Non-binding agreements in dynamic games with complete information*

Emiliano Catonini†

July 22, 2011

Abstract. In dynamic games, players may observe a deviation from a pre-play, possibly incomplete, non-binding agreement before the game is over. The attempt to rationalize the deviation may lead the players to revise their beliefs about opponents' behaviour in the continuation of the game. Here I study the effects of such rationalization on the credibility of the agreements itself, that is on the possibility that it will be respected and on the beliefs it is able to induce.

Keywords: Agreements, dynamic games with complete information, epistemic game theory, strong-delta-rationalizability

1 Introduction

When the players of a dynamic game are given the opportunity to communicate among themselves before the game starts, they are likely to exploit this opportunity to take a (possibly incomplete) agreement about how to carry on the game. In most cases, the context allows them to take only a non-binding agreement, a "cheap talk" among the players that cannot be enforced by an external court of law. All along the paper, I refer exclusively to such non-binding agreements. Yet, there may be a wide multiplicity of reasonable agreements. For instance, most of the games feature a multiplicity of subgame perfect equilibria which are not pareto-dominated among themselves; moreover players may be willing to agree only on an equilibrium path, without specifying the off-the-path behaviour even though all possible contingencies are commonly known to the players. Among this great variety of possible agreements, without investigating the pre-play bargaining issue, the aim of this paper is to shed some light about what the credible alternatives are, that is, which agreements will be

*Preliminary Draft. I am grateful to Pierpaolo Battigalli for introducing me to the mysteries of epistemic game theory and for precious discussions about my work. Thank you also to all the attendants of the third-year PhD workshop in January.

†Bocconi University, PhD Candidate, emiliano.catonini@gmail.com

able to induce the belief that they will be respected and, among them, which ones will actually be, given such beliefs. To fix ideas and understand the main issues, Section 2 focuses on 2-players, finitely repeated games, first through the discussion of an example, then through the class of *equilibrium paths that can be upset by a convincing deviation* introduced by Osborne [11], for which the beliefs leading players to deviate from them are clarified, whereas Osborne leaves this condition unspecified. The ideas introduced in this section can be easily extended to different games. Therefore, Section 3 formalizes and deepens the analysis in the framework of dynamic games with complete information. First, the possible nature of an agreement is discussed, focusing the attention to complete agreements and agreements concerning just a path to be played, like the ones of the previous section. Second, the beliefs that an agreement may be able to induce and their interaction with rationality beliefs are formalized in the framework of strong-delta-rationalizability [2] that will be used for the analysis. Section 4 analyzes the main issues: when are agreements credible and will also be respected? Beside this concept of self-enforceability, a weaker concept of enforceability is introduced to indicate the possibility that the agreement is respected under weaker beliefs than a full belief in it. It is shown however that an enforceable agreement is also a self-enforceable agreement if enforceability is obtained without leaving off-the-path restrictions (hence for all agreements that don't restrict behaviour at information sets which may not be reached although the remainder of the agreement is respected). Sharper results about self-enforceability and enforceability can be claimed for the class of path agreements and they are collected together in the second paragraph of the section. Section 5 concludes and anticipates the next steps of the research project: the inversion of the epistemic priority between the beliefs in the agreement and the beliefs in rationality and the issues to be tackled in the much wider framework of dynamic games with incomplete information.

2 Agreements in two-players, finitely-repeated games

2.1 The leading example

Consider to repeat twice the following prisoner dilemma with a Punishment action:

$A\backslash B$	C	D	P
C	$5,5$	$2,6$	$0,0$
D	$6,2$	$3,3$	$0,0$
P	$0,0$	$0,0$	$1,1$

There exists a subgame perfect equilibrium where the two players cooperate

in the first stage. Namely, (s_1^*, s_2^*) where:

$$s_i^*(h) = \begin{cases} C \ if \ h = h_0 \\ D \ if \ h = (CC) \\ P \ else \end{cases} \quad i = 1, 2.$$

Suppose that Ann and Bob, before starting the game, meet together and agree on playing such SPE. That is, they agree to play its equilibrium path and to punish each other in case the path wasn't played in the first stage. Suppose that Ann and Bob trust each other and that there is common belief in this trust. That is, they believe the opponent will respect the agreement, they believe the opponent believes in the respect of the agreement, and so forth. Moreover, suppose that Ann and Bob are rational and that there is common belief in rationality. That is, they choose their actions maximizing expected payoff conditional on their beliefs about the opponent's moves, they believe the opponent chooses actions in the same way, and so forth. Can all these beliefs hold together for Ann after she observes a deviation by Bob in the first stage? The answer is no. Clearly, she cannot still think that Bob is rational and believes in her respect of the whole agreement. Which beliefs will she drop then? If she drops the belief that the Bob is rational, but she keeps the belief that Bob is respecting the remainder of the agreement (although he violated it so far), she will think that Bob is going to proceed with P, hence she will reply with P. If Bob anticipates this, he won't deviate from the agreement. If the same reasoning symmetrically holds, Ann will cooperate in the first stage too, and the agreement will be respected. Instead, if Ann keeps the belief that Bob is rational (in such a case I will say that rationality beliefs have the priority over agreement-induced beliefs) while she drops the belief that Bob believes that she's respecting the agreement, she has to wonder how Bob will proceed playing given the other beliefs which she hasn't had to drop. So why would a rational Bob deviate from the agreement in the first stage? Either because he believes that Ann won't respect the agreement in the first stage, or because he believes that Ann starts respecting the agreement but, by revising her beliefs after his deviation, won't necessarily respect it anymore. The second hypothesis is the more appealing. Let's decompose the belief in the respect of the agreement by the opponent in different beliefs about what she would do at different information sets. Then Ann, after observing Bob's deviation and not willing to drop the belief that Bob is rational, can drop only the belief in Bob's belief about what she would do after the deviation and not about what she would do along the equilibrium path (including the root of the game). More intuitevely, this may correspond to believing in the attitude of "respecting the agreement until the other does" and to a reputation context in which "Ann is a reliable player and has never deviated from an agreement first". The, how should Ann interpret the deviation of Bob under the light of Bob trusting her and being rational? The trust in Ann's respect of the agreement until Bob does amounts to let Bob believe that the agreement path would be respected if he does. The rationality of Bob, together with his trust, implies that Bob wants to gain more than under the agreement path through his deviation. As a consequence of this, he must choose D after

the deviation and not P. And Ann best replies with D. If Bob anticipates this, he would find it profitable to deviate, because he can be confident that his deviation will signal to Ann the desire of gaining more than under the path, hence his intention to play D in the last stage. Therefore the agreement wouldn't be respected.[1] Of course, exactly the same reasoning applies when Ann and Bob agree on the equilibrium path only, without even specifying the reaction to an opponent's deviation from the path, possibly because they are aware that they won't be convinced that the opponent respects the agreement after that the agreement has already been violated.[2]

I moved from the assumptions to this conclusion with an instance of forward induction reasoning which does not rely only on the beliefs in rationality but also on the beliefs about the respect of the agreement (namely of the agreement path) and their interaction. If we were to rely only on rationality beliefs, we wouldn't be able to exclude that the two players will play the equilibrium path. The two symmetric strategies which form this SPE are strongly rationalizable, as defined by Battigalli and Siniscalchi [5], because they survive the iterated elimination of weakly dominated strategies. They survive it because:

1. they are best replies one to each other;

2. all other (non realization-equivalent) best replies to them do strictly worse either against the opponent's strategy which prescribes playing P at every history or against the opponent's strategy which prescribes playing D in the first stage and P in the second stage regardless of what has been played in the first stage;

3. the latter two strategies survive because they are best replies to themselves and all other best replies do strictly worse either against the other strategy or against the SPE strategy.

This example shows three important facts. The first is the relevance of the beliefs about the respect of a path for the analysis of the strategic interaction which follows from a non-binding agreement. Both as the potentially believed part of a complete agreement and as whole agreements themselves, path agreements will be defined and put at the center of the focus. The second is the importance of the assumptions about players' beliefs in understanding the behavioural consequences of an agreement. For this reason, these assumptions and their implications will be investigated in detail in the general discussion. The third fact concerns the non-triviality of the implications of an agreement for the development of the game. An equilibrium path, although generated by SPE strongly rationalizable strategies, may surely be violated by a player if at

[1] Anticipating this, could Ann still play C in the first stage? Yes if she believes that if she also plays D, then Bob would punish with P, and this may apply because once the belief that Bob respects the agreement path has been dropped, Ann's deviation is not a clear signal that she wants to play D also in the last stage.

[2] In turn, Ann and Bob may reject also agreeing on the equilibrium path because they are aware that it wouldn't be believed by the reasoning above.

least the belief that the opponent would have respected it holds, although the punishments are not ruled out a-priori. These three issues (nature of the agreement, structure of induced beliefs, respect and credibility of the agreement) will be the three drivers of the general analysis carried on from the the next section. Before that, the next paragraph recovers a class of equilibrium paths analyzed by Osborne still in the context of 2-players finitely-repeated games and uses the epistemic approach to specify under what assumptions on beliefs their label is justified and evaluate them as agreements.

2.2 Equilibrium path that can be upset by a convincing deviation

Although not explicitely dealing with agreements, the closest contribution to the research questions raised in the previous paragraph (to the best of my knowledge) comes from Osborne [11]. Osborne focuses on 2-players finetely-repeated coordination games. In this framework, he defines a class of equilibrium paths which "can be upset by a convincing deviation", where a supposedly expected path is dismissed by one of the two players in such a way that the opponent understands to what other path the deviator is aiming to.

