play weakly dominated strategies in the first stage game. There exists therefore three subgame perfect $\mathcal{P}$-equilibria: $((A_1; A_2); (\mathcal{P}_{1,AA}; \mathcal{P}_{2,BB}))$, $((B_1; B_2); (\mathcal{P}_{1,BB}; \mathcal{P}_{2,BB}))$, and $((A_1; A_2); (\mathcal{P}_{1,AB}; \mathcal{P}_{2,AB}))$.

4 Continuous strategy sets

In this section, we consider games with continuous strategy sets and investigate the conditions under which the players choose cooperative preferences. We firstly establish a link between supermodularity (respectively submodularity) and cooperation (competition) in the second stage game of two players games, and illustrate our results with a two-players collective action game.

4.1 Supermodularity and cooperation

We introduce the following notations and definitions:

[6] It is here sufficient that both players believe that two conditional cooperators — who know that they are both conditional cooperators — will always cooperate.

- The partial derivatives of a function $h_i : Y \mapsto Z$ are noted (the index i designates to whom player is associated the function h, and the index j and k the successive derivatives of the function h_i according to the strategy of the players j and k):

$$h_i^{jk}(y) = \frac{\partial^2 h_i}{\partial y_j \partial y_k}(y_1; \ldots; y_n). \tag{21}$$

- For a given $n \times n$ matrix $S \in \mathbb{R}^{n \times n}$, S_{ij} denotes a $(n-1) \times (n-1)$ matrix that results from deleting row i and column j of S.
- For a given $n \times n$ matrix $S \in \mathbb{R}^{n \times n}$, $C_{ij}^S = (-1)^{i+j} |S_{ij}|$ denotes the $(i; j)$ cofactor of S.
- The game Γ defined in section 2.1 is supermodular (respectively submodular) if and only if, $\forall i \in N$, the welfare functions w_i are of class C^2 and:

$$w_i^{ij}(x) \geq (\leq) 0 \qquad \forall x \in X, \quad \forall j \neq i. \tag{22}$$

For convenience, we will assume in this section that the set of possible preference relations in the first stage game is limited to the set of preference relations whose associated utility functions can be described as linear combinations of the welfare functions w_i. The possible utility functions of an individual are therefore of the following form, $\forall i \in N$:

$$u_i(x|S_i) = \sum_{j \in N} \sigma_{ij} w_j(x), \tag{23}$$

with $S_i = \{\sigma_{ij}\}_{i \in N}$ the set of weighting parameters of player i, i.e. the weight she gives to the other players in her preferences. In this restricted framework, each player chooses in the first stage game a vector of parameters $S_i \in \mathbb{R}^n$, and then maximizes in the second stage game a weighted sum of the welfare functions of all players. We investigate in this section the conditions under which the players choose positive parameters in the first stage game in order to maximize their welfare function, i.e. when the possiblity of choosing one's preferences generates cooperation.

For matters of simplicity, we make the following assumptions about the game G, $\forall i \in N$:

- the welfare function $w_i(x)$ is of class C^3;
- the utility function $u_i(x|S_i)$ is quasiconcave in x_i, $\forall S_i \in \mathbb{R}^n$;
- the Stackelberg function $\Psi_i(x|S)$ is quasiconcave in x_i, $\forall S = \{S_i\}_{i \in N} \in \mathbb{R}^{n \times n}$.

A direct implication of these assumptions is that, $\forall S \in \mathbb{R}^{n \times n}$, there exists at least one second stage equilibrium $\bar{x}(S) \in X$. We also assume that the second stage equilibrium $\bar{x} : \mathbb{R}^{n \times n} \mapsto X$ is of class C^2. We can check that, for $n = 2$, those conditions are verified when the welfare functions w_i are quadratic and concave.

We now assume that at each second stage game equilibrium $\bar{x}(S) \in X$, $\forall i \neq j$, the first order derivatives $w_i^j(\bar{x})$ and $w_j^i(\bar{x})$ (if non null) have the same sign (as in a Cournot oligopoly for instance). We show the following results:

Proposition 3. *Let Γ be a two-players game in normal form as defined in section 4.1. If the game Γ is submodular, then at least one player will choose a negative weight for the welfare function of the other player in her utility function at the first stage equilibrium of the two stages game G^*.*

Proposition 4. *Let Γ be a two-players game in normal form as defined in section 4.1. If the game Γ is supermodular, then both players will choose a positive weight for the welfare function of the other player in their utility functions at the first stage equilibrium of the two stages game G^*.*

Proposition 3 means that, if the players are able to make strategic commitments, then the submodularity property of the game Γ will create competitive behaviours in the second stage game (at the first stage equilibrium), although it is not certain that both players will choose a spiteful motivation. It means that players will maximize their own welfare as well as the difference between their welfare and the welfare of the other player. Conversely, proposition 4 means that the supermodularity property of the game Γ will generate cooperation in the second stage game (at the first stage equilibrium).

