We-thinking and vacillation between frames: filling a gap in Bacharach's theory

Smerilli, Alessandra

University of East Anglia

25 August 2010

WE-THINKING AND VACILLATION BETWEEN FRAMES: *Filling a gap in Bacharach's theory*

August 24, 2010

"Probability arises from an opposition of contrary chances or causes, by which the mind is not allow'd to fix on either side, but is incessantly tost from one to another, and at one moment is determin'd to consider an object as existent, and at another moment as the contrary. The imagination or understanding, call it which you please, fluctuates betwixt the opposite views..." *(Hume,(1739-1740)[1978]p. 440)*

1 Introduction

The idea of team-thinking or we-thinking is increasingly drawing the attention of economists. In its general formulation, it has been proposed by David Hodgson (1967), Donald Regan (1980), Margaret Gilbert (1989), Susan Hurley (1989), Raimo Tuomela (1995, 2007), and Martin Hollis (1998). Within this body of literature, Robert Sugden (1993, 2000, 2003) and Michael Bacharach (1995, 1997, 1999, 2006[1]) have developed analytical frameworks from an economic point of view.

We-thinking theories allow groups to deliberate as agents. A central concept in these theories is what has been called *team reasoning*: "Roughly, somebody 'team-reasons' if she works out the best feasible combination of actions for all the members of her team, then does her part in it" (Bacharach 2006, p. 121). In other words, when people we-reason they answer to the question: "What should we do?".

[1]The 2006 book was published after Bacharach's death. The editors, Natalie Gold and Robert Sugden, assembled all the existent materials Bacharach intended to put into the book and added their own discussion of Bacharach's plans for the chapters that were uncompleted when he died.

We-thinking theories have been introduced into the economic domain for at least three reasons:

- to give an account of a relational nature of human kind (see Sugden 2005, Bruni 2008, and Davis 2009). As Hollis puts it: "we need a more social conception of what persons are and a role-related account of the obligations which make the social world go round and express our humanity" (Hollis 1998, p. 104);

- to solve some puzzles that arise in game theory, especially linked to Hi-Lo[2] and Prisoner's Dilemma (PD) games, in which rational choice theory cannot explain cooperation or selection of the Pareto-superior equilibria;

- to explain experimental evidence about the previous games (see Tan and Zizzo 2008; Becchetti, Faillo, Degli Antoni 2009).

The main claim of scholars who analyze we-thinking is that it is a coherent mode of reasoning people may use when they face a decision problem. In fact, experimental evidence shows that, especially in some kind of games, such as coordination games, people do endorse we-thinking.[3] But, if there is a general agreement on the existence of the we-mode of reasoning and on the fact people endorse it, scholars have different opinions about the way in which we-thinking arises and how it brings people to behave in a particular way. Then different authors have proposed different analyses of the issue, and, what is more, none of these analyses is entirely satisfactory from a game theoretic point of view.

In this paper I address the issue by proposing a simple model of vacillation between the I and we-modes of reasoning, as a way in which we-thinking can arise in the face of a decision problem. The model is based on a not fully developed intuition - the double-crossing problem in the PD game - of Bacharach, whose theory is the most developed from an analytical point of view.

But, first of all, let us see how philosophers and economists have dealt with the question of how we-thinking arises.

For Gilbert (1989) and Tuomela (1995), for example, group formation, which leads to we-thinking, is a result of a mutual commitment; in particular,

[2]Hi-Lo game is in general a n player game in which each player chooses one item from the same set of alternatives. Each alternative ia associated with a prize, and one alternative's prize is greater than the others. If all players choose the same alternative they get the associated prize, if not nobody gets anything.

[3]See Tan and Zizzo (2008) for a review of experiments.

both authors refer to preference formation based on collectively accepted attitudes of a group. This approach, however, while dealing with preference formation in a well shaped theory, lies outside game theory.

Susan Hurley (1989) offers a theory of the 'rationality' of we-thinking. She identifies as units of agency the subsystem ('each') or the system ('we') and claims that those units have not to be taken as fixed. In fact, in the face of a decision problem an agent firstly must ask herself: which is the objective in this situation? Subsequently she can choose the unit of agency that is the most appropriate for the objective. But in rational choice theory, as Hollis and Sugden point out, a choice is rational in relation to the desires or preferences of the agent who is making the choice: "a choice can be rational only for a particular agent" (Hollis and Sugden 1993, p. 13). It follows that a theory related to standard rational choice, as game theory is, cannot give an account of the formation of the unit of agency.

Gauthier (1975) proposes an approach that has some similarities with Hurley's: he allows players to choose between alternative descriptions of a game, in particular each player should choose the description which could lead to the best result for everyone, provided that everyone chooses the same description. Bacharach, whose idea is similar to Hollis and Sugden's, that in rational choice theory decisions must be defined prior to choice, criticises Gauthier's approach, by saying "It seems to me that one can't just go round changing one's own description for convenience; this is like changing beliefs; surely you must describe the world as you find it"(Bacharach 2006 p. 29, quoted from his preliminary notes).

Differently from Hurley, who claims that there must be agent-neutral goals to be pursued, Elisabeth Anderson (2001) states that the determination of personal identity, which can be plural or individualistic, precedes the choice of the kind of reasoning to be adopted: "what principle of choice it is rational to act on depends on a prior determination of personal identity, of who one is" (p.30). Following the previous principle, Anderson shows that either acting on maximization of expected utility or on team reasoning is a rational act, depending on regarding oneself as an isolated individual or a member of a team. In Anderson's account, then, the determination of personal identity comes before the decision of what principle of choice is in play. Unfortunately, also Anderson's theory is not formalized, and so can not be applied to game theory.

The two main contributors from an economic point of view are Bacharach and Sugden, who approach the topic in different ways. Sugden's aim is to show that we-reasoning is a consistent and logical way of thinking, but he does not face the problem of how we-reasoning can arise. He gives only some

intuitions about a psychological background based on Smith's analysis of correspondence of sentiments (Sugden 2005).