Definition 1 *[Osborne, [11]]. Suppose that the following condition holds for the pure Nash equilibrium outcome path $P = (\overline{a}^1, ..., \overline{a}^T)$ in the T-fold repetition of an arbitrary two-player strategic game. There is a deviation by player i in some period τ which generates the outcome $d^\tau \neq a^\tau$ in period τ, with the property that there is precisely one sequence of outcomes $(d^{\tau+1}, ..., d^T)$ in the remaining periods for which player i is at least as well off in $(d^\tau, ..., d^T)$ as he is in $(\overline{a}^\tau, ..., \overline{a}^T)$, and player i is in fact better off in $(d^\tau, ..., d^T)$ than he is in $(\overline{a}^\tau, ..., \overline{a}^T)$. Further, player j's payoff is higher when she adheres to the path $(d^{\tau+1}, ..., d^T)$ than when she deviates from this path,* whatever *sequence of outcomes her deviation induces. In this case I say that the path P can be* upset by a convincing deviation.

Osborne formalizes the condition under which an outcome path is upset by a convincing deviation, and asserts that such outcome path is not stable, in the sense of Kohlberg and Mertens [10]. Let b^i and c^i be the first- and second-ranked outcomes of G for player i $(= 1, 2)$.

Proposition 2 *[Osborne, [11]] Let $P = (\overline{a}^1, .., \overline{a}^T)$ be a pure Nash equilibrium outcome path of G^T. Suppose that there exist $\tau \in \{1, ..., T-1\}$, $i \in \{1, 2\}$, and $\widetilde{a}_i \in A_i$ such that*

$$u_i(\overline{a}^\tau \backslash \widetilde{a}_i) + u_i(c^i) + (T - \tau - 1)u_i(b^i) < \sum_{t=\tau}^{T} u_i(\overline{a}^t) < u_i(\overline{a}^\tau \backslash \widetilde{a}_i) + (T - \tau)u_i(b^i) \quad (1)$$

and

$$(T-\tau)u_j(b^i) > \max_{a_j} \left\{u_j(b^i \backslash a_j) : a_j \in A_j \text{ and } a_j \neq b_j^i\right\} + (T-\tau-1)u_j(b^j), \quad (2)$$

where $j \neq i$. Then the outcome path P is not stable.

In the analysis of Osborne it is understood that player i can confidently deviate from the equilibrium path and upset it because player j will interpret the deviation in the desired way[3]. This is an instance of forward induction reasoning, but the beliefs upon which it is grounded are not explicit in the analysis. Beliefs in rationality are not enough to induce this instance of forward induction reasoning. Player i might have played $\widetilde{a}_i$ at stage τ simply because she did not believe that j would have actually played $\overline{a}_j^\tau$ in τ. Then, it may be rational for player i to go back to the original path and player j may have good reasons to continue playing $\overline{a}_j^t$ $(t = \tau+1, ..., T)$ even after observing that player i played $\widetilde{a}_i$ at stage τ. Hence player i could not be sure that j would interpret the deviation in the desired way.[4]

The existence of further beliefs is necessary to justify the instance of forward induction reasoning which leads player i to upset the path. Natural candidate beliefs are those which concern the respect of the path on which the players have agreed. The following proposition shows precisely which these beliefs are.

Proposition 3 *Consider a path that can be upset by a convincing deviation by player i. Then, conditionally on history $(\overline{a}^1, ..., \overline{a}^{\tau-1})$, player i will actually deviate from the path if the following epistemic assumptions hold at $(\overline{a}^1, ..., \overline{a}^{\tau-1})$:*

1. *player i is rational.*
2. *player i believes that player j is rational.*
3. *player i believes that player j believes that player i is rational.*
4. *player i believes that player j would respect the path.*
5. *player i believes that player j believes that player i believes that player j would respect the path.*

Proof. Define $U : Z \to \mathbb{R}$ as $U((a^1, ..., a^T)) = \sum_{t=1}^{T} u(a^t)$. $\forall s_1 \in S_1$ *such that* $\exists s_2 : \zeta(s_1, s_2) = (\overline{a}^1, .., \overline{a}^\tau \backslash \widetilde{a}_1, a^{\tau+1}, ..., a^T)$ *for some* sequence $(a^{\tau+1}, ..., a^T) \in A^{T-\tau}$ *and* $\nexists s_2 \in S_2 : \zeta(s_1, s_2) = (\overline{a}^1, .., \overline{a}^\tau \backslash \widetilde{a}_1, b^i, ..., b^i)$ (i.e. all strategies of

[3] A similar idea of signaling future actions is in the work by Ben-Porat and Dekel[7], but there the forward induction reasoning which does the job is simply based on rationality beliefs and not on other beliefs which may arise from an agreement.

[4] The formal proof of this fact is that the strategy that prescribes to play at any given stage any given action which is featured by a Nash equilibrium of the stage game, regardless of what has been played in the previous stages, is a strongly rationalizable strategy.

player one deviating in τ but not aiming to her best prosecution afterwards), s_1 is incompatible with assumptions 1 and 4 jointly.

Hence, by assumptions 2, 3 and 5, player i cannot believe that player j will play any $s_2 \in S_2$ *such that* $\exists s_1 : \zeta(s_1, s_2) = (\overline{a}^1, .., \overline{a}^\tau \backslash \widetilde{a}_i, a^{\tau+1}, ..., a^T)$ *for some* seq *uence* $(a^{\tau+1}, ..., a^T) \in A^{T-\tau}$ *and* $\nexists s_1 \in S_1 : \zeta(s_1, s_2) = (\overline{a}^1, .., \overline{a}^\tau \backslash \widetilde{a}_i, b^i, ..., b^i)$.

Hence, by assumptions 1 and 4, player i will play a $s_1 \in S_1$ *such that* $\exists s_2 \in S_2 : \zeta(s_1, s_2) = (\overline{a}^1, .., \overline{a}^\tau \backslash \widetilde{a}_1, b^i, ..., b^i)$.

The proposition, other than explaining under which conditions on beliefs it is legitimate to talk of a "convincing deviation", provides a first answer to the question whether such an agreement would be respected. If players agree on a path that can be upset by a convincing deviation by player i and the agreement induces in player i the belief that player j will respect the agreement and the belief that player j believes this fact (other than the rationality beliefs), then player i will actually violate the agreement by upsetting the path when the convincing deviation is available. Notice that depending on player j's beliefs, it is possible that player j will not follow the path until the point where player i can implement the convincing deviation. Anyway, under the epistemic assumptions of proposition 3, the agreement will not be respected. Yet, on the other side, the path may be played by strategically sophisticated players if some assumption of proposition 3 fails to hold, since the path can be generated by strongly rationalizable strategies. Moreover, a belief of the opponent that the deviatior would have instead respected the path is clearly at odds with the set of assumptions of proposition 3, and although it seems more reasonable to drop such belief, one could also think that some other belief assumed by proposition 3 was not generated by the agreement. Does it open up the possibility that the agreement can induce a set of beliefs under which the agreement will be respected for sure?

The aim of the prosecution of is paper is to answer such questions beyond the narrow framework investigated so far. The concept of equilibrium path that can be upset by a convincing deviation can be easily extended to any dynamic game. For instance, the equilibrium path of the leading example is a kind of path that can be uspset by a convincing deviation by extending the definition in the most natural way to any finitely repeated game. Moreover, an intermediate concept of deviation, after which the coordination of players is possible but not necessary like here, could also be introduced to deal with agreements which are able to induce all the beliefs in their respect, but under these beliefs the respect of the agreement is unsure, because it depends on whether the potential deviator has sharp beliefs about opponent's reaction or not. But instead of looking for other peculiar situations and agreements, a more general approach will be taken, to develop a methodology which allows to evaluate agreements in any dynamic game with complete information.

3 Agreements and beliefs in dynamic games with complete information

3.1 Agreements, complete agreements and path agreements.

When the players of a game are given the opportunity to communicate among themselves before the game starts, they will probably exploit this situation to coordinate their moves in the game. The process may involve a relevant bargaining component when, as it is usually the case, some alternative is more favourable to some player and some other alternative is more favourable to some other player. Here I don't investigate the bargaining process; I consider instead any alternative that the players may finally choose as their final agreement and investigate its possible implications for the development of the game. Formally, I expect players to come up with an *agreement* of the following form.

Definition 4 (Agreement) *An agreement is a profile of correspondences* $(f_i)_{i \in N}$ *where* $\forall i \in N$, $f_i : H \to 2^{A_i(h)}$.

That is, an agreement specifies pure actions among which players are expected to choose at the information sets where they are called to act (for the sake of simplicity, the agreement also specifies the "dummy" action at the information sets where the player is not active). If for a player the agreement assigns to an information set the whole set of available actions, $A_i(h)$, it means that the agreement is silent about how the player should play at that information set.

Of course, players may anticipate the implications of an agreement they are going to take, and this will have a feedback effect on the evaluation of the candidate agreement. As a starting point, however, it seems quite natural to take into consideration as a basis for the agreement the *equilibria* of the game, according to some notion of equilibrium. Whatever notion of equilibrium we are considering, equilibria are natural candidates for an agreement because rational players have the incentive to respect it if the opponents do. In static games, once a suitable equilibrium has been identified by the players, all they can do is taking the agreement on it. In dynamic games, the agreement can incorporate an equilibrium to different extents. One very natural possibility is to agree just on what to do from the start to an end of the game, understood that the agreement has been followed so far. For instance, in repeated games it may make a lot of sense, and remind of many real life situations, to agree on what to do in the stage-game but make clear that the agreement doesn't apply anymore after that it has been broken up by someone. This is the situation depicted in the previous section, both in the example and in the discussion of Osborne's class of paths. Or, in any dynamic game, people may just agree on the outcome of the game they want to achieve. This has a univocal implication for the moves that players are expected to perform as long as the expectations about opponents' moves are fulfilled, while it has no implication about how the players should react to a deviation. I will call such an agreement a *path agreement.* Formally:

Definition 5 (Path Agreement) *A path agreement is an agreement where* $\exists z = (a^0, a^1, ...) \in Z$ *such that* $\forall i \in N$, $\forall h \in H : h \prec z$, $f_i(h) = a_i^{lenght(h)}$ *and* $\forall i \in N$, $\forall h \in H : h \not\prec z$, $f_i(h) = A_i(h)$.