Proof. Propositions 3 and 4 require quite similar proofs, and we will therefore use the same reasoning for both. Our proof consists of three steps: (1) we firstly explicit the first order conditions of the first and second stage game equilibria; (2) we then determine the best reply function as defined in section 3.1; and (3) we focus on the case of two players games and show that the supermodularity property of the game G is preserved for the second stage game at the first stage equilibrium.

Step 1: for any $S \in \mathbb{R}^{n\times n}$, the utility function $u_i(x|S_i)$ is continuous in x and quasiconcave in x_i, $\forall i \in N$. There therefore exists a second stage equilibrium $\bar{x} \in X$ that verifies, $\forall i \in N$:

$$u_i^i(\bar{x}|S_i) = 0, \tag{24}$$

$$\sum_{j\in N} \sigma_{ij} w_j^i(\bar{x}) = 0. \tag{25}$$

The second stage equilibrium $\bar{x} : \mathbb{R}^{n\times n} \mapsto X$ being of class C^2 as well as the welfare function $w_i(x)$, the indirect welfare function $v_i(S)$ is also of class C^2. As shown in the previous section, maximizing the indirect welfare function in the first stage game is equivalent to maximizing the Stackelberg function $\Psi_i(\bar{x}_i(S))$ evaluated at the second stage equilibrium. We have at the first stage equilibrium:

$$\frac{\partial v_i}{\partial \sigma_{ij}}(\bar{x}_i(S)) = \Psi_i^i(\bar{x}_i(S)) \frac{\partial \bar{x}_i}{\partial \sigma_{ij}}(\sigma_{i1}; \ldots; \sigma_{in}) \;=\; 0, \qquad \forall j \in N. \tag{26}$$

We must therefore verify at the first stage equilibrium:

$$\text{either} \qquad \frac{\partial \bar{x}_i}{\partial \sigma_{ij}}(\sigma_{i1}; \dots; \sigma_{in}) = 0, \qquad \forall j \in N, \tag{27}$$

$$\text{or} \qquad \Psi_i^i(\bar{x}_i(S)) = 0, \qquad \forall \bar{x}_i \in X_i. \tag{28}$$

We suggest now highlighting that the condition (27) is quite uncommon, and that we can reasonably assume that the condition (26) is verified if and only if (28) is verified. We need therefore to characterize the second stage equilibrium $\bar{x} \in X$. Since we assumed the existence of a twice differentiable solution for (25), we identify the best reply of player i, $\forall i \neq j$, when a player j unilaterally changes her parameters $\sigma_{j.}$. We consider here the differential of u_i^i, and look for the reactions $\mathrm{d}x_i$ that verify $\mathrm{d}u_i^i(\bar{x}) = 0$, $\forall i \in N$. We have the following relations:

$$\mathrm{d}u_i^i(\bar{x}) = 0, \qquad \forall i \in N, \tag{29}$$

$$\sum_{j \in N} \left[u_i^{ij}(\bar{x}) \ \mathrm{d}x_j + w_j^i(\bar{x}) \ \mathrm{d}\sigma_{ij} \right] = 0, \qquad \forall i \in N. \tag{30}$$

We solve this system of linear equations in $\mathrm{d}x_i$:

$$\begin{pmatrix} u_1^{11}(\bar{x}) & \dots & u_1^{1n}(\bar{x}) \\ \dots & & \dots \\ u_n^{n1}(\bar{x}) & \dots & u_n^{nn}(\bar{x}) \end{pmatrix} \begin{pmatrix} \mathrm{d}x_1 \\ \dots \\ \mathrm{d}x_n \end{pmatrix} + \begin{pmatrix} \sum_{j \in N} w_j^1(\bar{x}) \, \mathrm{d}\sigma_{1j} \\ \dots \\ \sum_{j \in N} w_j^n(\bar{x}) \, \mathrm{d}\sigma_{nj} \end{pmatrix} = 0, \tag{31}$$