Bacharach proposes a formal theory of games with I-reasoners and we-reasoners, with the mode of reasoning taken as given. A fundamental point in Bacharach's theory is that the determination of mode of reasoning is a psychological matter, prior to rational choice, and is given by frames. So, as he recognizes, to complete the theory he needs to build a theory of which mode of reasoning will be in play. This means to endogenize I/we determination. This part of Bacharch's theory is less developed, although he suggests some intuitions. He tries to complete his theory following two different approaches: the concept of the *harmony* of the game, which has been further developed by Tan and Zizzo (2003, 2008), and the *interdependence hypothesis,* which links to an underdeveloped intuition about vacillation between frames. Because of his death, he never achieved his aim of endogenising the determination of the mode of reasoning.

In the present paper, I shall suggest a way to complete Bacharach's theory, generalising the interdepence hypothesis and building on his intuition about vacillation. I propose a formal model of vacillation between frames, which allows individuals to switch from I to we mode of reasoning and viceversa (section 4 and 5). In order to develop my proposal, Bacharach's theory of team-reasoning will be analysed in section 2, by taking into account published and unpublished material. In section 3 I propose a discussion of some not fully developed intuitions of Bacharach, and section 6 presents the conclusions.

2 Bacharach's theory of we-thinking

"The answers to fundamental questions about coordination and cooperation... lie in the agent's conception not of the objects of choice, nor of the consequences, but of herself and of the agents with whom she is interacting" (Bacharach 2006, p. 70). This sentence is the starting point of Bacharach's analysis of we-reasoning. We-reasoning is seen as a powerful 'mechanism' (in Bacharach's words) for solving puzzles about coordination and cooperation in game theory (i.e. games like Hi-Lo and PD). On the whole, in his work Bacharach tries to demonstrate, by showing some evidence,[4] that we-reasoning is a valid mode of reasoning and that people do endorse it.

[4]Bacharach claims that there are five kinds of evidence: logical, introspective, evolutionary, transcendental and experimental. In particular he gives an account of experiments one conducted by himself and Guerra, and another one made with Bernasconi, in which they provide some behavioural evidence that group identification leads people to

Bacharach's main purpose is to explain cooperation, seen as a successful group activity (ib p. 69), and the core mechanism for doing that comprehends 'framing', 'common purpose', and 'cooperation':

> "(i) we frame ourselves as members of groups; (ii) ... perceived agreement of individual goals among a set of individuals favours framing as members of a group with this common goal; (iii) the group framing tends to issue in efficient cooperation for the group goal" (ib p. 90).

In what follows, I illustrate the building blocks of Bacharach's theory, but, first of all, I give an account of how and when Bacharach developed the idea of we-thinking. This is because the particular pattern he followed could offer hints for developing some of his intuitions, remaining faithful to his thought.

2.1 Development of Bacharach's thought

Bacharach started by building Variable Frame Theory (Bacharach 1993), when in parallel he was developing a theory of cooperation. Variable Frame Theory (VFT) is an analysis of choices in games in which frames are taken into account. VFT allows games with descriptions of players' frames. Concisely, in VFT a player can intentionally choose an object, or an action, if she has a way of thinking about that object or that action, i.e. she has a frame. Frames can be more or less salient or available, depending on a probability measure on them.

Bacharach's aim in developing VFT was to explain the choice of focal points in games: by making use of VFT he could turn focal point problems into Hi-Lo games. We-thinking theory, as proposed by Sugden (1993), helped him to explain the selection of Pareto-superior equilibria in Hi-Lo and in coordination games more generally. He started then, to develop his own theory of we-reasoning.

In 1995 he introduced the category of 'fellow member reasoner':

> "Someone who is a member of a natural type T and chooses a certain strategy if she is sufficiently sure that her interactants are also member of T" (1995, p.1).

In this context he tries to link T-membership to VFT and, at the same time he introduces the 'we' category:

we-reasoning (see Bacharach 2006, pp.145-146, and Bacharach and Bernasconi 1997).

> "The present paper has made type T membership an issue which type T members think about, and nuanced their capacity to recognize it. An alternative development would make T membership a variable element in players' frames in the sense of variable frame theory: that is, a player might or might not think about the game in terms of whether she and her coplayers belong to T. In the case in which T is the player set, we may put this by saying that a player may or may not think in 'we' terms about how to play the game. The more inclined a player is to 'we' thinking, and the more inclined she takes coplayers to be, the more will fellow-member reasoning be favoured" (p.17).

In 1997 Bacharach formally introduces we-thinking, in an unpublished paper whose title is: "'We' Equilibria: A Variable Frame Theory of Cooperation". The first published paper in which Bacharach formalizes his theory is an article published in 1999 about 'interactive team reasoning'. In it Bacharach introduces some elements that we can find in the book, such as, group identification, team reasoning as the effect of group identification, unreliable team interaction (which in the book becomes cirscumspect team reasoning), etc. Between the 1999 article and the book we may find some lecture notes, in which the concepts of agency and 'superagency' begin to appear. The book represents an (incomplete, because of his death) attempt to build a complete, and at the same time simple, theory of we-thinking: I shall present in the following subsections the theory as it appears in the book.

2.2 I-reasoning and we-reasoning

First of all, Bacharach allows for the existence of both I and we modes of reasoning. Each is seen as rational maximization of a von Neumann - Morgenstern utility function. I-reasoning is represented by a standard utlity function. We-reasoning, instead, requires a team utility function (W) : "a game-theoretic treatment of agents who may group-identify must... determine a payoff function to represent the group objective"(Bacharach 2006 p. 87). In order to clarify what the group identification process implies about what the players want as a group, or, in other words, in order to clarify what W is, Bacharach proposes that W must satisfy the 'Paretianness' condition: if a profile of actions p is weakly Pareto-superior to p', then $W(p) \geq W(p')$. Examples of group utility functions include the utilitarian function and weighted utilitarian functions. This means that group objectives are related to personal ones. Another important point for Bacharach

is to allow principles of symmetry and fairness between individual payoffs[5] to be embedded in W.

2.3 Frames

For Bacharach, modes of reasoning are not chosen rationally: it is a psychological process that determines which mode of reasoning will be in play.

This process is based on frames: if the we-frame comes to mind, the subject will group identify and then she will start to we-reason. A frame can be defined as a set of concepts that an agent uses when she is thinking about a decision problem. It cannot be chosen, and how it comes to mind is a psychological process:

> "Her frame stands to her thoughts as a set of axes does to a graph; it circumscribes the thoughts that are logically possible for her (not ever but at the time). In a decision problem, everything is up for framing... also up for framing are her coplayers, and herself" (ib. p. 69).