Yet, even if players agree on playing a certain equilibrium path and one believes that the opponents are going to respect the agreement, she may find it profitable to deviate from the path anyway, depending on what she believes about the opponents' reaction to her deviation. A way to overcome this problem may be discussing anticipately also about off-the-path behaviour and agree on a whole subgame perfect equilibrium. More generally, I call *complete agreement* an agreement which prescribes to the players what to do exactly in every contingency.

Definition 6 (Complete Agreement) *A complete agreement is an agreement where* $\forall i \in N$, $\forall h \in H$, $|f_i(h)| = 1$.

In a subgame perfect equilibrium, what is prescribed off-the-path prevents players from deviating from the path. But this poses two types of problems. First, players may not be willing to discuss how to act in case someone violates the agreement. Undertaking such discussion may signal a lack of trust in the opponents or the mere fact of taking into consideration the idea that the agreement may be violated could be perceived as an undesirable way to start. Moreover, players may reject to anticipate any coordination with opponents who have already proved being untrustworthy with a deviation from an agreement in a past game. Furthermore, players may find it too costly to discuss what to do in all possible contingencies, that is, all possible violations of the agreement and subsequent prosecutions of the game, although throughout all this paper I assume that players have common knowledge of the whole game. These observations justify the interest for path agreements. Second, even if players do take a complete agreement on how to play, this does not necessarily imply that player will believe that the agreement will be respected by the opponents at every information set of the game. As observed in the previous section, a player who observes a deviation from a SPE complete agreement may find it more reasonable to drop the belief that the deviator believes in the agreement rather than the belief that the opponent is rational. Instead of inverting the epistemic priority between rationality and agreement believes (an issue discussed in the concluding section), one can assume that the deviator does not believe in the prescribed reaction to the deviation. This amounts to let players believe in the agreement only along the path: in this case the analysis of path agreements performed in the devoted section applies also to incomplete agreements whenever they are not able to induce further beliefs beyond the ones regarding the path.

However, before restricting the focus again on path agreements, any conceviable agreement will be taken into consideration for the definition of the beliefs that an agreement may be able to induce and for the investigation of whether an agreement will be believed and respected.

3.2 Beliefs induced by an agreement and their interaction with rationality beliefs

The only way a non-binding agreement can affect the behavior of players is through the beliefs it is able to induce in their minds. When taking a non-binding agreement, one is usually willing to trust that the opponents are going to respect it. In static games, if the players agree upon a given equilibrium, it is quite natural to believe that the opponents are going to respect the agreement and hence, accordingly with own rationality, respect it too. In dynamic games, instead, the issue may become very subtle. Players can observe the violation of an agreement before the game is over and this makes them revise their beliefs. Anticipating this, violating an agreement can become a promising strategy if the deviation conveys a useful signal to the opponents. Hence, in dynamic games, the simple belief in the agreement plus the rationality of the players (i.e. expected utility maximization given those beliefs), although it prescribes to play part of an equilibrium (for instance, the path), is not sufficient to induce players to respect the agreement itself. To see what the behavioral implications of the agreement are, a much deeper analysis of the beliefs that the agreement is able to induce is required.

The formalization of the concepts that follow will be done in the same framework of Battigalli and Prestipino [4]. The framework gives a representation of beliefs and strategies of players in a state-space which is suited for dynamic games with incomplete information. Although the incomplete information dimension is dropped here, this framework has been chosen for reference in future work, as the last section proposes.

First of all, the agreement may be able to induce in the players the *first-order belief* that the opponents are going to respect it. Here I assume that such belief holds at every information set of the game; of course the relevant part of these beliefs concerns the part of the agreement which regards the continuation of the game. This simply amounts to assuming that players don't change their their beliefs about the opponents behavior in the continuation of the game while the game unfolds. This rules out the attitude of believing in the respect of an off-the-path portion of the agreement until the path is respected and forming doubts about it once the path has actually been violated, which could be of interest from a psychological point of view. In this framework, the first-order belief of a player in the respect of the agreement by the opponents is represented by a restricting the set of Conditional Probability Systems [13](CPS) among which the player can choose her conjecture with respect to all conceivable ones:

Definition 7 *Consider an agreement $a = (f_i)_{i \in N}$. I call the cartesian set of conditional probability systems corresponding to the agreement (or the first-order*

beliefs restrictions corresponding to the agreement) $\Delta^a := \prod_{i=1}^{N} \Delta_i^a$ *where:*

$$\forall i \in N,\ \forall h \in H,\ S_i^{a,h} := \left\{ s_i \in S_i(h) : \forall \widehat{h} \gtrsim h,\ s_i(\widehat{h}) \in f_i(\widehat{h}) \right\} \text{ and}$$
$$\forall i \in N,\ \Delta_i^a = \left\{ \mu_i \in \Delta^{H_i}(S_{-i}) : \forall h \in H_i,\ \text{supp}\mu_i(h) \subseteq S_{-i}^{a,h} \right\}.$$

But first-order beliefs may not be enough to explain the strategic behaviour of the players after an agreement. For more sophisticated strategic reasonings, higher-order beliefs must be taken into consideration. In this paper, I assume that such higher-order beliefs, when they are induced by the agreement, they all reflect the beliefs of lower order. That is, I assume that the second-order beliefs of player j about the first-order beliefs of player i put probability one on the first-order beliefs among which the latter actually chooses (Δ_i^a). And the same holds for the n-order beliefs of a player about the $n-1$-order beliefs of an opponent, up to any order of beliefs. This assumption is called *transparency* of the first-order beliefs restrictions. It means that players hold the restrictions, believe that the restrictions hold, believe that they believe that the restrictions hold, and so on. Differently from the assumption that players don't change their mind throughout the game, which is psycologically binding but of minor interest here, the assumption that "players believe the true beliefs" is quite demanding and doesn't allow to explain some violations of agreements in reality. In the leading example, for instance, if Ann and Bob agree on the whole subgame perfect equilibrium but Ann believes in the transparency of the belief in the whole agreement while Bob believes in the transparency of the belief in the equilibrium path, Bob is going to deviate from the path, feeling sure to signal the intention to defeat again in the second stage, but Ann won't understand it and will play the punishment. Nevertheless, in many situations it is reasonable to think that when the first-order belief restrictions corresponding to the agreement can hold (why they couldn't is explained later), players will be aware they actually hold. Hence the trasparency of Δ^a applies and in this framework is defined by taking the portion of the state-space where it holds, which is formally indicated as follows:

$B^*([\Delta^a]) := \bigcap_{n \geq 0} B^n([\Delta^a])$, where $B(E) = \prod_{i=1}^{N} B_i(E)$ and $[\Delta^a]$ is the subset of the state-space where each player i holds restrictions Δ_i^a.

Yet, like the first-order restrictions have no bite without the rationality assumption, higher-order restrictions have no bite without the belief in rationality and higher-order beliefs in rationality. In the analysis of equilibrium paths that can be upset by a convincing deviation, for instance, the belief of one player that the deviator believes that she would respect the path wouldn't ensure that the she interprets the deviation as a way to gain more than the agreement payoff if she wouldn't believe in the rationality of the deviator. Therefore, the interaction of beliefs in the agreement and beliefs in rationality is the crucial point of the analysis of how agreements affect the behavior of players. In order to investigate deeper into this issue, we must first make precise what kind of beliefs in

rationality we do take in dynamic games. We think that the most appropriate notion of belief in rationality for dynamic games is the *strong belief* (SB) one. A player strongly believes in the rationality of the opponent if she believes that the opponent is playing some rational strategy of hers (i.e. a strategy which is a best reply to some conjecture about opponents' moves) at every information set that can be reached by some rational strategy of hers, while she believes in some other strategy at the other information sets. Here, I consider the strong belief in jointly rationality and the transparency of first-order beliefs restrictions induced by the agreement. This implies that a player, at an information set that cannot be reached by a rational opponent who holds the first-order belief restrictions, can believe that the latter is playing any other strategy which is compatible with her own first-order belief restrictions, and not necessarily a rational one (but not best reply to any conjecture in delta). Notice here the incompatibility that can arise among the beliefs that I have already set down. A rational and believing in the agreement opponent may choose strategies which are out of the agreed ones. Hence the first-order beliefs restrictions of a player can be at odds with her belief in rationality and belief restrictions of the opponent. In this case, the set of first-order-belief restrictions corresponding to the agreement cannot be transparently believed by players who are rational and strongly believe in rationality and transparency of the restrictions. If such an agreement is taken anyway, the problem of which beliefs the agreement actually induces is open. As it will be discussed in the next section, it's hard to think of the transparency of restrictions which have not been openly discussed by the players, but among the various possibilities of looser-than-the-agreement restrictions, a mediator or the players themselves may recognize the need to point them out before playing. Higher order beliefs concerning rationality and the transparency of the first-order beliefs restrictions are built on the conjunction of all the lower-order ones. For instance, I don't assume that players just strongly believe in the strong belief in rationality and in the transparency of the restrictions, but I assume that players strongly believe in the conjuction of the strong belief in rationality and in the transparency of the restrictions with rationality and the transparency of the restrictions themselves. Although counterintuitive, there are very strong reasons for this choice. First, a strong believe in an event which is a mere belief and does not include rationality will hold at every information set and therefore the "strong" definition loses any meaning. A belief is an epistemic event and as such can never be falsified by the observation; an event which includes rationality instead can be falsified by the observation because not all information sets are compatible with it (the projection of the event on the space of strategies is not the whole set of strategy profiles). Second, it is easy to check that even without considering the existence of restrictions, in a game there could be no element of the state-space which is coherent, for instance, with strong belief in rationality and strong belief in strong belief in rationality. At some information set the former may select couples of first-order beliefs and second-order beliefs providing the rational grounding for the first which are absent from the couples of second-order beliefs and third-order beliefs selected by the latter, because the second-order believed strategies may be irrational for the corresponding

