A COMMITMENT FOLK THEOREM

ADAM TAUMAN KALAI, EHUD KALAI, EHUD LEHRER, AND DOV SAMET

Abstract. Real world players often increase their payoffs by voluntarily committing to play a fixed strategy, prior to the start of a strategic game. In fact, the players may further benefit from commitments that are conditional on the commitments of others.

This paper proposes a model of conditional commitments that unifies earlier models while avoiding circularities that often arise in such models.

A commitment folk theorem shows that the potential of voluntary conditional commitments is essentially unlimited. All feasible and individually-rational payoffs of a two-person strategic game can be attained at the equilibria of one (universal) commitment game that uses simple commitment devices. The commitments are voluntary in the sense that each player maintains the option of playing the game without commitment, as originally defined.

1. Introduction

We study the following commitment folk theorem for a general finite two-person strategic game G: When an appropriate set of voluntary commitment devices $\mathfrak{D}$ is made available to the players, the Nash equilibria in the game with commitments, $G^{\mathfrak{D}}$, span all the individually-rational correlated-strategies payoffs of the original

Date: April 11, 2007.
This paper replaces "Meta-Games and Program Equilibrium," an earlier version presented at the Second World Congress of the Game Theory Society in Marseilles (2004) and at the 15th Annual International Conference on Game Theory, Stony Brook (2004). The research of the first two authors is partially supported by the National Science Foundation Grant No. SES-0527656
Since we wrote this paper, we learned of independent related findings reported in Monderer and Tennenhotz (2006).

game G. In particular, in a decentralized manner the players may commit to individual devices that lead to fully-cooperative (Pareto efficient) individually-rational outcomes of the game.

A direct implication is that players do not have to resort to infinite (or any) repetitions in order to avoid conflicts (of the prisoners' dilemma type) between cooperative and noncooperative solutions. The availability of sufficiently rich set of individual commitment devices is enough to resolve such conflicts.

We emphasize that, at this stage of the research, our goal is only to map out the mathematical possibilities of commitment devices. The commitment devices we use are mathematical constructs, designed to illustrate the folk theorem above.

Further development of "natural" commitment devices is necessary for use in a variety of real-life applications. The discovery or construction of such natural commitment devices may, in some cases, directly improve the welfare of people and organizations engaged in strategic interaction.[1]

Another implication is that voluntary commitment devices can be more effective than correlation devices, see Aumann (1974,1987). Correlated equilibria also offer Pareto improvements over the Nash equilibria of a game. However, unlike the commitment equilibria presented in this paper, they fall short of being able to attain full cooperation in many cases. This paper is restricted to the simple setting of two-person complete-information games, even though extensions to more general settings seem plausible, see discussion in the concluding section.

While the illustration of the above folk theorem requires nothing beyond elementary mathematics, it introduces two modelling innovations. First, it avoid pitfalls

[1]There is a need to first study what conditions make devices natural. Such research, which may involve issues from psychology, bounded rationality, etc., is left for future work.

and circularities of conditional commitments by incorporating into the model a simple notion of a well-defined commitment space. Second, in order to obtain the full generality, especially in games that have no Pareto-efficient *pure-strategies* individually-rational payoffs (unlike the standard prisoners' dilemma game, for instance), our commitment space permits the use of jointly controlled lotteries, see Blum (1983) and Aumann and Maschler (1995).

Referring to the main observation of this paper as a *folk* theorem is appropriate for two reasons. First, this observation describes the same set of possible payoffs as the repeated-game *folk* theorem. Second, (and again in parallel to repeated games) this type of phenomenon has been known to many authors in different contexts. The earlier literature on commitments, however, only established possibilities of *partial* cooperation in special cases, the current paper presents a general complete folk theorem in a simple unifying model.

We next discuss commitments in real life and in some of the earlier theoretical literature. Since the subject of commitments is too large for a full survey, we selected examples that are helpful in explaining the contribution of the current paper.