$$J \, \mathrm{d}x + \mathrm{d}A = 0, \tag{32}$$

with $\mathrm{d}x = {}^t\{\mathrm{d}x_i\}_{i \in N}$ the column vector of strategies' variations; $\mathrm{d}A = {}^t\{\mathrm{d}A_i\}_{i \in N}$; and J the $n \times n$ Jacobian matrix (evaluated at the second stage equilibrium) of the function $\partial u : X \mapsto \mathbb{R}^n$ that associates to any strategy profile $x \in X$ the marginal utility function of each player. We make the assumption that J and its minors J_{ii} are generically non singular $\forall S \in \mathbb{R}^{n \times n}$. The system (32) is therefore a Cramer system and the unique solution is given by:

$$\mathrm{d}x_i = \frac{\left| J^i \right|}{|J|} \qquad \forall i \in N, \tag{33}$$

with J^i a $n \times n$ matrix identical to J, except for the i^{th} column which is replaced by $-\mathrm{d}A$. We deduce the following relations:

$$\mathrm{d}x_i = - \frac{\sum_{k \in N} C_{ki}^J \, \mathrm{d}A_k}{|J|}, \tag{34}$$

$$\Longrightarrow \qquad \frac{\partial \bar{x}_i}{\partial \sigma_{ik}}(S) = - \, w_k^i \frac{C_{ii}^J}{|J|}(\bar{x}(S)) \qquad \forall S \in \mathbb{R}^{n \times n}. \tag{35}$$

We can therefore see the condition (27) implies:

$$w_k^i(\bar{x}) = 0 \qquad \forall k \in N. \tag{36}$$

This last condition means that the strategy profile that maximizes the utility function of player i also maximizes her own welfare function w_i as well as the welfare of all the other players $j \neq i$ (or minimizes, according to the sign of the second order derivative). Since the condition (27) is quite uncommon, we make the additional assumption that:

$$\frac{\partial \bar{x}_i}{\partial \sigma_{ik}} = 0, \quad \forall k \in N \qquad \Longrightarrow \qquad \Psi_i^i(\bar{x}_i(S)) = 0. \tag{37}$$

We can therefore rewrite the first order condition of the first stage equilibrium (26):

$$\sum_{j \in N} w_i^j(\bar{x}_i(S)) f_j^i(\bar{x}_i(S)) = 0. \tag{38}$$

For $n = 2$, the first order conditions of the first and second stage equilibrium are therefore:

$$\begin{cases} \sigma_{ii} w_i^i(\bar{x}) + \sigma_{ij} w_j^i(\bar{x}) = 0, & \forall i, j \in N,\ i \neq j, \\ w_i^i(\bar{x}(S)) + f_j^i w_i^j(\bar{x}(S)) = 0, & \forall i, j \in N,\ i \neq j. \end{cases} \tag{39}$$

If we assume that each player gives a non null weight to her own welfare function (i.e. $\sigma_{ii} \neq 0$), we obtain:

$$\sigma_i = f_j^i \frac{w_j^i}{w_i^j}(\bar{x}(S)), \tag{40}$$

with $\sigma_i = \frac{\sigma_{ij}}{\sigma_{ii}}$. Since we assumed that w_j^i and w_i^j have the same sign at the second stage equilibrium, it means that σ_i has the same sign than $f_j^i(\bar{x})$. Our purpose is to show under which conditions the individuals will cooperate or not in the second stage game, and therefore to determine the sign of σ_i at the first stage equilibrium: we now must determine the best reply function $f_j(x_i)$.

Step 2: we now determine $f_j(x_i)$, the best reply functions as defined in section 3.1. Consider that all players but i are maximizing their utility functions, i.e. that they play their best reply strategy for x_i; if player i changes her strategy such that $\mathrm{d}x_i \neq 0$, then, we must verify, $\forall j \neq i$ (the different functions are evaluated in $(f_1(x_i); \ldots; f_n(x_i))$, i.e. when all players but i maximize their utility functions):

$$\mathrm{d}u_j^j(x) = 0, \tag{41}$$

$$u_j^{ji}\ \mathrm{d}x_i + \sum_{k \neq i} u_j^{jk}\ \mathrm{d}x_k = 0. \tag{42}$$