player[5]. But there is no clear reason why people shouldn't be able to believe in rationality in some sense up to the third order, therefore this possible incompatibility is something to be avoided, and it is avoided posing the higher-order strong beliefs on the lower-order ones: in the reasoning before, a strong belief on the conjunction of rationality and strong belief in rationality would be silent at that information set, hence the needed second-order belief would be free to hold. Third, the conjuction of $n-$order strong beliefs on the conjuction of the previous $n-1$, is shown by Battigalli and Prestipino [4] to have behavioural implications that are equivalent to the strong-delta-rationalizability procedure introduced by Battigalli [2], the augmentation of the strong-rationalizability solution concept introduced by Battigalli and Siniscalchi [5] as a form of rationalizability for extensive form games.[6] The strong-delta-rationalizability solution concept will be indeed used to investigate the behavioural implications of the beliefs induced by the agreement in the next section of the paper; let me first introduce the structure of beliefs discussed above as again a subset of the state-space which is formally indicated as follows:

$$CSB^{\infty}((R\bigcap B^{*}([\Delta^{a}])) := \bigcap_{n\geq 0} CSB^{n}(R\bigcap B^{*}([\Delta^{a}]))^{7} where CSB(E) = E\bigcap SB(E).$$

The projection of this event on the sub-space of strategies is actually the epistemic characterization of the strong-delta-rationalizability solution concept. The ultimate definition of the strong-delta-rationalizability procedure is the following:

DEFINITION OF STRONG-DELTA-RATIONALIZABILITY AND CARACTERIZATION.

4 Credibility and respect of the agreements

4.1 Self-enforceability and enforceability

Once the players have taken some kind of agreement, to which it corresponds a tentative set of beliefs, it has to be investigated whether such beliefs can truly

[5] Consider a dynamic game where Ann chooses first between an outside option granting her 2 and leaving the move to Bob. If Bob is called to move, he can choose between action D that gives 1 to both and action S which gives 0 to him and 3 to Ann. If Bob strongly believes that Ann is rational, at the information set when he's called to play he should believe that Ann believes that he would play S. If Bob strongly believes that Ann strongly believes that Bob is rational, he should believe that Ann believes that he would play D. The two strong beliefs therefore cannot hold together.

[6] the first attempt in Pearce ([12], 1984), epistemically analyzed by Battigalli ([1], 1997).

[7] Battigalli and Prestipino [4] show that $Proj_{S}CSB^{n}((R\bigcap B^{*}([\Delta])) = Proj_{S}CSB^{n}(R\bigcap[\Delta^{a}])$, where the latter is the epistemic characterization of strong-delta-rationalizability given by Battigalli and Siniscalchi [6]. That is, transparent restrictions and restrictions having the same epistemic priority as rationality have the same behavioural implications, precisely because restrictions to higher-order beliefs than rationality ones have no bite.

be induced in the sophisticated players and, if yes, whether such beliefs lead players to respect the agreement. Of course, the most desirable implication of an agreement is the fact that players will surely play as the agreement prescribes. In case the Δ^a restrictions cannot hold (i.e. the agreement is not credible), some other belief restrictions will be induced by the agreement. Later it will be shown that even for a non credible agreement, some restrictions looser than Δ^a, if made transparent, might lead to the respect of the agreement. Yet, it is hard to assume that restrictions different than Δ^a will ever become transparent to players by the mere fact that Δ^a has failed to hold. For this reason, it is important that the agreement is respected under the Δ^a restrictions, and only in this case I will claim that the agreement is "self-enforceable". Thus, I define my concept of *self-enforceabilty* of the agreement accordingly.

Definition 8 (self-enforceability) *Consider an agreement $g = (f_i)_{i \in N}$. The agreement is self-enforceable if:*

1. $CSB^{\infty}(R \bigcap B^*([\Delta^a])) \neq \emptyset$;
2. $\forall i \in N, \forall h \in H(\text{Proj}_S CSB^{\infty}(R \bigcap B^*([\Delta^a]))), \forall s_i \in \text{Proj}_{S_i} CSB_i^{\infty}(R \bigcap B^*([\Delta^a])) \bigcap S_i(h)$, $s_i(h) \in f_i(h)$;

Point 1 requires that the agreement is transparently believed by the players. When point 1 holds but point 2 not, I will say that the agreement is "believable". Point 2 requires explicitely that *every* strongly-delta-rationalizable strategy respects the agreement at the information sets which are compatible with it and with the strongly-delta-rationalizable strategy profiles. That is, players are going to respect the agreement for sure at every node of the game where they are called to play. Concerning information sets which are not compatible with any strongly-delta-rationalizable strategy profile, instead, point 1 only implies that among the highest-degree rationalizable strategies which are compatible with an information set, there is at least one which respects the agreement. But this is enough to guarantee that the off-the-path beliefs of the players are in line with the agreement, so that the desired behavioural consequences apply. It is reasonable to think that the agreement has obtained its goal only if players believe it and contemplate it would be played in case that information set would be reached.

Analogous concepts could be introduced by replacing in the definition above the "for every" quantifier at point 2 with a "there exists" one. But an agreement which can only be believed and possibly played is not considered here to be an interesting one; every strongly-rationalizable profile has this characteristics without imposing any restriction, hence without taking any agreement. Therefore I proceed with the analysis focusing on self-enforceability.

Examples of self-enforceable agreements are complete agreements corresponding to a subgame perfect equilibrium in a dynamic game with observable actions and no relevant ties[8]. As observed by Battigalli and Friedenberg [3] for a

[8] We conjecture that by extending the definition of EFBRS to N-players games in the natural way, the following results keep on holding.

2-players case, in this kind of games the realization equivalence class of a subgame perfect equilibrium is an Extensive Form Best Response Set, a solution concept introduced in the same paper. Then Battigalli and Friedenberg [3] show that an EFBRS can always be induced by strong-delta-rationalizability, where the restrictions correspond for each player to the set of CPS which sustain the strategies in the EFBRS.[9] But this is not enough to claim self-enforceability of an agreement corresponding to a EFBRS, hence of the SPE, because the restrictions may go beyond the EFBRS itself. Instead, if the EFBRS contains all the strategies which are best replies to some conjecture which strongly believes the EFBRS itself, this amounts to claim self-enforceability of the EFBRS, because then the restrictions which deliver the EFBRS correspond to the restrictions induced by the agreement (Δ^a) and because the EFBRS definition already requires that if a sequential best reply to a conjecture which strongly believes the EFBRS is in the EFBRS, then also any other sequential best reply to the same conjecture is in it (the "maximality" requirement). I first define this stricter version of EFBRS, then I show that an agreement corresponding to such EFBRS is self-enforceable; moreover I show that although they can't be mapped into a full EFBRS, agreements corresponding to subgame perfect equilibria in dynamic games with observable actions and no relevant ties are self-enforceable anyway.

Definition 9 (full EFBRS) *A full EFBRS Q is an EFBRS (Battigalli and Friedenberg [3]) where $\forall i \in N$, $\forall \mu_i$ which strongly believes Q_{-i}, $r_i(\mu_i) \in Q_i$.*

Now we show the self-enforceability of full EFBRS.

Proposition 10 *Consider an agreement $g = (f_i)_{i \in N}$. If $Q = \prod_{i=1}^{N} S_i^a$, where $S_i^a := \left\{ s_i \in S_i : \forall \widehat{h} \in H,\ s_i(\widehat{h}) \in f_i(\widehat{h}) \right\}$, is a full EFBRS, then the agreement is self-enforceable.*

Proof: By definition, $\forall i \in N$, Δ_i^a is the set of all and only the CPS which strongly believe Q_{-i}. Since Q is a full EFBRS, then $\forall i \in N$, $S^1_{i,\Delta^a} = Q_i$. But then the set of CPS which strongly believe S^1_{-i,Δ^a} coincides with Δ_i^a; hence $\forall i \in N$, $S^2_{i,\Delta^a} = Q_i$. Hence, by induction, $S^\infty_{i,\Delta^a} = Q_i$. Recalling that $S^\infty_{i,\Delta^a} = Proj_{S_i} CSB^\infty(R \bigcap B^*([\Delta^a]))$, we can conclude that the agreement corresponding to Q is self-enforceable.■

Now we show that a subgame perfect equilibrium is self-enforceable.

[9] Thus, EFBRS represent strategy profiles which can be induced by *some* agreements, but not necessarily by agreements corresponding to them. The analysis of agreements performed in this paper doesn't assume that players first think of an EFBRS than to the agreement that will induce it because first this reverse problem may be extremely difficult to solve, second many interesting and natural agreements cannot be generated in this way.