In Bacharach's framework a person may start to we-reason only if she has 'we' concepts in her frame. If the we-frame is active in a subject she begins to think of herself as a part of a collective actor, then she begins to team-think, and this means that in the face of a decision problem she will answer the question: "What shall we do?". In Bacharach's theory then, to see the we-frame implies to endorse that frame.

In his theory group identification is a framing phenomenon that determines choices by "changing the logic by which people reason about what to do" (ib). If, by reasoning in the individual standard mode (I-reasoning), an agent looks at a decision problem by thinking what it would be best for her to do, when there is group identification, the agent will think: "What would best be for us to do?". Basically then, "Somebody 'team reasons' if she works out the best feasible combination of actions for all the members of her team, then does her part in it"(ib 2006 p.121).

2.4 Circumspect team reasoning

One of Bacharach's aims is to explain situations in which some people may 'we'-reason and some others may not. In order to model these situations, he assumes that the 'we' frame comes to mind with probability 'ω', which represents the probability that a subject group-identifies. The probability ω

[5]"Such as those of Nash's axiomatic bargaining theory"(Bacharach 2006, p. 88).

is common knowledge:[6] "in coming to frame a situation as a problem 'for us', an individual also gains some sense of how likely it is that another individual would frame it in the same way" (ib p. 163). A context in which some people may group-identify and some may not is seen by Bacharach as an *unreliable* coordination context, and team reasoning in this context is called *circumspect* team reasoning. Briefly, people who we-reason in an unreliable coordination context look for the best available profile o - the combination of actions - that maximizes W given that each person will choose to do her part in o with probability ω, or will fail - i.e. act on I-reasoning - with probability $(1\text{-}\omega)$.

One problem which remains open in Bacharach theory is the endogenization of ω: he sees the need for endogenization and proposes some speculations, but he did not complete this part of the theory, as we shall see later.

2.5 Variable frame theory and 'vacillation'

Bacharch's (never reached) aim was to explain we-reasoning in terms of Variable Frame Theory (VFT), which he had developed in a earlier stage of his investigation[7].

The intersection between VFT and we-thinking would have been called by Bacharach 'Variable Agency Theory' (Bacharach 2006, p.59). However, the completion of the description of we-reasoning in terms of VFT raised problems that he had not solved at the time of his death. Let us see these problems.

In Bacharach's circumspect team reasoning, as I have said before, if people group identify, then the we-frame comes to their mind and they start to we-reason. It seems as though in Bacharach's framing theory there are two aspects that are deeply linked: in framing a situation, the first step is to recognize a frame, that is coming to see it; the second step is endorsing that frame, i.e. reasoning as the frame allows you to do. In Bacharach's theory group identification means not only coming to see a particular way of reasoning, but also endorsing it. The 'compression' between the two aspects of framing is due to the VFT. However, in the original form of VFT, changing frame does not mean to change the way of reasoning, and the decision prob-

[6] In a previous work (1999), Bacharach has developed a more formalized model, in which each agent can participate or lapse in a team and everyone, before choosing, receives a signal knowing the joint probability distribution of this signal and agent's state (i.e. an agent's signal includes her participation state).

[7] See Bacharach 1993, 2001.

lem for a subject is fully determined by the interplay of his frame and the objective world. VFT was originally thought as a way to allow of a player to frame different situations differently, but frames were not related to different agencies. In constructing his Variable Agency Theory, Bacharach was trying to use VFT in a new way, but, because of this 'compression', he could not allow people to use more than one frame at a time. In a certain sense, as it has been noticed by Gold and Sugden (Bacharach, Gold and Sugden 2006), in we frame people become committed to we-reason:

> "In the theory of team reasoning, an individual who reasons in the 'We' frame is aware of the 'I' frame too (as one that other players might use) but acknowledges only 'We' reasons. It seems that group identification involves something more than framing in the sense of variable frame theory: the group-identifier does not merely become aware of group concepts, she also becomes committed to the priority of group concepts over individual ones" (p.199).

In one of his unpublished papers (Bacharach, 1997), Bacharach allowed for the possibility of the existence of three frames: the I frame, the We frame and the 'S' (superordinate) frame. We and I are called simple frames: "players in them begin their reasoning with the two basic conceptualization of the situation, as 'what shall we do?' problem and 'what shall I do?' problem respectively" (p. 5). A S frame is active when someone manages "during deliberation to see the problem from both the we and the I/she perspectives"(p.14). Although Bacharach allows for the existence of S, based on psychological attainments, he states that We and I perspectives cannot be held simultaneously: "Although we can switch self-identities rather easily, we appear to be unable to inhabit more than one at a time" (p.15). He assumes that I thoughts in the S frame generate a personal evaluation, whereas we thoughts generate a group evaluation. The solution concept in the model roughly states that the cooperative option is chosen by a player in S if it is the best in group evaluation and not worse than the other option in personal evaluation.

The S-frame intuition of the 1997 unpublished paper, however, disappeared in subsequent pieces of work.

Later on, in developing the VFT Bacharach faces the issue of integrability of frames. He says that normally frames are integrable:

> "It is easy to integrate frames which consist of classifiers such as shape, colour and position: we can easily see a mark as a triangle,

> as a blue triangle, as a blue triangle on the left,... on the other hand... a person can see the marks as letters and as geometric shapes, but not at the same time – you can't integrate these two perceptions" (2001 a, p.6).

There exist frames, then, that are non-integrable. 'I' and 'we' frames appear non integrable in Bacharach's words, and when this happens, "the agent may find herself vacillating between the judgments that she should do" (ib.). The idea that an agent can 'vacillate' between the two frames was so important for Bacharach that one of his (not realized) desires was to have Rubin's vase (Figure 1) on the front cover of the book.[8]

I shall suggest later that it is possible to take into account what Bacharach called 'personal' and 'group' evaluation, by reasoning in terms of deviation from an equilibrium and not in terms of frames. Or better, it is possible to do that, if we separate the two aspects of framing: how a frame might come to mind and how a person endorses a particular frame when she sees more than one frame.