We can rewrite this system of linear equations with $\mathrm{d}x_{-i} = {}^t\{\mathrm{d}x_k\}_{k \neq i}$, and $B_i = {}^t\{u_k^{ki}\,\mathrm{d}x_i\}_{k \neq i}$:

$$J_{ii}\,\mathrm{d}x_{-i} + B_i = 0. \tag{43}$$

Since we assume that J_{ii} is non singular, the system (43) is a Cramer system and has a unique solution, with J_{ii}^j a $(n-1)\times(n-1)$ matrix identical to J except for the column made of u_k^{kj}, $\forall k \neq i$ which is replaced by $-B_i$, and without row i and column i:

$$\mathrm{d}x_j = \frac{\left|J_{ii}^j\right|}{|J_{ii}|}. \tag{44}$$

We can develop the determinant of J_{ii}^j (we arbitrarily suppose that $i<j$):

$$\left|J_{ii}^j\right| = \begin{vmatrix} u_1^{11} & \dots & u_1^{1,i-1} & u_1^{1,i+1} & \dots & u_1^{1,j-1} & -u_1^{1i}\,\mathrm{d}x_i & u_1^{1,j+1} & \dots & u_1^{1n} \\ \dots & & \dots & \dots & & \dots & \dots & \dots & & \dots \\ u_{i-1}^{i-1,1} & \dots & u_{i-1}^{i-1,i-1} & u_{i-1}^{i-1,i+1} & \dots & u_{i-1}^{i-1,j-1} & -u_{i-1}^{i-1,i}\,\mathrm{d}x_i & u_{i-1}^{i-1,j+1} & \dots & u_{i-1}^{i-1,n} \\ u_1^{i+1,1} & \dots & u_{i+1}^{i+1,i-1} & u_{i+1}^{i+1,i+1} & \dots & u_{i+1}^{i+1,j-1} & -u_{i+1}^{i+1,i}\,\mathrm{d}x_i & u_{i+1}^{i+1,j+1} & \dots & u_{i+1}^{i+1,n} \\ \dots & & \dots & \dots & & \dots & \dots & \dots & & \dots \\ u_n^{n1} & \dots & u_n^{n,i-1} & u_n^{n,i+1} & \dots & u_n^{n,j-1} & -u_n^{ni}\,\mathrm{d}x_i & u_n^{n,j+1} & \dots & u_n^{nn} \end{vmatrix} \tag{45}$$

We can rewrite this determinant as follows:

$$\left|J_{ii}^j\right| = -\,\mathrm{d}x_i \begin{vmatrix} u_1^{11} & \dots & u_1^{1,i-1} & 0 & u_1^{1,i+1} & \dots & u_1^{1,j-1} & u_1^{1,i} & u_k^{k,j+1} & \dots & u_1^{1n} \\ \dots & \dots & \dots & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ u_{i-1}^{i-1,1} & \dots & u_{i-1}^{i-1,i-1} & 0 & u_{i-1}^{i-1,i+1} & \dots & u_{i-1}^{i-1,j-1} & u_{i-1}^{i-1,i} & u_{i-1}^{i-1,j+1} & \dots & u_{i-1}^{i-1,n} \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ u_1^{i+1,1} & \dots & u_{i+1}^{i+1,i-1} & 0 & u_{i+1}^{i+1,i+1} & \dots & u_{i+1}^{i+1,j-1} & u_{i+1}^{i+1,i} & u_{i+1}^{i+1,j+1} & \dots & u_{i+1}^{i+1,n} \\ \dots & \dots & \dots & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ u_n^{n1} & \dots & u_n^{n,i-1} & 0 & u_n^{n,i+1} & \dots & u_n^{n,j-1} & u_n^{ni} & u_n^{n,j+1} & \dots & u_n^{nn} \end{vmatrix} \tag{46}$$

We can then invert the i^{th} with the j^{th} column, and we obtain:

$$\left|J_{ii}^j\right| = (-1)^{i+j}\,|J_{ij}|\,\mathrm{d}x_i. \tag{47}$$

We can now rewrite the relation (44):

$$\mathrm{d}x_j = \frac{C_{ij}^J}{C_{ii}^J}(f_1(x_i);\dots;f_n(x_i))\,\mathrm{d}x_i. \tag{48}$$

This last relation gives us the best reply of player j to a given variation of strategy of player i in order to maximize her utility function when all the other players but i are maximizing their utility functions. We can notice that the primitive of the best reply in terms of variation $\mathrm{d}x_j$ is the best reply function of player j, i.e. the strategy x_j which maximizes the utility function u_j for a given strategy of player i, knowing the best reply of the other players $k \neq i, j$:

$$f_j(x_i) = \int \frac{C_{ij}^J}{C_{ii}^J}(f_1(x_i);\dots;f_n(x_i))\,\mathrm{d}x_i. \tag{49}$$