Proposition 11 *In a game with observable actions and no relevant ties, any agreement which prescribes the actions of a subgame perfect equilibrium s is self-enforceable.*

Proof: Consider the Δ^a restrictions corresponding to the agreement. By observable actions, for every information set h we can define the subgame $\Gamma(h)$. Then by subgame perfection and no relevant ties, for every player the sequential best reply to s_{-i} after each information set is $s_i|h$. Hence $\forall i \in N$, $\forall \mu_i \in \Delta_i^a$, $r_i(\mu_i) = [s_i]$. Thus $\forall i \in N$, $S^1_{i,\Delta^a} = [s_i]$, hence for every player the set of CPS which strongly believe S^1_{-i,Δ^a} is a superset of Δ_i^a, which implies $S^2_{i,\Delta^a} = [s_i]$. Hence, by induction, $S^\infty_{i,\Delta^a} = [s_i]$. Recalling that $S^\infty_{i,\Delta^a} = Proj_{S_i} CSB^\infty(R \bigcap B^*([\Delta^a]))$, we can conclude that the agreement corresponding to s is self-enforceable.■

On the other hand, not just agreements corresponding to some subgame perfect equilibrium behaviour, even eventually incomplete, can be self-enforceable. Battigalli and Friedenberg [3] show that EFBRS not corresponding to any subgame perfect equilibrium behaviour can be generated by strong-delta-rationalizability for some restrictions: the agreement corresponding to those restrictions is then a self-enforceable one.

In there any hope that an agreement which can't be believed or which won't be surely respected when believed may be surely respected by making transparent some looser restrictions? Intuition would suggest no because of a simple but wrong countepositive argument. If an agreement is surely respected under some restrictions, it means that there are conjectures which believe the agreement and support strategies which respect the agreement. Then it would seem possible to restrict beliefs to match the agreement from the start and, consequently, respect it for sure. This naive view is proved to be wrong by two examples below. Before showing them, I formalize what I mean by respecting an agreement under looser restrictions. Agreements to which the concept applies are called "enforceable" rather than "self-enforceable" because the needed restrictions will be hardly made transparent by the agreement alone. But still, the possibility that an agreement is respected under some restrictions, which are less demanding than the ones corresponding to the agreement, is of great interest from an "agreement design" point of view. In such case, the players or a mediator who analyzes the game can suggest to take a looser "instrumental agreement" to obtain the more specific goal of the original agreement. For this reason I call the original agreements "enforceable":

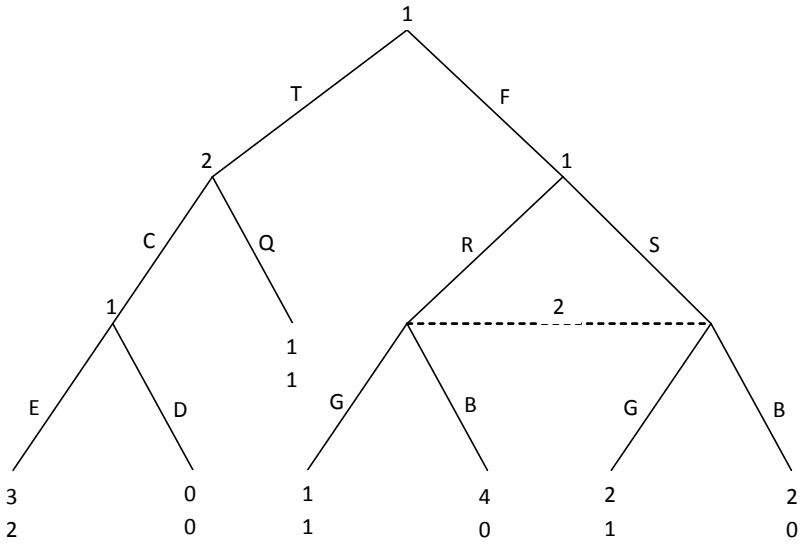
Definition 12 (enforceability) *Consider an agreement $g = (f_i)_{i \in N}$. The agreement is enforceable if $\exists \Delta \supseteq \Delta^a$ such that:*

1. $CSB^\infty(R \bigcap B^*([\Delta])) \neq \emptyset$;
2. $\forall i \in N, \forall h \in H(\text{Proj}_S CSB^\infty(R \bigcap B^*([\Delta]))), \forall s_i \in \text{Proj}_{S_i} CSB_i^\infty(R \bigcap B^*([\Delta])) \bigcap S_i(h)$, $s_i(h) \in f_i(h)$;

3. $\forall i \in N$, $\forall h \in H/H(\text{Proj}_S CSB^{\infty}(R \bigcap B^*([\Delta])))$, *defining*
$m := \max\left\{n \in \mathbb{N} \bigcup \{\infty\} : \text{Proj}_{S_i} CSB_i^n(R \bigcap B^*([\Delta])) \bigcap S_i(h) \neq \emptyset\right\}$,
$\exists s_i \in \text{Proj}_{S_i} CSB_i^m(R \bigcap B^*([\Delta])) \bigcap S_i(h)$ *such that* $s_i(h) \in f_i(h)$
and $\nexists s_i \in \text{Proj}_{S_i} CSB_i^m(R \bigcap B^*([\Delta])) \bigcap S_i(h)$ *such that* $s_i(h) \notin f_i(h)$
and such that $\exists j \neq i$, $\exists \mu \in \Delta_j$, $\exists \widetilde{h} \precsim h$ *such that* $s_i \in \text{supp}\mu(\widetilde{h})$

Notice the differences with the self-enforceability definition. The obvious one concerns the fact that the respect of the 3 requirements is asked not for the restrictions corresponding to the agreement but for looser restrictions. The less obvious one is in the further requirement at point 3. Since I want the players to believe in the agreement also off-the-path, so that the desired behavioural consequences apply, the surviving strategies must not be compatible with the delta restrictions when they are not compatible also with the agreement (which is weakly stricter).

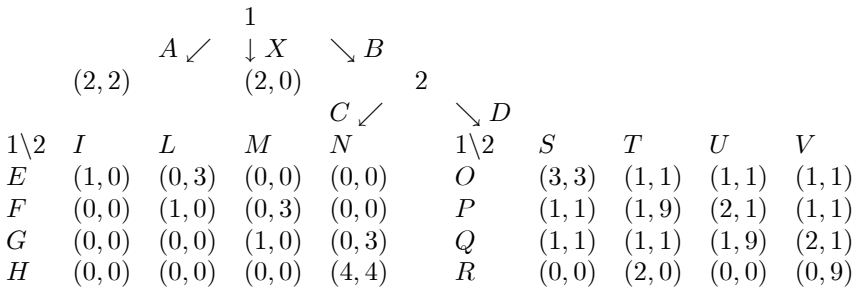
The first example provides a case of an agreement which is not self-enforceable because it cannot be fully believed by sophisticated players, but is actually respected by just removing one restriction corresponding to the agreement.



Suppose players agree that player 2 should play C and that player 1 should play S in case the respective nodes are reached. Players cannot transparently believe in the agreement and hold a strong belief in rationality and the first-order-beliefs restrictions corresponding to the agreement. Indeed, player 2 cannot believe that player 1 will play S after F if she strongly believes that player 1 is rational and that believes she will play C after T. Nevertheless, notice what

happens if we drop the belief of player 1 that player 2 will play C after T. We obtain anyway that player 2 will play C as agreed and we don't run anymore into contradiction. Moreover the "most rationalizable" strategy of player 1 which is compatible with history F prescribes to play S.

The next example shows that when an agreement is believable but not self-enforceable, it cannot be excluded anyway that it is enforceable for some looser restrictions.



1\2	I	L	M	N
E	$(1,0)$	$(0,3)$	$(0,0)$	$(0,0)$
F	$(0,0)$	$(1,0)$	$(0,3)$	$(0,0)$
G	$(0,0)$	$(0,0)$	$(1,0)$	$(0,3)$
H	$(0,0)$	$(0,0)$	$(0,0)$	$(4,4)$

1\2	S	T	U	V
O	$(3,3)$	$(1,1)$	$(1,1)$	$(1,1)$
P	$(1,1)$	$(1,9)$	$(2,1)$	$(1,1)$
Q	$(1,1)$	$(1,1)$	$(1,9)$	$(2,1)$
R	$(0,0)$	$(2,0)$	$(0,0)$	$(0,9)$

Consider the following agreement: the players agree to play in such a way that both get a strictly positive payoff. That is, 1 shouldn't play X and should play H in the left subgame and not play R in the right subgame; 2 should play N in the left subgame and anything in the right subgame. If we implement all restrictions corresponding to the agreement, the strong-delta-rationalizable procedure delivers X or A, meaning that the agreement is transparently believed (we don't have the empty set), but not self-enforceable, because X should be excluded[10]. The reason is that by the belief in H, player 2 excludes to play S, which makes player 1 exclude to play O and, in turn, player 2 to play C.[11] If instead we keep only the restriction that player 1 shouldn't play R, we obtain as unique strongly-delta-rationalizable path $(B, C, (H, N))$.[12]. The CPS

[10] The existence of the tie between A and X is for the sake of simplicity but it could be avoided putting after X a $2x2$ game where player 1 could get more or less than 3. Hence the assumption of no relevant ties doesn't prevent the existence of counterexamples like this one.

[11] The complete deletion procedure of reduced-form strategies goes as follows: at the first round, all strategies of player 1 with E, F and G are eliminated, while CI, CL, CM, DV and DS are eliminated for player 2; at the second round, all strategies of player 1 with O and with Q are eliminated, and none for player 2; at the third round no strategy is eliminated for player 1, while strategies CN and DU are eliminated for player 2 (only DT is left); at the fourth round all strategies with B are eliminated for player 1.