1.1. **Commitments and conditional commitments.** The observation that a player in a strategic game can improve his outcome through the use of a commitment device goes back to Schelling (1956 and 1960). For example, when a player in a game delegates his play to an agent, with irreversible instruction to play strategy X, the agent may be viewed as a device that commits the player to the strategy X. The strategic delegation literature, see for example Katz and Shapiro (1985) and Fershtman and Judd (1987) study implications of strategic delegation in economic

applications. Fershtman, Judd and Kalai (1991) provide a partial delegation folk theorem for a special class of games.

Indeed, real players often use agents and other commitment devices strategically. Sales people representing sellers, lawyers representing buyers, and sports agents representing athletes are only a few examples. Early price announcements, in newspapers, on the internet and in stores, are commitments to terms of sale by retailers. Money-back guarantees are commitment devices used by sellers to overcome informational asymmetries that may prevent trade. A limited menu of options on an airline's web page is a device that commits the airlines to not discuss certain options that customers may wish to raise.

But real life examples display the use of more sophisticated, conditional, commitment devices. For example, when placing an ad that states "we will sell X at a price of $500, but will match any competitor's price,"a retailer commits itself to a conditional pricing strategy. Such conditional commitment can be more efficient. For example, in oligopoly pricing games match-the-competitors clauses make the monopolist price be a dominant strategy for all sellers, see Kalai and Satterthwaite (1986) and Salop (1986).

Legal contracts are another example of effective conditional commitment devices. Each player's commitment to honor the contract is conditioned on his opponent's commitment to honor the contract. As Kalai (1981) and Kalai and Samet (1985) show, under dynamic use of contracts, refined Nash equilibria must converge to partially efficient outcomes.

Recently, Tennenholtz (2004) introduced a sophisticated model of conditional delegation, called program equilibrium.[2] In his model, every player in a game delegates the choice of his strategy to a computer program. Each player's selected program reads the opponents' selected programs and outputs a (mixed) strategy that plays the game on behalf of the player. Equilibria in the game of choosing programs, called program equilibria, are more efficient than the unmodified Nash equilibria of the game. But they fail to reach full efficiency.

In general, however, conditioning requires caution, as conditional commitments may fail to uniquely determine the outcome, lead to circular reasoning, or generate programs that fail to terminate. For example, imagine that each of two retailers places the following ad in the paper: "we sell X at a price of $500, but will undercut any competitor's price by $50." Obviously, no pair of prices charged by the two competitors is consistent with their ads, because each of the prices should be $50 lower than the other price.

Another example is the prisoners' dilemma game. If both players commit to matching the strategy of the opponent then there are two possible outcomes: both cooperate and both defect. But if one player commits to *match* and the other commits to *mismatch* then there are no possible outcomes consistent with such commitments.

Howard (1971) initiated a study of conditional commitments through the construction of metagames. In order to avoid the contradictions and circularities above, he constructed hierarchical spaces in which higher levels of commitments

[2]See McAfee (1984) for an earlier treatment of such concepts.

are defined inductively over lower ones.[3] However, perhaps due to its complexity, Howard's model of hierarchical commitments does not seem to lead to many applications.

1.2. **Our approach and main finding.** The current paper offers two main contributions. First, we propose a general model that encompasses the various approaches and questions above without getting trapped in the definitional difficulties of conditional commitments. Second, by proving a general *full* folk theorem, we show that the *potential* of conditional commitment devices is essentially unlimited.

In our model each player may delegate her play to one of many conditioning devices that selects her strategy in the game. We require that such a device conditions on the *conditioning device* chosen by the opponent and not on the *strategy* realized from the opponent's conditioning device. To illustrate, in the newspaper ad example this requirement means that every ad must determine a unique selling price for every possible *ad* of the opponent (and not for every price that may computed from ads of the opponent). This requirement is important for having a well-defined model that avoids the circularities discussed above. But it also reflects similar real-life restrictions. In the design of contracts, for example, parties use lawyers with the hope that the outcomes of their commitments are well defined under all conceivable circumstances.

For an arbitrary finite two-person strategic game we construct a simple voluntary complete device space. Being voluntary means that each individual player may choose to commit to a device ahead of time, but may also choose not to commit and just play the game as originally defined.