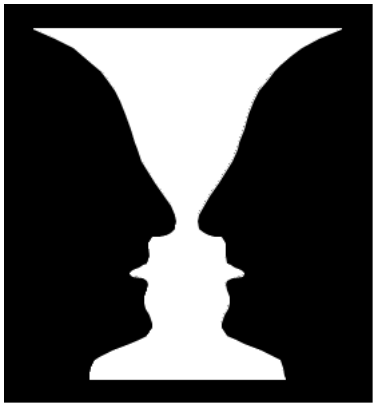

Figure 1: Rubin's vase

3 The determination of mode of reasoning

To complete the theory, Bacharach needs to endogenize the determination of mode of reasoning (this means the endogenization of ω). He tries to endogenize ω, because he sees that the fact that ω is exogenous as a lacuna in his theory[9]. As we have seen in section 2.2, the theory of the determination of mode of reasoning has to be not a theory of rational choice.

[8] This follows from a personal communication with Robert Sugden, who inferred this desire by managing Bacharach's incomplete manuscript, which displayed Rubin's vase image on the first page.

[9] "The unreliable team explanation of co-operative behavior I outlined in this paper contains an important lacuna. The distribution of agents over teams and the probability that they are active, are exogenous" (Bacharach 1999 p. 144).

In his earlier works (Bacharach, 1997, 1999) he proposes that the possibility of team reasoning is related to having 'scope for cooperation' and to the 'harmony of interests'. Harmony is a non-strategic assesment of the game:

> "To endogenize ω, and other features of Ω, one must show that the payoffs and other constitutive features of the basic game make collective identity salient or otherwise tend to induce team-thinking. The laboratory evidence is promising, as it suggests that group identification may be induced by the 'common problem' mechanism'. In addition, it is plausible that ω may be an increasing function of certain quantitative features of the payoff structure, such as 'scope for co-operation' and 'harmony of interest'" (1999, p.144).

A step forward on this topic has been made by Tan and Zizzo:[10] in their work there is an attempt to investigate the relationship between harmony of interests ('game harmony' for them), group identification and cooperation. They claim that game harmony is a good predictor of the extent of cooperation or conflict in games. They postulate that "game harmony increases cooperation by increasing the probability of team reasoning on the part of different players" (Zizzo 2004, p.20). Game harmony, defined as "a generic property describing how harmonious or disharmonious the interests of players are, as embodied in the payoffs" (Tan and Zizzo 2008, p. 3), is based on the correlation coefficient between payoff pairs - Pearson or Spearman correlation coefficient between the payoffs of the players for each state of the world for two player games[11]. This measure is the best existent proxy for what Bacharach has called 'the harmony of interest', and it is entirely derived from the payoffs of the game. It is a potential solution of Bacharach's problem of endogenization of ω. However, some of Bacharach's intuitions about vacillation cannot be expressed by the game harmony approach.

In fact, Bacharach tries a second line in order to endogenize ω. This is the (strong) Interdependence Hypothesis, that roughly states: perceived interdependence prompts group identification. The perception of interdependence between two agents in a game is given by three factors:

- common interest (the agents have common interest in some s* over s, if both prefer s* to s, where s*, s are possible states of affairs, or, in a game, possible outcomes)

[10] See Tan, J. and D. Zizzo (2008), Zizzo D. and Tan, J. (2003), and Zizzo D. (2004)

[11] More in general, this measure is an average of Pearson (Spearman) correlation coefficients among payoff pairs.

- copower (nobody can reach s* alone, but both can together)
- standard solution (basically the existence of a Nash equilibrium that realises s).

Basically, if an outcome that can be reached by an individual way of reasoning (standard solution) is Pareto-dominated by another outcome achievable only by thinking as a group, there is space for group identification.

The interdependence hypothesis uses I-reasoning as a default, makes use of opportunities for we-deviations that are good for 'us' and treats these opportunities as prompting we-reasoning. Interdependence fits with the intuition Bacharach had about the vacillation between frames, but it seems to give an account only of we-deviation from I-thoughts. What about the opposite, that is from we to I?

Bacharach offers only an informal conjecture about deviation from we to I: the 'double-crossing intuition'. Taking the most famous game in terms of cooperation, the PD game, as an example, Bacharach says:

> "In a Prisoner's Dilemma, players might see only, or most powerfully, the feature of common interest and reciprocal dependence which lie in the payoffs on the main diagonal" (Bacharach 2006, p.86).

If this happens, players do cooperate. But, it might be the case that

> "they might see the problem in other ways. For example, someone might be struck by the thought that her coplayer is in a position to double-cross her by playing D in the expectation that she will play C. This perceived feature might inhibit group identification" (ib).

Here Bacharach seems to have in mind some psychological process which inhibits group identity and which is not quite represented by his own concept of interdependence – the idea of 'double-crossing'. The reason this idea does not fit his framework is that double-crossing is the incentive to act on individual reasoning when one believes the other is acting on team reasoning. And, what is more, double-crossing is a reason for a person who we-reasons to switch to the I-frame. A player, in order to recognize the 'double-crossing' threat, should be allowed to imagine herself in a we-frame, and then deliberating to cooperate, but at the same time she should use the I-frame by thinking that the other player would take advantage of her. In the first player's conjecture, the other player too should use the we-frame in order

to think that the first player could choose to cooperate, and, at the same time, she should use I-frame in order to think how to 'double cross' the first player. We may formalize what the statement 'i double-crosses j' means, i and j being the two players:

- i defects; i believes that j will cooperate; i believes that j believes that i will cooperate. $[a]$

And 'j thinks that i will double-cross j' means:

- j thinks $[a]$.

It is now clearer that j's thoughts include: i acting on I-reasoning; i attributing we-reasoning to j; i attributing to j: attributing we-reasoning to i.

In the theory of we-thinking the way in which a person reasons (I-mode or we-mode) is a consequence of the perceived frame. She may switch from the I-mode of reasoning to we-reasoning (if the we-frame comes to mind), or not. Bacharach, then, does not seem to take into account the possibility that once we are in the we-frame, we may switch to the I-mode of reasoning, or better, he allows the possibility of switching frame, but does not allow a person to be able to visualize switching frames. And this is why he cannot represent his 'double-crossing' intuition. It seems also, that when the "we" frame is perceived, it is also perceived as the correct frame or dominant frame, so that once a person sees the world this way she cannot visualize going back to seeing it the other way (compare illusions, myths, lies – 'the scales fell from my eyes').