[12] The deletion procedure of reduced-form strategies goes as follows. In the first round, no strategy of player 1 is eliminated, while DV and CI are eliminated for player 2; in the second round, all strategies with E and all the strategies with Q are eliminated for player 1, and none for player 2; in the third round, no strategy is eliminated for player 1, while DU and CL are eliminated for player 2; in the fourth round, all strategies with P and all the strategies with F are eliminated for player 1, none for player 2; in the fifth round, no strategy is eliminated for player 1, while strategies DT and CM are eliminated for player 2; in the sixth round strategy A, X and all strategies with G and R are eliminated for player 1 (only BHO is left), none for player 2. In the seventh round player 2 decides playing CN. The agreement is enforced.

compatible with these restrictions are a superset of those compatible with the restrictions analyzed before. Yet, the behavioral implications are totally different and, more surprisingly, the looser restrictions lead to the respect of the agreement while the stricter ones don't.

One could argue that this problem is due to a badly-designed agreement, referring in particular to the exclusion that player 1 wouldn't play R, which in the subgame is a best reply to another action that player 2 may actually choose. However it is very difficult to find a meaningful condition on agreements such that this problem is avoided. Ex-post conditions could be established on the most rationalizable substrategies in the off-the-path subgames, but this would complicate heavily the picture[13]. Therefore, I prefer to claim that enforceability implies self-enforceability whenever under the Δ restrictions which deliver enforceability, no restriction is left off-the-path. Formally:

Theorem 13 *Consider an agreement $g = (f_i)_{i \in N}$. If the agreement can be enforced by delta restrictions Δ^E and $\forall h \notin H(Proj_S CSB^\infty(R \cap B^*([\Delta^E])), \forall i \in N, f_i(h) = A_i(h)$ then $\forall \Delta \subseteq \Delta^E, \Delta \supseteq \Delta^a, \{z \in Z : \exists s \in Proj_S CSB^\infty(R \cap B^*([\Delta])),\ \zeta(s) = z\} = \left\{z \in Z : \exists s \in Proj_S CSB^\infty(R \cap B^*([\Delta^E])),\ \zeta(s) = z\right\}$.*

Proof: see Appendix.

The theorem actually says two further things other than the self-enforceability of such enforceable agreements. The first regards delta restrictions between the ones which deliver enforceability and the ones corresponding to the agreement. Also for them the result holds. The other regards more precisely the behavioural implications of the restrictions. They are exactly the same under all the restrictions which enforce the agreement, including the ones corresponding to the agreement. Notice that the theorem does not require that for a delta lying between the one which satisfies the definition of enforceability and the one corresponding to the whole agreement the strongly-delta-rationalizable strategy profiles respect the agreement also at information sets which are excluded by the players. Yet, as long as the behavioral implications of the off-the-path beliefs are totally the same, it loses any importance whether those beliefs are fully in line with the agreement or not. Notice moreover that in the example the theorem does not apply because the restriction the player 1 shoud play R in the right subgame is left off-the-path with respect to the sole strongly-delta-rationalizable path delivered by the Δ^E restrictions which delivered enforceability.

The importance of the result is better understood through its negative formulation. If for believable but not self-enforceable agreements the agreement is not respected, then there is no hope to obtain the respect of the agreement by loosing the restrictions, unless leaving part of the agreement ending-up off-the-path. For agreements which don't restrict behaviour at information sets

[13] These conditions can be expressed as the non-emptiness of strategy profiles which derive from running strong-delta-rationalizability in the subgames not reached under Δ and Δ^a, with the further restriction to believe in the most rationalizable substrategy profiles obtained under Δ and Δ^a. This procedure puts under further stress the off-the-path restrictions and avoids that an action like R in the counterexample is restricted off.

which may not be reached even when the agreement is respected at the preceding information sets, this is kind of "revelation principle" result for agreements design. If players (or a mediator between them) wants to obtain that behaviour (but they are not willing to agree further than that) they can't do better than making the whole agreement transparently believed. Unfortunately, this is a narrow class of agreements. Yet, this class includes path agreements, for which sharper predictions can be made and are the object of the next paragraph.

4.2 Self-enforceability of path agreements

The importance of path agreements has already been pointed out throughout the paper. Here I apply the concepts defined in the previous paragraph to path agreements, in order to understand their desirability as credible and effective agreements.

I start with an important claim which excludes the mere enforeceability of path agreements.

Proposition 14 *If a path agreement is enforceable, then it is also self-enforceable.*

Proof. As discussed above, this is just an application of the previous theorem. If a path agreement is enforced, no agreement restriction is left off-the path, hence the theorem applies for $\Delta := \Delta^a$.

Hence path agreements have a desirable property from the "agreement design" point of view. If players or a mediator want to realize a given outcome of the dynamic game, without willing or trusting the possibility to threaten in advance any sort of punishment in case of deviation, they could just try to let everyone transparently believe that the path will be played. Leaving some mystery about the beliefs on the path moves cannot be of any help for the goal. A corollary of this result provides another important answer to the question about which agreements will be self-enforceabile in a dynamic game, besides complete agreements and agreements corresponding to an EFBRS.

Corollary 15 *If all the strongly-rationalizable strategy profiles of a dynamic game induce the same path, the respective path agreement is self-enforceable.*

Proof: if all the strongly-rationalizable strategy profiles induce the same path, the respective path agreement is enforceable for $\Delta := \prod_{i=1}^{N} \Delta^{H_i}(S_{-i})$ (i.e. without restrictions to first-order beliefs). Hence I can apply the proposition. ■

What about non-strongly rationalizable paths? They can never be self enforceable. But the two things cannot be put in a direct relationship, as it would be intuitive to think, because of the non-monotonicy of the CSB operator in

general. Yet, it is possible to prove that the CSB is monotonic with respect to "path restrictions", i.e. the restrictions corresponding to an agreement on a path.

Theorem 16 *Let $P \in Z$ and $\forall i \in N$, let $\Delta_i(h) = \Delta(\Sigma_{-i}(P))$ if $h \prec P$, $\Delta_i(h) = \Delta(\Sigma_{-i}(h))$ else. Then, $\{z \in Z : \exists s \in Proj_S CSB^\infty(R \cap B^*([\Delta])),\ \zeta(s) = z\} \subseteq \{z \in Z : \exists s \in Proj_S CSB^\infty(R),\ \zeta(s) = z\}$.*

Proof: see appendix.

As a consequence of this property of path restrictions, which holds also for partial path restrictions whenever the path is strongly-delta-rationalizable under them, we have the following result.

Corollary 17 *Non-strongly-rationalizable path are not self-enforcing*

Proof: Suppose by contrapposition that a path P is self-enforcing. Then $P \in \{z \in Z : \exists s \in \text{Proj}_S CSB^\infty(R \cap B^*([\Delta])),\ \zeta(s) = z\}$. Hence by the theorem $P \in \{z \in Z : \exists s \in \text{Proj}_S CSB^\infty(R),\ \zeta(s) = z\}$, i.e. tha path is strongly rationalizable. ■

What about strongly-rationalizable paths when there is more than one? Here there is no sharp answer. Looking back at the first example in the first section of the paper one can see that there can be both self-enforceable and non self-enforceable strongly rationalizable path. As a self-enforcing path, one can consider $((D, D), (D, D))$; it is easy to see that believing in it, and excluding that the opponent can ever play C in the second stage, there is no incentive to deviate from the path. Instead, the path analyzed in the section is an instance, in the same game, of a strongly-rationalizable path which is not self-enforcing. As already argued, it can be regarded as an "equilibrium path that can be upset a convincing deviation" by extending the definition in the obvious way beyond repeated coordination games; and as the next proposition shows, those paths are not self-enforceable[14].

Proposition 18 *Let $P = (\overline{a}^1, ..., \overline{a}^T)$ be a path that can be upset by a convincing deviation. $\forall i \in \{1, 2\}$, define $\Delta_i : H_i \to \Delta(S_{-i})$ as follows: $\forall h \in H_i : h \prec P$, $\Delta_i(h) = \Delta(\{s_{-i} \in S_{-i} : \exists s_i \in S_i, \zeta(s_i, s_{-i}) = P\})$; $\forall h \in H_i : h \not\prec P$, $\Delta_i(h) = \Delta(S_{-i}(h))$. Let $\Delta := \Delta_1 \times \Delta_2$. Then $CSB^\infty(R \cap B^*([\Delta])) = \emptyset$.*

Proof: Define $U : Z \to \mathbb{R}$ as $U((a^1, ..., a^T)) = \sum_{t=1}^{T} u(a^t)$. Call $(\overline{a}^1, ..., \overline{a}^{\tau-1}) =: h_0$, $(\overline{a}^1, .., \overline{a}^\tau \backslash \widetilde{a}_1) =: h_1$, $(\overline{a}^1, .., \overline{a}^\tau \backslash \widetilde{a}_1, b^i, ..., b^i) =: h_2$. $\forall s_1 \in S_1(h_1)$, $s_1 \notin S_1(h_2)$

[14] The proposition provides also an epistemic characterization of Osborne's solution concept, in the same spirit of the intuitive criterion[9] characterization provided by Battigalli and Siniscalchi [6]

(i.e. all strategies of player 1 deviating in τ but not aiming to her best prosecution afterwards) $s_1 \notin \text{Proj}_{S_1}(R \cap B^*([\Delta]))$ because $\forall \mu \in \Delta_1, \forall s_2 \in \text{supp}(\mu(h_0)), \exists s_1' \in S_1 : \zeta(s_1', s_2) = P$ and, by (1), $U_1(P) > U_1(h_2) \geq U_1(\zeta(s_1, s_2))$, thus $s_1 \notin \rho(\mu)$ (i.e.such strategies are not sequentially rational given any conjecture in delta). Instead, $\forall s_1 \in \text{Proj}_{S_1}(R \cap B^*([\Delta])) : s_1 \in S_1(h_1)$, $s_1(h) = b_1^1$ *for* $h_0 \prec h \prec h_2$.