[3]Klemperer and Meyer (1989) and Epstein and Peters (1999) represent economics literature that deals with related issues; but also with issues that are beyond the scope of the present paper.

The completeness of the space means that a full folk theorem is obtained through the play of its (one-shot) induced commitment game. The equilibria of the one-shot commitment game span the convex hull of *all* the individually-rational correlated payoffs of the game without commitments. In particular, our folk theorem includes all the distributions over payoff profiles obtained from mixed correlated strategies. As illustrated by the example later presented in Figure, this set can be much larger than that achievable by earlier approaches.

2. A model of commitment devices

In what follows we restrict ourselves to a fixed 2-person finite strategic game, defined by a triple $G \equiv (N = \{1, 2\}, S = S_1 \times S_2, u = (u_1, u_2) : S \to \mathbb{R}^2)$.

$N = \{1, 2\}$ is the set of players, each S_i is a non-empty finite set describing the feasible *strategies* of player i, and each u_i is the *payoff function* of player i. We use the standard convention where for every player i, player $-i$ denotes the other player.

A *mixed strategy* of player i is a probability distribution σ_i over S_i, with $\sigma_i(s_i)$ describing the probability that player i chooses the strategy s_i. A pair of independent mixed strategies $\sigma = (\sigma_1, \sigma_2)$ induces a probability distribution on S with $\sigma(s_1, s_2) = \sigma_1(s_1)\sigma_2(s_2)$. A *correlated strategy* is a probability distribution γ over S. Clearly, every pair of independent mixed strategies induces the product distribution described above, which is in particular a correlated strategy, but there are correlated strategies that cannot be obtained this way.

For a correlated strategy γ we define the (expected) payoffs in the natural way, $u(\gamma) = E_\gamma(u)$.

A *pure strategy Nash equilibrium* is a pair of strategies s, such that for every player i, $u_i(s)(= u_i(s_i, s_{-i}))\geq u_i(\overline{s}_i, s_{-i})$, for any alternative strategy $\overline{s}_i$ of player i. A *mixed strategy Nash equilibrium* is a vector of mixed strategies $\sigma = (\sigma_1, \sigma_2)$, with the same property, i.e., no player can increase his expected payoff by unilaterally switching to a different mixed strategy.

We say that a correlated strategy γ is *individually rational* if for all $i \in N$, $u_i(\gamma) \geq \min_{\sigma_{-i}} \max_{\sigma_i} u_i(\sigma_1, \sigma_2)$. For each player i let ψ_i be some fixed member of $\operatorname{argmin}_{\sigma_i}(\max_{\sigma_{-i}} u_{-i}(\sigma_1, \sigma_2))$, which we will call his *minmax strategy*. So when player i's strategy is ψ_i, then player $-i$'s payoff is at most her *individual rational payoff*.

2.1. **Commitment devices and commitment games.** In the model below, sophisticated players choose their conditioning devices optimally against each other. For example, for a pair of devices (d_1^*, d_2^*) to be an equilibrium, d_1^* must be the best device that player 1 can select against the device d_2^* of player 2, taking into account the known responses of d_2^* to hypothetical alternatives to d_1^*.

A non-empty set D_i describes the *conditional commitment devices* (or just devices) available to player i. With every device $d_i \in D_i$ there is an associated *device response function*: $r_{d_i} : D_{-i} \to S_i$ where $r_{d_i}(d_{-i})$ denotes the strategy that d_i selects for player i, if it plays against the device d_{-i} of the opponent.

However, to ease the discussion we use a more compact represention for the response functions. The responses of the various devices of player i are aggregated into one *(grand) response function* $R_i : D_1 \times D_2 \to S_i$, where $R_i(d_i, d_{-i}) = r_{d_i}(d_{-i})$ describes the strategy chosen by the device d_i of player i when matched against the device d_{-i} of the opponent. The two response functions together describe a *joint*

response function $R(d_1, d_2) = (r_{d_1}(d_2), r_{d_2}(d_1))$ where $R(d_1, d_2)$ describe the pair of strategies selected by the devices when they respond to each other.