For two players, we have therefore:

$$f_j^i(x_i) = -\frac{u_j^{ji}}{u_j^{jj}}(x_i; f_j(x_i)). \tag{50}$$

We can now show our proposition 3. Since the condition (40) implies that σ_i has the sign of $f_j^i(\bar{x}_i)$ at the first stage equilibrium, and since we assumed that u_i is always quasiconcave in x_i, $\forall i \in N$, we know that the solution $\bar{\sigma}_i$ has the same sign than $w_j^{ji}(\bar{x})$. Suppose that there exists an equilibrium for which $\bar{\sigma}_i > 0$ $\forall i \in N$. If G is submodular, then we have $w_i^{ij}(\bar{x}) < 0$: it implies that both functions $u_i^{ij}(\bar{x})$ and $u_j^{ji}(\bar{x})$ are negative by construction. This contradicts the positivity of the solution $\bar{\sigma}_i$ assumed above. It means that at least one player put a negative weighting on the welfare function of the other player in the second stage game at the first stage game equilibrium: we have therefore shown our proposition 3. It is however not certain that both players will put a negative weight on the welfare of the other player: it is indeed possible that a player chooses a sufficiently negative weight such that the other decides to unilaterally cooperate: the rationale of this equilibrium is that, since player i maximizes the difference between $w_i(x)$ and $w_j(x)$, player j can benefit from maximizing $w_i(x)$ too (maximizing the difference like player i would indeed lead to a deteriorated situation), and therefore relatively decreases the pressure exerted on her welfare by the other player.

We can now notice that $f_j^i(x_i)$ (and then σ_i) will be positive at the first stage equilibrium if the utility function u_i is supermodular. We therefore study now the conditions under which the supermodularity of the welfare function w_i is preserved at the first stage equilibrium, i.e. under which the utility function u_i is supermodular too.

Step 3: we now show that the supermodularity of the game G is preserved for the utility functions at the first stage equilibrium. Let $\phi_i : \mathbb{R} \mapsto \mathbb{R}$ denotes the best reply function of player i for the first stage game, i.e. the function that maximizes the indirect utility function of player i for a given strategy $\sigma_j \in \mathbb{R}$. Thanks to the equation (49), we know that this best reply function is (with V the Jacobian matrix of the marginal indirect welfare function):

$$\phi_i(\sigma_j) = \int \frac{C_{ji}^V}{C_{jj}^V}(\phi_i(\sigma_j); \sigma_j)\, \mathrm{d}\sigma_j, \tag{51}$$

$$V = \left(\Psi_i^{ii} \frac{\partial x_i}{\partial \sigma_i} \frac{\partial x_i}{\partial \sigma_j} \right)_{i,j \in N}. \tag{52}$$

We have therefore:

$$\phi_i(\sigma_j) = \int \frac{u_i^{ij}}{u_j^{jj}} \frac{w_i^j}{w_j^i}(\phi_i(\sigma_j); \sigma_j)\, \mathrm{d}\sigma_j. \tag{53}$$

We suggest now studying the function $\Phi_i(\sigma_i) = \phi_i \circ \phi_j(\sigma_i)$. We can indeed notice that a first stage equilibrium $(\bar{\sigma}_i)_{i \in N}$ necessarily verifies:

$$\Phi_i(\bar{\sigma}_i) = \bar{\sigma}_i, \qquad \forall i \in N. \tag{54}$$

The supermodularity is preserved for the utility functions if and only if[7]:

$$\bar{\sigma_i} \geq -\frac{w_i^{ij}}{w_j^{ji}}(\bar{x}(\bar{S})) \qquad \forall i,j \in N, j \neq i. \tag{55}$$

We therefore suggest showing that $\Phi_i(\sigma_i)$ is bounded below by $-\frac{w_i^{ij}}{w_j^{ji}}(\bar{x}(\sigma_i;\phi_j(\sigma_i)))$, $\forall i \in N$. In this case, the condition (55) will indeed be verified, and the second stage game with the utility functions $u(x|\bar{S})$ will be supermodular. The supermodularity of the game G would then imply that both players choose positive parameters $\sigma_i \in \mathbb{R}$ in the first stage game and therefore cooperate in the second stage game.
We now determine the minimum of $\Phi_i : \mathbb{R} \mapsto \mathbb{R}$. Since we assume the continuity and quasi-concavity of the Stackelberg function $\forall S \in \mathbb{R}^2$, we know that the best reply functions ϕ_i and therefore Φ_i are continuous $\forall \sigma_i \in \mathbb{R}$. The minimum of Φ_i can therefore be reached only for $\sigma_i \to \pm\infty$ or for σ_i such that $\Phi_i^i(\sigma_i) = 0$. The condition (40) implies that:

$$\lim_{\sigma_i \to \infty} \phi_j(\sigma_i) = 0, \tag{56}$$

$$\implies \lim_{\sigma_i \to \infty} \Phi_i(\sigma_i) = -\frac{w_j^{ji}}{w_j^{jj}} \frac{w_i^j}{w_j^i}(\bar{x}(\sigma_i, 0)). \tag{57}$$