Hence $\forall s_2 \in S_2(h_1)$, $s_2 \notin S(h_2)$ (i.e. all strategies of player 2 following the path until τ and not replying with 1's best prosecution after 1's deviation in τ), $s_2 \notin \text{Proj}_{S_2} CSB(R \cap B^*([\Delta]))$ because $\forall \mu \in \Delta_2 : \mu(h_1) \in \Delta(\text{Proj}_{S_1}(R \cap B^*([\Delta])) \cap S_1(h_1) \neq \emptyset)$, $\forall s_1 \in \text{supp}\mu(h_1)$, $\exists s_2' \in \text{Proj}_{S_2}(R \cap B^*([\Delta])) : \zeta(s_1, s_2') = h_2$ and $U_2(h_2) > U_2(\zeta(s_1, s_2))$ by (2), thus $s_2 \notin \rho(\mu)$. Instead, $\forall s_2 \in \text{Proj}_{S_2} CSB(R \cap B^*([\Delta])) \neq \emptyset$, $s_2(h) = b_j^i$ *for* $h_0 \prec h \prec h_2$.

Hence $\forall s_1 \in S_1(P)$, $s_1 \notin \text{Proj}_{S_1} CSB^2(R \cap B^*([\Delta]))$ because $\forall \mu \in \Delta_1, \mu(h_0) \in \Delta(\text{Proj}_{S_2}(CSB(R \cap B^*([\Delta]))), \forall s_2 \in \text{supp}\mu(h_0), \exists s_1' \in \text{Proj}_{S_1} CSB(R \cap B^*([\Delta])) : \zeta(s_1', s_2) = h_2$ and $U(\zeta(s_1, s_2) = P) < U(h_1)$ by (1), thus $s_1 \notin \rho(\mu)$.

Hence, $\nexists \mu \in \Delta_2 : \text{supp}(\mu(h^0)) \in \text{Proj}_{S_1} CSB^2(R \cap B^*([\Delta]))$, hence $CSB^3(R \cap B^*([\Delta])) = \emptyset$. ■

I conjecture that analogously, paths which are not generated by a subgame perfect equilibrium of the game cannot be self-enforceable. Among the ones that are, it is also possible that pure subgame perfect equilibrium paths are all self-enforceable or all not self-enforceable. For instance, they are all self-enforceable in the twice repetition of the following simple coordination game:

$A \backslash B$	L	R
U	$1, 1$	$0, 0$
D	$0, 0$	$1, 1$

Whatever subgame perfect equilibrium we consider, a deviation from it in the first period is not a useful signal because players can't hope to get more than under the equilibrium path in the second period. Instead, consider to repeat twice the following game. The two players must perform a task which gives a profit of 3 to each of them at the total effort cost of 2. If at least one player works, the task is performed; if only one player has worked, she will pay the total cost of effort, if instead they both have worked, they share the effort cost 1 per each.

$A \backslash B$	$Work$	$FreeRide$
W	$2, 2$	$1, 3$
FR	$3, 1$	$0, 0$

No pure subgame perfect equilibrium path is self-enforceable. If the path prescribes the same Nash equilibrium in both stages, the unhappy player can signal with a deviation the intention to switch to the preferred equilibrium in the second stage. If the path prescribes to play one Nash equilibrium in the first stage and the other Nash equilibrium in the second stage, the player whose preferred equilibrium is played in the first stage can deviate from it to signal the intention to play it in the second stage.

5 Conclusions and further research

In dynamic games, the consequences of a pre-play non-binding agreement are not as immediate as in static games, especially when the agreement is only partial. The belief in the agreement, assumed to be transparent to the players, is not enough to guarantee that the agreement will be respected, even when the agreement is part of an equilibrium of the game. The reason is that instances of forward induction reasoning, based not just on rationality beliefs but also on the beliefs in the agreement and the interaction of the two, allow the players to convey the signal that they want to gain more than under the agreement when they deviate from it. A player will surely or possibly deviate from the agreement when all or some of the best replies of the opponents to their revised conjectures allow her to reach the goal of improving the payoff with respect to what she could possibly get under the agreement. In both cases, the agreement won't be considered as "self-enforceable": in the second it will be believed but possibly not respected; in the first it won't even be believed by the strategically sophisticated players. In this second case, the problem of which beliefs then the agreement induces is open. Since there exists a molteplicity of possibilities, it is hard to think that the induced beliefs will be transparent to the players, unless they (or a mediator) recognize the unbelievability of the agreement in advance but, being willing to obtain the same result, seek for a looser instrumental agreement that can be believed and induces the same behavior of the original agreement. When this is possible, the agreement is called to be "enforceable", but it is shown that if the induced beliefs respect a reasonable requirement enforceability is impossible in absence of self-enforceability. All these concept are defined and analyzed in the framework of strong-delta-rationalizability (Battigalli, [2]), which allows to put at work the hypothesis of transparency of the belief in the agreement and the process of forward induction reasoning that strategically sophisticated players carry on. All the analysis becomes particularly sharp for an important class of agreements, path agreements. Path agreements are important because for many reasons people are not willing to discuss how to behave in case someone deviates from the agreement, or because a complete agreement may induce only the firm belief in its path, not because it is unbelievable per se (if the game has observable actions, subgame perfect equilibria are always self-enforceable in the sense defined in the paper) but because it may be transparent to the players that after a deviation no-one is willing to trust the agreement anymore, at least not up to the point of restricting the beliefs to it. What is clear already from the starting example is that the respect of an equilibrium path can be unbelievable and not just uncertain although the punishment which follow deviations in the equilibrium are not ruled out a-priori. The class of equilibrium paths that can be upset by a convincing deviation, introduced by Osborne [11] is shown to have precisely this feature, although they may the paths of strongly-rationalizable (Battigalli and Siniscalchi, [5]) equilibria. As it is shown, instead, paths which are not induced by a strongly-rationalizable strategy profile cannot be self-enforcing and not self-enforcing paths cannot be

enforced in any way (while if all strongly-rationalizable strategy profiles induce the same path, then the path is self-enforceable).

Although all the tools of the analysis can be applied to dynamic games with incomplete information, the focus of the paper has been kept on complete information for an interpretative reason. It is reasonable to think that in an incomplete information enviroment players, although taking an agreement at the interim stage, do not "promise" to be of one type or the other in the bargaining process. Yet, the agreement that players achieve surely suggests something about the type of the opponents. Then, belief restrictions induced by the agreement should not concern just the moves of the players, but also their pay-off relevant types. I conjecture that restricting a-priori the beliefs about opponents' types is not equivalent to restrict them throughout the strong-delta-rationalizability procedure, just like it happens for strategies. Therefore, there is the need to investigate deeply into the way that beliefs about types can be shaped by the agreement which is achieved, and on their consequences on the outcome of the game.

Other than expanding the analysis with respect to the class of games, it seems to be promising to relax some hypothesis of the paper. For instance, the restrictions to first-order beliefs may be not believed when they are at odds with the beliefs in rationality of some order. This idea has already been proposed in the paper to justify why path restrictions may be the only ones to hold even when players try to threaten each other agreeing on an off-the-path behaviour. Instead of eliminating directly all the off-the-path restrictions, it may be more reasonable to assume that players keep on believing also in the off-the-path restrictions if they are coherent with some opponents' strategies which are compatible with the observed behaviour and a common correct strong belief in rationality. This means that players hold the restrictions and believe in the restrictions only to break the uncertainty about the different rationalizable strategies of the opponent, that is, they put epistemic priority on rationality rather than on the restrictions as assumed in this paper[15].

Another assumption which may be reasonable to drop is the highest strategic sophistication of players. If at the other extreme players are just rational and hold the first-order belief restrictions induced by the agreement, any Nash equilibrium of the game becomes self-enforceable. On the other hand, despite it becomes less likely to have self-enforceability of parsimonious agreements, the behavioural implications of any self-enforceable agreement (in the sense of this paper) can still be obtained with just rational players by restricting their first-order beliefs to the final EFBRS delivered by strong-delta-rationalizability, with delta induced by the self-enforcing agreement. Therefore, the welfare opportunities for non-sophisticated players may be wider.

[15] The epistemic representation of the inverted epistemic priority would be given by $CSB^n([\Delta] \bigcap CSB^\infty(R))$; the pitfall of this representation is that it coincides with an empty set when the belief of some order in the restrictions is at odds with common correct strong belief in rationality, but it doesn't automatically drop the restrictions at the information sets where this incoherency arises (so for instance it wouldn't deliver the implications of path restriction instead of complete restrictions for the first example of the paper).

6 Appendix

Proof of the theorem.

The first inclusion to be proved is the following:

$$\{z \in Z : \exists s \in \text{Proj}_S CSB^\infty(R \cap B^*([\Delta])),\ \zeta(s) = z\} \supseteq \left\{z \in Z : \exists s \in \text{Proj}_S CSB^\infty(R \cap B^*([\Delta^E])),\ \zeta(s) = z\right\}.$$

This inclusion implies the absence of off-the-path restrictions under Δ. Then, the other inclusion can be shown in analogous way.

To simplify notation, I will refer to games with observable actions, so that an information set h coincides with a history and we can define the subgame $\Gamma(h)$. The assumption is not crucial for the theorem.

The external structure of the proof is by induction, but the inductive step requires to claim a lemma which is proved by contradiction through a re-iteration of a similar inductive proof, which in turn requires to claim the lemma again for smaller subgames and the proof becomes recursive. In any direction, the iterations stop when the lemma is claimed for subgames of depth 1.