Note, however, that any function $R\colon D_1 \times D_2 \to S$ is a possible joint response function. This reasoning motivates the simple definition of a commitment space below.

Definition 1 (Device Space)**.** *A* space of commitment devices *(also a device space) of* G *is a pair* $\mathfrak{D} \equiv (D = D_1 \times D_2, R\colon D \to S)$.

Each D_i *is a nonempty set describing the possible* devices of player i*, and* R *is the* joint response function*. The associated* device response functions *are defined (as above) by* $r_{d_i}(d_{-i}) = R_i(d_i, d_{-i})$.

A device space $\mathfrak{D}$ induces a two-person *commitment game* $G^{\mathfrak{D}}$ (or device game) in the following natural way. The feasible pure strategies of player i are the devices in the set D_i and the payoff functions are defined by $u(d) = (u_1(R(d)), u_2(R(d)))$ (we abuse notation by using the letter u to denote both the payoffs in G and the payoffs in $G^{\mathfrak{D}}$).

Definition 2 (Commitment-Device Equilibrium)**.** *A* commitment-device equilibrium *(or* device equilibrium*) of the game* G *is a pair* $(\mathfrak{D}, \sigma)$*, consisting of a device space* $\mathfrak{D}$ *and an equilibrium* σ *of the device game* $G^{\mathfrak{D}}$.

Clearly, the pair of payoffs of any pair of mixed strategies in the device game, including any device equilibrium, are the payoffs of some correlated strategy in G.

Of special interest to us are the equilibrium payoffs in *voluntary* commitment spaces. These allow each player i to play the game G as scheduled, without making any advanced commitment. In other words, he can choose any G strategy $s_i \in S_i$

without conditioning on the opponent's choices and with the opponent not being able to condition on s_i. Formally, we incorporate this into a device space by adding to it *neutral* (non committal) devices.

Definition 3 (Voluntary)**.** *The device space $\mathfrak{D}$ is* voluntary for player i *if for every strategy $s_i \in S_i$, his set of devices, D_i, contains one designated* neutral device $s_i^{\mathfrak{D}}$ *with the following two properties.*

(1) *Unconditioned play: for every $d_{-i} \in D_{-i}$, $r_{s_i^{\mathfrak{D}}}(d_{-i}) = s_i$.*

(2) *Private play: for every $d_{-i} \in D_{-i}$, and $s_i, \overline{s}_i \in S_i$, $r_{d_{-i}}(s_i^{\mathfrak{D}}) = r_{d_{-i}}(\overline{s}_i^{\mathfrak{D}})$.*

A voluntary device space *is one that is voluntary for both players.*

3. Elaboration on the model.

A trivial example of a voluntary commitment space is the game itself, with each $D_i = S_i$, where $G^{\mathfrak{D}} = G$. But all the examples discussed in the introduction, delegation to agents, newspaper ads, contracts, program equilibrium, and many more can be effectively described by the model above. The next example illustrates this point.

Example 1 (Price competition)**.** *Consider two retailers,* 1 *and* 2*, preparing to compete in the sales of X in the upcoming weekend. The game G is described by the (per-unit) prices that each retailer may charge, and the payoff of each retailer is the profit realized after informed buyers choose who to buy from. Assume, for simplicity, that there is a known demand curve, that buyers buy from the less expensive retailer, and that if their prices are the same, the demand is equally split.*

As discussed in the introduction, this game lends itself to the use of commitment devices in the form of newspaper ads posted in Friday's newspaper. To fit into the

formal model above, we may let D_1 and D_2 describe (respectively) all the ads that the two retailers are allowed to post. With the D_i's specified, it is straightforward to verify that ads lead to well-defined prices: one must check that for every ad of player i, d_i, there is a well define price of retailer i, $r_{d_i}(d_{-i})$, resulting from every competitor's ad, $d_{-i} \in D_{-i}$. This formulation disallows vague ads, like "I will undercut opponents' prices by \$50," which fail to specify a response price to an identical competitor's ad.