In order to complete Bacharach's theory in a more formal way, we need a model of vacillation, with deviation both from I to We and from We to I.

I shall present a first step in the next section, where I propose a representation of the double-crossing intuition.

4 Representing the 'double-crossing' intuition: reasoning in terms of deviations from equilibrium

In what follows, I shall present my analysis in terms of individual and collective rationality as two alternative ways of approaching a decision problem, and in particular I shall focus on reasons for deviating from an equilibrium. For simplicity I am considering two-player games, but the analysis could be easily extended to n-player games.

First of all, I suppose that the group utility function of a combination of actions (a1, a2), when the players are P1 and P2, is W(a1, a2) = (u1(a1, a2) + u2(a1, a2))/2,[12] where u1(a1, a2) and u2(a1, a2) are the player's payoffs. A player who team reasons, first computes which is the best profile for the group[13], and then he does its part in it. A player who 'I'-reasons follows the standard theoretic predictions of game theory.

It is possible to classify games in terms of reasoning about deviations. The basic idea is that a person may reason in the standard I-mode, or in we-mode, but she may have both frames (I and we) in mind (perhaps not at the same time, if the non-integrability hypothesis is correct, but vacillating between them). In standard game theory an equilibrium is defined as a combination of actions in which no player has anything to gain by changing unilaterally her own strategy. In we-reasoning theory, an equilibrium is instead defined as a combination of actions in which the whole group cannot gain anything by switching from this combination to another. Deviation is seen then as a test for the existence of an equilibrium, no matter if I or we-equilibrium. As a first step in my analysis, I shall simply test games in search for equilibria that hold from both I and we points of view.

Table 1: game A

	L	R
U	***3,3***	4,1
D	1,4	2,2

Take, for example, the game A (table 1). The combination of actions (U, L) is a Nash equilibrium. Neither row player nor column player has reason to unilaterally deviate from that combination of actions. But the same combination is also a we-equilibrium: as a group both players cannot do better by switching to another combination[14].

[12]This formulation is the most used one in literature, although any W which satisfies the Paretiannes condition is acceptable.

[13]In this version of the model, I am not taking account of the problems of 'unreliability' that Bacharach models by mean of circumspect team reasoning. The focus here is on vacillation, and at this stage I want to keep the model as simple as possible.

[14]The utility U for the group is 3 in the (U, L) combination, 2.5 in both (U, R) and (D, L), and 2 in (D, R).

Table 2: game B

	L	R
U	***2,2***	3,0
D	0,3	***2,2***

Game B shows a unique Nash equilibrium, (U, L) and two we-equilibria, (U, L) and (D, R), but only the (D, L) combination is an equilibrium at the same time for I and we-reasoners.

Table 3: game C

	L	R
U	**3,3**	1,1
D	1,1	***2,2***

Game C is Hi-Lo game, and as is well known, it has two Nash equilibria, i.e. (U, L) and (D, R), but only one we-equilibrium, that is (U, L).

Table 4: game D

	L	R
U	***3,3***	1,4
D	4,1	**2,2**

Game D is a PD game, it has one Nash equilibrium (D, R) and one we-equilibrium (U, R), but these do not coincide.

If an equilibrium survives both I and we deviation tests, it is particularly strong, in the sense that it allows for the existence of both ways of reasoning. At the same time such an equilibrium could be seen as a refinement when more than one equilibrium exists. I shall call this equilibrium: *I-we equilibrium*. In game B, for example, there are two we-equilibria, but if we allow players to see the game endorsing both I and we concepts, this could help them to recognize that the (U, L) equilibrium is the prominent one, because it passes both deviation tests. In this case, having an I thought helps we-reasoners to select an equilibrium. But the opposite can happen as in the Hi-Lo game, where there are two Nash equilibria and we-thoughts can help I-reasoners to choose the (U, L) equilibrium.

This double test for deviation could also be seen in terms of deliberations, and not only as a method for testing the existence of an equilibrium. It can represent a model of transition between modes of reasoning, and as

a component of a model of vacillation between them. The scheme in figure 2 represents a possible way to classify the previous games in terms of deliberation, or vacillation.

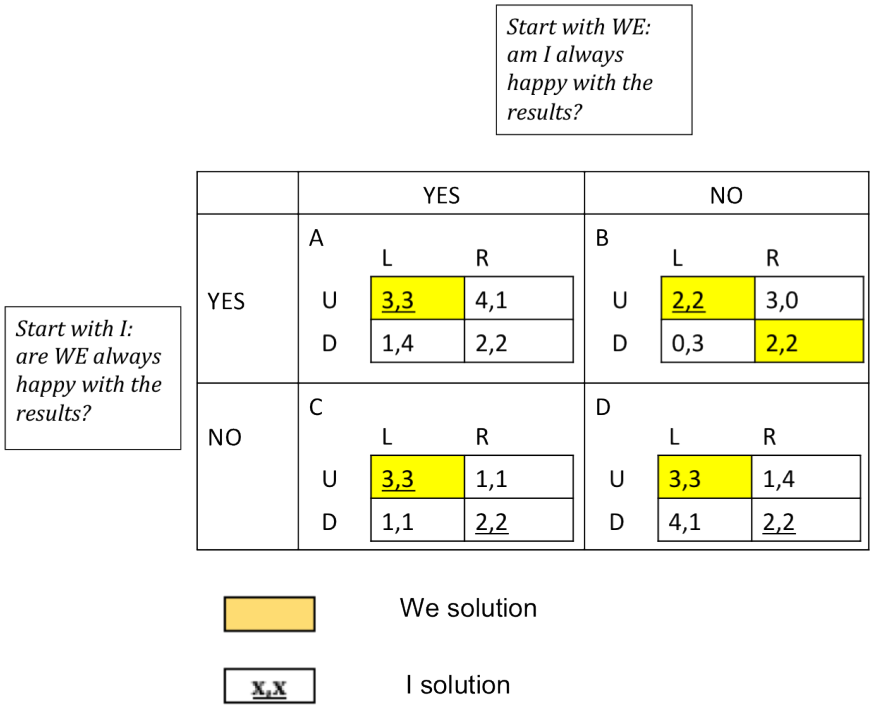

Figure 2: Reasoning about deviations and deliberations

Take for example game A: in this game, if I start to reason in the standard I-mode, we as a group will be happy with the result (U, L), i.e. we shall not want to deviate jointly from the I-reasoning 'solution'. Conversely, if I group identify, and then I look for the best solution for the group, I as an individual will be happy with the result, i.e. I will not want to deviate unilaterally from the we-reasoning solution. So, in this game, the same result will be reached, independently of the particular way of reasoning. We may say that I or we-reasoning are observationally indifferent or equivalent, because they give the same result in terms of choice.