The relation (57) implies that, if G is supermodular, then Φ_i is positive by construction when $\sigma_i \to \pm\infty$.
Consider now the first order derivative of Φ_i:

$$\Phi_i^i(\sigma_i) = \phi_i^j(\phi_j(\sigma_i))\phi_j^i(\sigma_i), \tag{58}$$

$$\Phi_i^i(\sigma_i) = \frac{u_i^{ij}}{u_j^{jj}}(\phi(\sigma_i);\phi_j(\sigma_i))\frac{u_j^{ji}}{u_i^i}(\sigma_i;\phi_j(\sigma_i)) \tag{59}$$

The minimum of Φ_i can therefore be reached for:

- $\tilde{\sigma}_i$ such that $u_j^{ji}(\sigma_i;\phi_j(\sigma_i)) = 0$, i.e. such that $\phi_j(\tilde{\sigma}_i) = -\frac{w_j^{ji}}{w_i^{ij}}(\sigma_i;\phi_j(\sigma_i))$. We would have in this case $\Phi_i(\tilde{\sigma}_i) = 0$;
- $\tilde{\sigma}_i$ such that $u_i^{ij}(\phi(\sigma_i;\phi_j(\sigma_i)) = 0$, i.e. such that $\Phi_i(\tilde{\sigma}_i) = -\frac{w_j^{ji}}{w_i^{ij}}(\sigma_i);\phi_j(\sigma_i))$.

We have shown that the possible minima for $\Phi_i(\sigma_i)$ are all greater than $-\frac{w_j^{ji}}{w_i^{ij}}(\bar{x}(\sigma_i;\phi_j(\sigma_i)))$. This condition is therefore also verified at the first stage equilibrium:

$$\bar{\sigma_i} \geq -\frac{w_j^{ji}}{w_i^{ij}}(\bar{x}(\bar{\sigma_i};\phi_j(\bar{\sigma_i}))). \tag{60}$$

[7]For convenience, we make the assumption that $w_i^{ij}(\bar{x}) \neq 0$, $\forall i,j \in N$.

The supermodularity of the game is therefore preserved at the first stage equilibrium, i.e. when the players decide to maximize the utility functions $u_i(x|\bar{S})$. Both players will therefore choose positive weightings σ_i and cooperate in the second stage game. We have therefore shown our proposition 4.

4.2 Collective action game

We now consider as an illustration a two-players collective action game $\Gamma = \langle \{1;2\}; \mathbb{R}^+ \times \mathbb{R}^+; \{w_1; w_2\} \rangle$, with the following welfare functions:

$$w_i(q) = aQ + \frac{b}{2}Q^2 - \frac{c}{2}q_i^2, \qquad a, c > 0 \text{ and } 4b < c, \tag{61}$$

with $q_i \in \mathbb{R}^+$, $Q = (q_1 + q_2)$ if $b \geq 0$ and $Q = \min\{(q_1 + q_2); |a/b|\}$ if $b < 0$ (this last condition ensures that the function $B(Q)$ is always increasing). We have therefore a collective action game, in which both players contribute to a collective benefit and support individual costs. We now assume that both players choose their utility function in a first stage game as a weighted sum of the welfare functions. We have therefore, $\forall i \in N$:

$$u_i(q|\sigma_{ii}, \sigma_{ij}) = \sigma_{ii} w_i(q) + \sigma_{ij} w_j(q). \tag{62}$$

We can easily check that $\sigma_{ii} = 0$ cannot be a first stage equilibrium (if $b > 0$, player i produces $q_i \to +\infty$ and gets her worst level of welfare; if $b < 0$, player i produces $q_i = |a/b|$ and supports all the costs). We normalize therefore each parameter σ_{ii} by 1, and consider the following utility functions, $\forall i \in N$, $j \neq i$:

$$u_i(q|\sigma_i) = (1 + \sigma_i)(aQ + \frac{b}{2}Q^2) - \frac{c}{2}q_i^2 - \sigma_i \frac{c}{2}q_j^2. \tag{63}$$

A strategy profile $\bar{q} \in \mathbb{R}^+ \times \mathbb{R}^+$ is a second stage equilibrium if and only if (we assume the existence of an interior solution $\bar{q}_i > 0$, $\forall i \in N$), $\forall i \in N$:

$$\begin{cases} (1 + \sigma_i)(a + b\bar{Q}) - c\bar{q}_i = 0, \\ (1 + \sigma_i)b - c \leq 0. \end{cases} \tag{64}$$