Notice that the mathematical need for a well-defined model reflects a legal real-life need for coherence. Indeed, we often see ads of the type, "we will meet any *posted price* of our competitors." A restriction of this type may be described by a model in which an ad consist of two items, a posted price, p, and a rule, h, that responds only to posted (not computed) prices of the opponent. In this case, the device set of player i consists of all such pairs (p_i, h_i), and if retailers 1 and 2 place the ads $d_1 = (p_1, h_1)$ and $d_2 = (p_2, h_2)$ then the selling prices are $R(d_1, d_2) = (h_1(p_2), h_2(p_1))$.

3.1. **More effective model.** Earlier attempts to deal with sophisticated conditional commitments (without the use of well defined commitment device spaces) lead to difficult models. Howard (1971) wanted to describe a notion of a meta strategy, one that conditions its choice of an action based on the action chosen by the opponent. For example, a player in a one shot prisoners' dilemma game should be able to match-the-opponent, and in effect induce a tit-for-tat strategy in the one shot game.

But this plan proved to be difficult due to the issue of timing. How can a player react to his opponent's choice, if they play simultaneously?

Howard's solution was to construct an infinite hierarchical structure of reaction rules: At the lowest level each player chooses a strategy in the underlying game, and at level $t+1$ he specifies response rules to his opponent level t rules.

The model of the current paper offers a simpler, more manageable solution to the apparent contradiction between timing and commitment. This is possible because a player's device conditions on the device chosen by the opponent, and not on the strategy produced by the opponent's device. The following is a simple illustration of this simple useful idea.

Example 2 (Divorce-settlement)**.** *The game is a simple model of divorce between two players, he and she. The underlying game is exactly like the standard Prisoners' Dilemma game with cooperative (c) and aggressive (a) strategies.*

But assume now that each player has the option of choosing a lawyer to represent him in the game and that lawyers are of two possible types: flexible (fl) and tough (tl) (and lawyers know the types of other lawyers).

No matter who they face, tls choose the strategy a. But fls choose the strategy c when they face an opponent of type fl, and choose the strategy a against all others.

A voluntary commitment-device space for the above situation may be described by $\mathfrak{D} \equiv (D = D_1 \times D_2, R\colon D \to S)$ as follows. Each $D_i = \{fl, tl, c^{\mathfrak{D}}, a^{\mathfrak{D}}\}$ and the response function R is described by the table below.

Notice that the R-table describes the behavior of the lawyers. For example, as can be seen in the top row, if player one commits to an fl device, he ends up cooperating against an fl device of the opponent, but aggressing against all the other opponent's devices. The $c^{\mathfrak{D}}$ and $a^{\mathfrak{D}}$ devices satisfy the conditions of neutral devices. For example, when Player 1 "commits" to $c^{\mathfrak{D}}$, he ends up cooperating

unconditially (no matter what device is chosen by the opponent). Moreover, his choice is private, as non of the devices of Player 2 (in choosing an action for Player 2) ever differentiate between the devices $c^{\mathfrak{D}}$ and $a^{\mathfrak{D}}$ of Player 1 (since the second entries in the two bottom cells of every column are identical).

If one substitutes the prisoners' dilemma payoffs in the sixteen cells in the table (assuming that the lawyers fees are negligible :), it is easy to see that fl,fl is a dominant strategy equilibrium. In effect, this equilibrium employs a tit-for-tat type of strategy to get cooperation in this one shot prisoners' dilemma game: a player deviating from fl causes the opponent's device to switch from c to a.

		Pl 2			
		fl	tl	cd	ad
Pl 1	fl	c,c	a,a	a,c	a,a
	tl	a,a	a,a	a,c	a,a
	cd	c,a	c,a	c,c	c,a
	ad	a,a	a,a	a,c	a,a

4. A commitment folk theorem

4.1. **Technical subtleties.** Unlike the example above, where the construction of a cooperative commitment equilibrium is easy, the proof of a general folk theorem is more subtle. Before continuing with the formal part, we point to two of the technical difficulties.