But there could be different situations. Let us look at game B: in this case, if I start with the I-mode, there will be a unique Nash equilibrium (U, L), which is also one of the two possible (and indifferent) we-solutions. If I start with the I-mode, we shall then be happy with the result. If we group

identify and we-reason, if we-reasoning gets us to (U, L), I am happy. But if it gets us to (D, R), I am unhappy. I may then turn to the I-mode of reasoning. In a vacillation process if, when reasoning in one mode, the conclusion is not endorsed by reasoning in the other mode, there is some tendency to switch to that other mode. So, in this case, the end of the vacillation process is the outcome (U, L), either by we-reasoning or by I-reasoning. This result is observationally equivalent to I-reasoning but not to we-reasoning, because the latter allows (D, R).

Game C, instead (the Hi-Lo game), is a mirror image of game B and will prompt we-reasoning: if we start by we-reasoning there will be a unique we-equilibrium (U, L), which is also one of the two possible Nash equilibria. So if we start with the we-mode, I will be happy with the result, and we shall not move from the (U, L) equilibrium. If instead I start with the I-mode we shall not always be happy: if the solution is (U, L), we shall be happy, but if it is (D, R) we shall not be happy, and we may turn to the we-mode of reasoning.

The last game, the PD game, is the most interesting: if I start with I-reasoning, we shall not be happy (the Nash equilibrium is Pareto-dominated by the we solution). But if we group identify the we solution is not good for me (I would be better off by playing the other strategy). In this case there can be a continuous switching or vacillation from a frame to another: this could be an explanation of the empirical evidence on behaviour in PD games. In fact, in experiments on the PD game, we observe a rate of cooperation of about 50% (see Sally 1995). Following Bacharach's interdependence hypothesis, the PD, as we have seen, is one of the typical games that can lead to we-reasoning, although Bacharach himself was aware of double-crossing threat. In the framework I have presented, the double-crossing intuition is taken into account, and this generates perpetual shifts between modes of reasoning, and then we-reasoning is only one of the two equally possible solutions. It is plausible that in cases like this, the salience of frames will play a key role in the selection of the solution of the game.

This way of looking at a decision problem does not tell us which frame is more likely to appear. But, if a frame comes to mind, within this classification, we may see, depending on the kind of game the subject is facing, if the frame will be stable or not, or, in other words, we might see if that frame is an absorbing state in a model of transition or vacillation between frames.

In order to say something more about games with conflicting frames, in the next section I propose a formalization of the intuitions embedded in the previous classification of games.

5 A more formal vacillation model

The classification I proposed in the previous section represents a first step towards a generalization of Bacharach's model and intuitions. In the present section I show a possible way to generalize the previous results: I sketch a simple model based on I and We temptations to deviate from an equilibrium, and on a possible refinement of equilibria.

We suppose that there are two players: 1, 2

S_1, S_2 are the strategies chosen by players 1, 2.

Let us define the following finite utility functions:

$U_1(S_1, S_2) =$ 1's individual utility

$U_2(S_1, S_2) =$ 2's individual utility

$W(S_1, S_2) =$ We-utility.

Individual and group utilitiy functions have the characteristics specified in section 2.2., i.e. the individual utility is represented by a standard von Neumann - Morgenstern utility function, and group utlity is represented by a team utility function which satisfies the Paretiannes condition.

Considering any equilibrium (S_1^*, S_2^*) from the viewpoint of player 1, we define:

- Own temptation to deviate $= \underset{S_1}{max} [U_1(S_1, S_2^*) - U_1(S_1^*, S_2^*)] \equiv T_1(S_1^*, S_2^*) \geq 0$
- Other's temptation to deviate $= \underset{S_2}{max} [U_2(S_1^*, S_2) - U_2(S_1^*, S_2^*)] \equiv T_2(S_1^*, S_2^*) \geq 0$
- Our temptation to deviate $= \underset{S_1, S_2}{max} [W(S_1, S_2) - W(S_1^*, S_2^*)] \equiv T_W(S_1^*, S_2^*) \geq 0$

We have a Nash equilibrium when the following conditions hold:

$$\begin{cases} T_1(S_1^*, S_2^*) = 0 \\ T_2(S_1^*, S_2^*) = 0 \end{cases} \qquad [1]$$

A We-equilbrium is given when:

$T_W(S_1^*, S_2^*) = 0 \qquad [2]$

An I-we equilibrium exists when both [1] and [2] conditions hold.

I-we equilibirum can be seen as:

- (i) a refinement of Nash equilibirum
- (ii) a refinement of We equilibrium

An I-we equilibrium helps to refine I-equilibria from a we point of view and we-equilibria from an I point of view, as we have seen in the classification in figure 2.

But there could be cases, as in the PD game, in which an I-we equilibrium does not exist, because the conditions [1] and [2] can not both be met. At the same time it is possible to have cases with more than one Nash or we equilibrium. When this happens, we can imagine other kinds of refinements. For example, the following could be a refinement of Nash equilibrium:

- (a) choose the Nash equilibrium which minimizes $T_W\left(S_1^*, S_2^*\right)$

Or, in case of more than one we-equilibria, a refinement could be:

- (b) choose the we-equilibrium which minimizes $f\left(T_1\left(S_1^*, S_2^*\right), T_2\left(S_1^*, S_2^*\right)\right)$ where $f\left(...\right)$ is an increasing and finite (for finite T_1, T_2) function, with $f\left(0,0\right) = 0$.