Now I show formally the external proof by induction and the first iteration of the lemma. First, observe that $\forall i \in N$, the strategies in $\text{Proj}_{S_i} CSB^\infty(R \cap B^*([\Delta^E])$ are best replies to some conjecture which strongly believes $\text{Proj}_{S_{-i}} CSB^\infty(R \cap B^*([\Delta^E])$ itself. Following Battigalli and Prestipino [4] and replacing σ, Σ and indexes θ_i with respectively s, S, and i to represent completeness of information, $Proj_{S_i} CSB^n(R \cap B^*([\Delta])) = S^n_{i,\Delta}$ so I can use the latter procedure of strong-delta-rationalizability.

Inductive hypothesis: $\forall i \in N$, $\forall \widetilde{s}_i \in S^\infty_{i,\Delta^E}$, $\exists \widehat{s}_i \in S^n_{i,\Delta}$ such that $\forall h \in H(S^\infty_{\Delta^E}) \bigcap H(\widetilde{s}_i)$, $\widehat{s}_i(h) = \widetilde{s}_i(h)$.

Basis step: Consider any strategy $s_i \in S^\infty_{i,\Delta^E}$ and the conjecture $\mu \in \Delta^E_i$ which supports (justifies) it in S^∞_{i,Δ^E}. By the enforceability of the agreement for Δ^E, $\mu \in \Delta_i$, so that $s_i \in S^1_{i,\Delta}$ and the inductive hypothesis is trivially satisfied.

Inductive step. Consider any strategy $\overline{s}_i \in S^\infty_{i,\Delta^E}$ and the conjecture $\overline{\mu} \in \Delta^E_i$ which supports (justifies) it in S^∞_{i,Δ^E}. By the enforceability of the agreement for Δ^E, by the inductive hypothesis and by the lemma below, $\exists \widehat{\mu} \in \Delta_i$ (by enforceability), which strongly believes $S^1_{-i,\Delta}, ..., S^n_{-i,\Delta}$ and for which it holds that:

$$\begin{aligned} \forall \widetilde{h} &\in H_i(S^\infty_{\Delta^E}), \forall \overline{s}_{-i} \in supp\overline{\mu}(\widetilde{h}), \exists \widehat{s}_{-i} \in supp\widehat{\mu}(\widetilde{h}) \text{ such that} \quad \text{(C)} \\ \overline{\mu}(\widetilde{h})[\overline{s}_{-i}] &= \widehat{\mu}(\widetilde{h})[\widehat{s}_{-i}] \text{ and } \forall \widehat{h} \in H(S^\infty_{\Delta^E}), \overline{s}_{-i}(\widehat{h}) = \widehat{s}_{-i}(\widehat{h}) \text{ (by the i.h.)}, \end{aligned}$$

and such that:

$$\forall h \notin H(S^\infty_{\Delta^E}), h \in H(S^\infty_{-i,\Delta^E}), r_i(\widehat{\mu}) \notin S_i(h) \text{ (by the lemma)}.$$

Hence, considering that what is conjectured after histories which do not belong to $H(S^\infty_{-i,\Delta^E})$ is instead irrelevant to determine the best reply outside those

subgames, it holds that $\exists \widehat{s}_i \in S_{i,\Delta}^{n+1}$ such that $\forall h \in H(S_{\Delta^E}^{\infty}) \bigcap H(\widetilde{s}_i)$, $\widehat{s}_i(h) = \overline{s}_i(h)$.

Thus, by using the induction theorem, we can claim that $\forall \widetilde{s}_i \in S_{i,\Delta^E}^{\infty}$, $\exists \widehat{s}_i \in S_{i,\Delta}^{\infty}$ such that $\forall h \in H(S_{\Delta^E}^{\infty}) \bigcap H(\widetilde{s}_i)$, $\widehat{s}_i(h) = \widetilde{s}_i(h)$. This implies that $\{z \in Z : \exists s \in \text{Proj}_S CSB^{\infty}(R \cap B^*([\Delta])),\ \zeta(s) = z\} \supseteq \left\{z \in Z : \exists s \in \text{Proj}_S CSB^{\infty}(R \cap B^*([\Delta^E])),\ \zeta(\sigma) = z\right\}$.

Lemma 19 *Suppose that all the hypothesis of the theorem are satisfied and consider any $\overline{\mu} \in \Delta_i^E$ which supports S_{i,Δ^E}^{∞}. Then*

$\forall h \notin H(S_{\Delta^E}^{\infty})$, $h \in H(S_{-i,\Delta^E}^{\infty})$, $\forall m \in \mathbb{N}$,

if $\exists \widehat{\mu} \in \Delta_i$ which strongly believes $S_{-i,\Delta}^1, ..., S_{-i,\Delta}^m$ and for which condition (C) holds,

then $\exists \widehat{\mu} \in \Delta_i$ which strongly believes $S_{-i,\Delta}^1, ..., S_{-i,\Delta}^m$ and for which condition (C) holds such that $r_i(\widehat{\mu}) \bigcap S_i(h) = \emptyset$.

Proof.

Suppose by contradiction that $\exists m \in \mathbb{N}$ and $\exists h \notin H(S_{\Delta^E}^{\infty})$, $h \in H(S_{-i,\Delta^E}^{\infty})$ such that $\forall \widehat{\mu} \in \Delta_i$ which strongly believes $S_{-i,\Delta}^1, ..., S_{-i,\Delta}^m$ for which condition C holds w.r.t. $\overline{\mu}$ (and this premise is not void), it holds $r_i(\widehat{\mu}) \bigcap S_i(h) \neq \emptyset$. Then, in the subgame $\Gamma(h)$, by running strong-delta-rationalizability where $\forall j \in N$, Δ_j is the set of CPS which strongly believe $S_{-j,\Delta}^{m+1}|h$, we obtain a set of substrategy profiles $S^{h,\infty}$. $S^{h,\infty}$ is not the empty set because the restrictions do not rule out to strongly believe in any strategy which is a best reply to the conjectures in the own restricted set (this is true thanks to the absence of restrictions after h, the fact that players different from i can be surprised of reaching h at step $m+1$ and the fact that player i, by the contradictive hypothesis, goes to h playing any possible sequential best reply after it at step $m+1$). But then, by reiterating the proof above in the procedure which delivered the previous set of strategy profiles (which in this first iteration is $S_{\Delta^E}^{\infty}$, hence the procedure is simply strong-delta-rationalizability) with $\left\{s \in S(h) : s|h \in S^{h,\infty}\right\}$ in place of S_{Δ}^{∞}, it can be shown that $h \notin H(S_{\Delta^E}^{\infty})$ is contradicted (recall that the conjecture $\widehat{\mu}$ was derived from the conjecture $\overline{\mu}$ which player i can make at every step of strong-delta-rationalizability with Δ^E, and such conjecture can be modified to let player i strongly believe in $\left\{s_{-i} \in S_{-i}(h) : s_{-i}|h \in S_{-i}^{h,\infty}\right\}$ from h on). But the reiteration of the proof above requires to claim the lemma for some histories which follow h, for instance h'. $S^{h',\infty}$ is used to contradict that $h' \notin S^{h,\infty}$ by reiterating the proof above in the procedure which delivered $S^{h,\infty}$ (this time strong-delta-rationalizability until step m, then the procedure in the subset $\Gamma(h)$)

In any direction, the iteration stops when we reach histories of depth one. There the generated subset has the best reply property, which supports the contradiction without the need to claim the lemma again (there are no sub-histories to consider).■

Proof of the theorem.

If $CSB^{\infty}(R \cap B^{*}([\Delta])) = \emptyset$, the theorem is automatically true. Otherwise, we have that $P \in \zeta(\text{Proj}_{S} CSB^{\infty}(R \cap B^{*}([\Delta])))$ so that we have no off-the-path restrictions under Δ. Therefore we can run the proof of the theorem above for the opposite inclusion with $\Delta^{E} := \prod_{i=1}^{N} \Delta^{H_i}(S_{-i})$, without the need to have the inclusion proved above. We obtain the result of the theorem. ■

References

[1] Battigalli, P., "On Rationalizability in Extensive Form Games,"Journal of Economic Theory, 74, 1997, 40-61.

[2] Battigalli, P., "Rationalizability in infinite dynamic games with incomplete information", *Research in Economics,* 57, 2003, 1-38.

[3] Battigalli, P. and A. Friedenberg, "Forward induction reasoning revisited", Theoretical economics, forthcoming.

[4] Battigalli, P. and A. Prestipino, "Transparent restrictions on beliefs and forward induction reasoning in games with asymmetric information", 2011

[5] Battigalli, P., and M. Siniscalchi, "Strong Belief and Forward-Induction Reasoning," Journal of Economic Theory, 106, 2002, 356-391.

[6] Battigalli, P. and M. Siniscalchi, "Interactive epistemology in games with payoff uncertainty", *Research in Economics,* 61, 2007, 165-184.

[7] Ben-Porath, E., and E. Dekel, "Signaling Future Actions and the Potential for Sacrifice," Journal of Economic Theory, 57, 1992, 36-51.

[8] Brandenburger, A., and A. Friedenberg, "Self-Admissible Sets," Journal of Economic Theory, 145, 2010, 785-811.

[9] Cho I.K. and D. Kreps, "Signaling Games and Stable Equilibria", Quarterly Journal of Economics, 102, 1987, 179-222.

[10] Kohlberg, E. and J.F. Mertens, "On the Strategic Stability of Equilibria," Econometrica, 54, 1986, 1003-1038.

[11] Osborne, M., "Signaling, Forward Induction, and Stability in Finitely Repeated Games", Journal of Economic Theory, 50, 1990, 22-36.

[12] Pearce, D., "Rational Strategic Behavior and the Problem of Perfection," Econometrica, 52, 1984, 1029-1050.

[13] Renyi, A., "On a New Axiomatic Theory of Probability,"Acta Mathematica Academiae Scientiarum Hungaricae, 6, 1955, 285-335.