We obtain the following solutions for the second stage game:

$$\bar{q}_i(\sigma_1, \sigma_2) = \frac{a(1 + \sigma_i)}{c - (2 + \sigma_1 + \sigma_2)b}, \tag{65}$$

$$\bar{Q}(\sigma_1, \sigma_2) = \frac{a(2 + \sigma_1 + \sigma_2)}{c - (2 + \sigma_1 + \sigma_2)b}. \tag{66}$$

We can firstly check that at the second stage equilibrium, if $b < 0$ then we have well $\bar{Q} < |a/b|$. We can then deduce the partial derivative of the indirect welfare function $v_i(\sigma_1, \sigma_2) = w_i(\bar{q}(\sigma_1, \sigma_2))$:

$$\frac{\partial v_i}{\partial \sigma_i}(\bar{\sigma}_1, \bar{\sigma}_2) = \frac{a^2 c(b(1+\bar{\sigma}_1)(1+\bar{\sigma}_2) - c\bar{\sigma}_i)}{(c-(2+\bar{\sigma}_1+\bar{\sigma}_2)b)^3} \quad = \quad 0, \qquad \forall i \in N. \tag{67}$$

We get the following symmetric solutions at the first stage equilibrium:

$$\bar{\sigma}_1 = \bar{\sigma}_2 = \frac{c-2b-\sqrt{c(c-4b)}}{2b}, \tag{68}$$

$$\bar{\sigma}'_1 = \bar{\sigma}'_2 = \frac{c-2b+\sqrt{c(c-4b)}}{2b}. \tag{69}$$

We can check that only the solution (68) verifies the positivity of q_i, and is then an interior solution. Moreover, we can check that the indirect welfare function is concave, whatever the sign of b is. We obtain therefore a unique subgame perfect $\mathcal{P}$-equilibrium $(\bar{q}; \bar{S}) \in \{\mathbb{R}^+\}^2 \times \mathbb{R}^2$:

$$\begin{cases} \bar{q}_i = \dfrac{2ab - c + \sqrt{c(c-4b)}}{2b\sqrt{c(c-4b)}}, & \forall i \in N, \\ \bar{\sigma}_i = \dfrac{c-2b-\sqrt{c(c-4b)}}{2b}, & \forall i \in N. \end{cases} \tag{70}$$

We can notice that the parameters $\bar{\sigma}_i$ have the same sign than the parameter b, i.e. that players will play cooperatively — and produce a higher output than at Nash equilibrium — if and only if the game is supermodular, even if the cooperation is not full (we have indeed $\bar{\sigma}_i < 1$). We can finally check that the profits of both players are superior to the profits at Nash equilibrium if and only if the game is supermodular. Conversely, for a game with a concave benefit function ($b < 0$), the players will be more competitive at the first stage equilibrium and will therefore get a lower outcome. Indeed, with a concave benefit function (i.e. with strategic substitutes), each player has an incentive to "blackmail" the other one — i.e. to unilaterally decrease her own output — in order to force the other player to increase her output. Since both players have the same logic, they enter in a vicious circle and end up with a deteriorated situation.

5 Conclusion

There exists a large literature on the possibility of making strategic commitments in games, but they generally focus on a specific type of commitment such as payoff distortions. We therefore suggested reasoning directly on the underlying preference relations of the players, and not only on their welfare functions. This enabled us to develop a more general framework of strategic commitment, in which we did not need any more a cardinal notion of utility. We argued that the rational choice of one's preferences is a consequence of individual rationality, and that a subgame perfect $\mathcal{P}$-equilibrium probably offers a more accurate representation of a rational behaviour than Nash equilibrium: the former corresponds to a strategy profile for which the individuals have maximized their welfare, knowing that the other players want also to maximize their welfare, whereas the latter is a strategy profile for which the individuals have maximized their welfare, knowing the strategy of the other players. Since it is the *objective* of welfare maximization which is common knowledge, and not the effective *strategy* implemented in order to maximize one's welfare, a subgame perfect $\mathcal{P}$-equilibrium is probably a more relevant notion than a Nash