Our research question now is if there is a way to remain faithful to Bacharach's intuitions about vacillation, by considering a way to refine the equilibria. We suggest a possible refinement:

- treat (a) and (b) as the candidate equilibria; we call
  candidate Nash equilibrium $\equiv(S_1^*, S_2^*)$
  candidate we equilibrium $\equiv(S_1^{**}, S_2^{**})$
- Then the probability that (S_1^*, S_2^*) is viewed as the solution by player 1, i.e. the probability of acting in I-mode, is given by:

$$pr\left[(S_1^*, S_2^*)\, is\, the\, solution\, for\, 1\right] = \mathrm{h}(T_W\left(S_1^*, S_2^*\right), f\left(T_1\left(S_1^{**}, S_2^{**}\right), T_2\left(S_1^{**}, S_2^{**}\right)\right)) \quad [3]$$

where $h\left(...\right)$ is decreasing in $T_W\left(S_1^*, S_2^*\right)$ and increasing in $f\left(T_1\left(S_1^{**}, S_2^{**}\right), T_2\left(S_1^{**}, S_2^{**}\right)\right)$.

In the same way we can say that the probability that (S_1^{**}, S_2^{**}) is viewed as a solution by player 1, i.e. the probability of acting in the we-mode, is given by:

$pr\left[(S_1^{**}, S_2^{**})\, is\ a\ solution\ for\ 1\right]$=1−$pr\left[(S_1^{*}, S_2^{*})\, is\ the\ solution\ for\ 1\right]$[4]

where $h\left(...\right)$ is decreasing in $T_1\left(S_1^{**}, S_2^{**}\right), T_2\left(S_1^{**}, S_2^{**}\right)$and increasing in $T_W\left(S_1^{*}, S_2^{*}\right)$.

The same holds for player 2.

The [3] and [4] combination of probabilities represents a refinement of both Nash and we-equilibria and we might call it *the vacillation* refinement.

The probabilities used in vacillation refinement could be expressed in terms of Markov transition processes. In fact, Markov chains, with their properties, seem to represent the best candidate for a model of vacillation. Let us see how.

Suppose that the state space is $\Omega = \{I, W\}$,where I is I-reasoning and W is we-reasoning, and that $(X_0, X_1, ...)$ is the sequence of possible states of the process. If we call p the probability of transition from we to I, and q the probability of transition from I to we, we may define the transition matrix:

$$P=\begin{pmatrix} P(I,I) & P(I,W) \\ P(W,I) & P(W,W) \end{pmatrix}=\begin{pmatrix} 1-q & q \\ p & 1-p \end{pmatrix}$$

Let π_t be the probability distribution at t: $\pi_t = \begin{bmatrix} \pi_{It} & \pi_{Wt} \end{bmatrix}$ is a row vector whose components are the probability of I-reasoning at t, and the probability of We-reasoning at t. It is known that:

$\pi_{t+1} = \pi_t P$

So that, for example, $\pi_{It+1} = \pi_{It}(1-q) + \pi_{WEt}p = \pi_{It}(1-q) + (1-\pi_{It})p$

By imposing the Markov property we have: $\pi_{t+1} = \pi_t = \pi$

and then (after a little algebra) we obtain:

$\pi = [\pi_I, \pi_W] = \left[\frac{p}{p+q}, \frac{q}{p+q}\right]$

Suppose (as a simplification) that p, the probability of transition from we to I is:

$p = \alpha f\left(T_1\left(S_1^{**}, S_2^{**}\right), T_2\left(S_1^{**}, S_2^{**}\right)\right) \quad$ with $\alpha > 0$

and q, the probability of transition from I to we is:

$q = \alpha T_W\left(S_1^{*}, S_2^{*}\right)$

Then

$\pi_I = pr\left[(S_1^{*}, S_2^{*})\, is\ a\ solution\ for\ 1\right] = \frac{f\left(T_1\left(S_1^{**}, S_2^{**}\right), T_2\left(S_1^{**}, S_2^{**}\right)\right)}{T_W\left(S_1^{*}, S_2^{*}\right) + f\left(T_1\left(S_1^{**}, S_2^{**}\right), T_2\left(S_1^{**}, S_2^{**}\right)\right)}$

which satisfies [3]

and

$\pi_W = pr\left[(S_1^{**}, S_2^{**})\, is\, a\, solution\, for\, 1\right] = \frac{T_W\left(S_1^*, S_2^*\right)}{T_W\left(S_1^*, S_2^*\right) + f\left(T_1\left(S_1^{**}, S_2^{**}\right), T_2\left(S_1^{**}, S_2^{**}\right)\right)}$

which satisfies [4].

The probabilities π_I and π_W, where $\pi_W = 1 - \pi_I$, are entirely derived from the temptations to deviate from an equilibrium, and represent, as we have already said, a refinement of equilibria when an I-we equilibrium does not exist, or there are more than one Nash or we equilibria. The model is then complete.

To see how these probabilities work in practice, take for example the following game:

Table 5:

	a	b	c
a	10,10	0,0	0,20
b	0,0	9,9	0,10
c	20,0	10,0	1,1

In this game (a,a) is the unique We-equilibrium, and (c,c) is the unique Nash equilibrium. They do not coincide.

As an illustration, suppose that

$f\left(T_1\left(S_1^{**}, S_2^{**}\right), T_2\left(S_1^{**}, S_2^{**}\right)\right) = T_1\left(S_1^{**}, S_2^{**}\right) + T_2\left(S_1^{**}, S_2^{**}\right)$

and that the group utility function W has the same form as the function used in section 4.

A simple vacillation model might give:

$\pi_I = pr\left[1\, plays\, c\right] = \left(\frac{10+10}{10+10+9}\right) = \left(\frac{20}{29}\right)$

$\pi_W = pr\left[1\, plays\, a\right] = \left(\frac{9}{10+10+9}\right) = \left(\frac{9}{29}\right)$

This game belongs to category D (the category which include the PD game) following the scheme proposed in the previous section: this means that the deliberation process leads to a continuous switching between frames. The structure of payoffs, however, and then the dimension of I and We temptations to deviate, allow us to infer that the strategy associated with the I solution is more likely to be selected by each player.

A further step of research could be a comparison, in terms of predictions between game harmony measure, the vacillation model, and other behavioural predictions about cooperation in games. It is worthy noticing that

some games, for example Game A (table 1 section 4) and Game D (table 4 section 4) have the same game harmony measure (in this case -0,8), but they are different in terms of reasoning about deviations, and therefore in terms of the vacillation model: in Game A I and we reasoning are observationally equivalent because both lead to the same solution, in Game D, instead, there will be a continuous switching between the I and we modes of reasoning, given that $\pi_I = \pi_W = 0,5$. At the same time, by slighlty changing the payoffs of Game A, harmony will change but not the way of reasoning. Let us see an example.