equilibrium for modelling rational behaviour. In particular, our framework can explain why — without any communication between the players — a player can unilaterally cooperate in a prisoner's dilemma, or how players can coordinate themselves in a Hi-Lo game, by adding the simple assumption that players do not play weakly dominated strategies in the first stage game.
We provided in this paper a general framework for the choice of one's preferences, without any restriction on the set of preferences the individual can choose. In particular, this implies that if the preferences of the players of a subgame perfect $\mathcal{P}$-equilibrium can be implemented thanks to a specific mechanism (such as contract with third parties), then this strategy profile should be an equilibrium in this restricted framework. Proposition 1 indicates that the players choose their preferences so that their satisfaction place them in a position of Stackelberg leadership. This means that a subgame perfect $\mathcal{P}$-equilibrium can be understood as a kind of Stackelberg disequilibrium. This explains our propositions 3 and 4, i.e. that the supermodularity structure of the game will generate cooperation, whereas submodularity will generate competition. Our framework can therefore give a theoretical justification to competitive behaviours — maximizing the difference between the outcomes for instance — since the best response of the other players can *in fine* increase the welfare of the competitive player. Our proposition 2 can seem a bit paradoxical, since it appears that, despite the fact that the players will generally choose a preference relation different from their welfare relation, if there exists a Nash equilibrium $\bar{x} \in X$ in the game Γ, then there always exists a set of preference relations $\bar{\mathcal{P}} \in \mathfrak{P}$ such that $(\bar{x}; \bar{\mathcal{P}})$ is a subgame perfect $\mathcal{P}$-equilibrium: it is indeed sufficient that the preferences of the players define $\bar{x}$ as an equilibrium in dominant strategies. The introduction of a rational choice of preferences therefore extends the set of solutions of the game. It should however be noticed that such a preference relation is quite likely to be weakly dominated by an other preference relation, such as for instance $\mathcal{W}_i$ in the Hi-Lo game.
Unlike strategic delegation or the indirect evolutionary approach which study the possible strategic commitments for a given method of commitment (either a contract or a selection over the time according to the efficiency of specific preferences), we suggested studying the possible strategic commitments that could be beneficial to the players without any restriction on the set of available preferences. It will therefore be interesting for future research to define mechanisms that would enable the implementation of a subgame perfect $\mathcal{P}$-equilibrium.

References

Fershtman, C. and Gneezy, U. (2001). "Strategic delegation: an experiment", *RAND Journal of Economics* **32**(2), 352-368.

Fershtman, C. and Kalai, E. (1997). "Unobserved delegation", *International Economic Review* **38**(4), 763-774.

Franck, R. (1987). "If homo economicus could choose his own utility function, would he choose one with a conscience?", *American Economic Review* **77**(4), 593-604.

Franck, R. (1988). *Passions within Reason – The Strategic Role of the Emotions*. New York: W.W. Norton.

Friedman, M. (1953). The Methodology of Positive Economics, in *Essays in Positive Economics*, 3-43. Chicago: University of Chicago Press.

Guth, W. and Yaari, M. (1992). "Explaining reciprocal behavior in simple strategic games: an

evolutionary approach", in Witt, U. (eds) *Explaining Forces and Changes: Approaches to Evolutionary Economics*. University of Michigan Press.

Hausman, D.M. (2012). *Preference, Value, Choice and Welfare*. Cambridge University Press.

Heifetz, A., Shannon, C. and Spiegel, Y. (2007). "What to Maximize if You Must", *Journal of Economic Theory* **133**, 31-57.

Hume, D. (2000 [1739]). *A Treatise of Human Nature*. Oxford University Press.

Kavka, G.S. (1983). "The Toxin Puzzle", *Analysis* **43**.

Mill, J.S. (1869 [1843]). *A System of Logic, Ratiocinative and Inductive*. Librairie philosophique de Ladrange.

Pareto, V. (1981 [1909]). *Manuel d'économie politique*, tome VII des Œuvres Complètes. Genève: Droz.

Pareto, V. (1968 [1916]). *Traité de sociologie générale*, tome XII des Œuvres Complètes. Genève: Droz.

Poulsen, A.U. and Roos, M.W. (2012). "Do people make strategic commitments? Experimental evidence on strategic information avoidance", *Experimental Economics* **13**, 206-225.

Samuelson, L. (2001). "Introduction to the evolution of preferences", *Journal of Economic Theory* **97**, 225-230.

Schelling, T. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Sengul, M., GImeno, J. and Dial, J. (2012). "Strategic Delegation: A Review, Theoretical Integration, and Research Agenda", *Journal of Management* **38**(1), 375-414.

Stackelberg (von), H. (1934). *Marktform und Gleichgewicht*. Vienna, Berlin: Springer.

Sugden, R. (1991). "Rational Choice: A Survey of Contributions from Economics and Philosophy", *The Economic Journal* **101**(407), 751-785.