Table 6:

	L	R
U	*4,4*	3,1
D	1,3	*2,2*

The game in table 6 belongs to category A, but now the game harmony has become positive: it is 0,2.

These are only examples, but they show that there is space for a comparison in terms of behavioral predictions between my proposal and the game harmony measure, as well as other behavioral predictions, deriving from theories of social preferences, which do not deal with we-reasoning.

6 Conclusions

In this paper I have analysed Bacharach's theory of we-thinking. This is a very well developed formal theory of games with I-reasoning and We-reasoning, with the mode of reasoning taken as given. A fundamental feature of the theory is that the mode of reasoning is prior to rational choice. So, as Bacharach himself recognises, to complete the theory there has to be a model of which mode of reasoning is used by the agents. This part is less formal and less developed by Bacharach, although he proposes many intuitions and suggestions. In particular I have described two approaches Bacharach attempted to use: the harmony approach, developed by Zizzo and Tan, and the interdependence idea, which contains an underdeveloped intuition about vacillation between frames, and is only one way - I to We- in its formal presentation, but it seems naturally two way - I to We and We to I - as we see in the double crossing intuition.

I have proposed a way in which the 'double-crossing' intuition may be taken into account: reasoning about deviation from equilibrium, where equi-

librium is seen both from an I and from a We point of view. I presented a classification of games, based on reasons to deviate from an equilibrium (I or We), suggesting that an I-We equilibrium could be represented by the intersection between I and We equilibria. However, a game can present more than one I-we equilibrium and in some games the intersection can be empty, and this leaves space for vacillation. In order to determine which mode of reasoning is more likely to be chosen in these cases, I sketched a more formal model, based on temptation to deviate from an equilibrium and on refinement of equilibria which can be induced by Markov transition processes. This seems to me to be faithful to Bacharach's intuitions and, at the same time, to be a development of his theory, in a way that makes we-reasoning more easily usable in game theory.

References

[1] ANDERSON, E. (2001): "Unstrapping the straitjacket of 'preference': a comment on Amartya Sen's contribution to Philosophy and Economics", Economics and Philosophy, 17, 21-38.

[2] BACHARACH, M. (1993): "Variable Universe Games," in Frontiers of Game Theory, ed. by K. Binmore, A. Kirman, and P. Tani. Massachusetts: MIT Press.

[3] BACHARACH, M. (1995): "Co-Operating without Communicating," London.

[4] BACHARACH, M. (1997): ""We" Equilibria: A Variable Frame Theory of Cooperation," Oxford: Institute of Economics and Statistics, University of Oxford, 30.

[5] BACHARACH, M. (1999): "Interactive Team Reasoning: A Contribution to the Theory of Cooperation," Research in Economics, 53, 30.

[6] BACHARACH, M. (2001): "Framing and Cognition in Economics: The Bad News and the Good," ISER Workshop, Cognitive Processes in Economics.

[7] BACHARACH, M. (2006): Beyond Individual Choice. Princeton University Press, Edited by N. GOLD, and R. SUGDEN.

[8] BACHARACH, M., BERNASCONI, M. (1997): "The Variable Frame Theory of Focal Points: An Experimental Study", in Games and Economic Behavior, 19 (1), 1-45.

[9] BECCHETTI, L., DEGLI ANTONI G., and FAILLO, M. (2009): "Common reason to believe and framing effect in the team reasoning theory: an experimental approach", Econometica Working Paper series, n.15, November 2009.

[10] BRUNI, L. (2008): Reciprocity, altruism and the civil society. Routledge.

[11] DAVIS, J. (2009): "Two relational conceptions of individuals: teams and neuroeconomics" in A Research Annual, Research in the History of Economic Thought and Metodology series, vol. 27 – A, 1-21.

[12] GAUTHIER, D. (1975): "Coordination"in Dialogue, 14, 195-221.

[13] GILBERT, M. (1989): On Social Facts. Routledge.

[14] GOLD, N., and R. SUGDEN (2008): "Theories of Team Agency," in Rationality and Commitment, ed. by P. Di Fabienne, and S. H.: Oxford University Press.

[15] HODGSON, D. H. (1967): Consequences of Utilitarianism. Oxford: Clarendon Press.

[16] HOLLIS, M. (1998): Trust within Reason. Cambridge: Cambridge University Press.

[17] HOLLIS, M. and SUGDEN, R. (2003): "Rationality in Action", Mind, 102 (405), 1-35.

[18] HUME, D. ([1739] 1978): A Treatise of Human Nature. Oxford: Oxford University Press.

[19] HURLEY, S. (1989): Natural Reasons. Oxford: Oxford University Press.

[20] REGAN, D. (1980): Utilitarianism and Cooperation. Oxford: Clarendon Press.

[21] SALLY, D. (1995): "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiment from 1958 to 1992," Rationality and Society, 7, 58-92.

[22] SUGDEN, R. (1993): "Thinking as a Team: Toward an Explanation of Nonselfish Behavior," Social Philosophy and Policy, 10, 69-89.

[23] SUGDEN, R. (2000): "Team Preferences," Economics and Philosophy, 16, 175-204.

[24] SUGDEN, R. (2003): "The Logic of Team Reasoning," Philosophical explorations, 16, 165-181.

[25] SUGDEN, R.(2005): "Fellow-Feeling," in Economics and Social Interactions, ed. by B. Gui, and R. Sugden: Cambridge University Press.

[26] TAN, J., and D. ZIZZO (2008): "Groups, Cooperation and Conflict in Games," The Journal of Socio-Economics, 37, 1-17.

[27] TUOMELA, R. (1995): The Importance of Us: A Philosophical Study of Basic Social Notions. Stanford University Press.

[28] TUOMELA, R. (2007): The Philosophy of Social Practices: A Collective Acceptance View. Cambridge University Press.

[29] ZIZZO, D., and J. TAN (2003): "Game Harmony as a Predictor of Cooperation in 2 X 2 Games: An Experimental Study," Oxford: Department of Economics, University of Oxford.

[30] ZIZZO, D. (2004): "Positive Harmony Transformations and Equilibirum Selection in Two-Player Games," Oxford: Department of Economics, University of Oxford.