Example 3. *(a modified prisoners' dilemma: fight-or-relinquish)*

		Pl. 2	
		fight	*relinq*
Pl. 1	*fight*	2,2	10,0
	relinq	0,10	0,0

Here, the minmax strategies guarantee each player a payoff of at least 2. But unlike in the standard prisoners' dilemma game, there is no pair of pure strategies that simultaneously yield each player a payoffs greater than 2. Yet a full folk theorem should have Nash equilibria generating every payoff profile in the convex hull of $\{(2,2),(2,8),(8,2)\}$, for example $(5,5)$.

In the repeated-game folk theorem this is not a problem since the players can alternate in playing the cells $(fight, relinquish)$ and $(relinquish, fight)$, and a trigger strategy will induce the correct incentives to do so. But such alterations are impossible if the game is played only once.

A second difficulty in the proof of a general commitment folk theorem is illustrated by the next example.

Example 4. *(a game with fighting options)*

		Copier		
		style A	*style B*	*relinq*
Trend setter	*style A*	1,3	3,1	10,0
	style B	3,1	1,3	10,0
	relinq	0,10	0,10	0,0

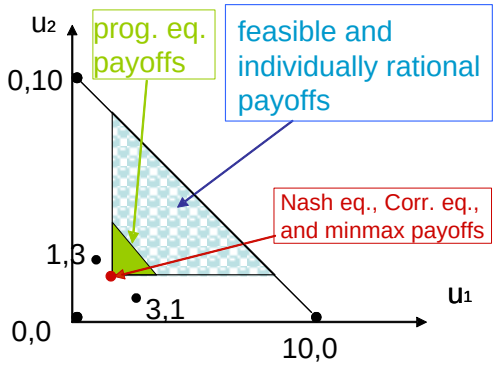

FIGURE 1. The payoffs achievable in Example 4. The full feasible individual rational region is significantly larger than that of the payoffs achievable by program equilibria. The unique Nash and correlated equilibrium is (2,2).

This game is similar to the modified prisoners' dilemma above, where in any cooperative outcome one of the players relinquieshes. But in this game *none* of the individually-rational feasible payoffs (i.e., the convex hull of $\{(2,2),(2,8),(8,2)\}$), including the minmax payoffs, are generated by pure strategies. Thus, when triggering to noncooperation a player must mix $.50-.50$ between his *style* A and *style* B strategies. But our definition of a commitment space prohibits commitment devices that randomize.

This second difficulty is overcome by moving the randomization from the stage of triggering to an earlier stage, when the player chooses a device. In the stage of choosing devices the player randomizes, and chooses with probabilities $.50-.50$ a device that triggers to *style* A or a device that triggers to *style* B. But doing this may still not suffice, since the chosen device is observable to the opponent's device, who would know whether the player plans to trigger him with the pure strategy *style* A or the pure strategy *style* B.

There are a variety of ways of dealing with this last difficulty. For example, a player may choose a device that punishes any opponent's device that conditions on the punshing strategy of the player.

The device space constructed in our proof below is carefully chosen to be rich enough in some aspects, but not so in others. It allows for jointly controlled lotteries, a la Blum (1983) and Aumann and Maschler (1995), to replace the alterations of an infinitely repeated game by one stage randomization. But it disallows devices that can react to certain pure choices made by the opponent ex-ante, as a way to avoid the second difficulty discussed above.[4]

4.2. **Formal statement and proof.**

Definition 4. *A space* $\mathfrak{C}$ *of commitment devices is* complete *for the game* G*, if the payoff profile of every individually-rational correlated strategy in the game* G *can be obtained at some (possibly mixed strategy) Nash equilibrium of the commitment game* $G^{\mathfrak{C}}$

Theorem 1 (Commitment-device folk-theorem)**.** *For any finite two-person game* G*, there is a complete voluntary space of commitment devices* $\mathfrak{C}$*.*

Proof. We first construct $\mathfrak{C}$. It will have an infinite set of devices, to be used as strategies in the commitment game $G^{\mathfrak{C}}$. The strategies of a player i are a triples, where the first part is an encoding of a correlated strategy, the second part is a number in the interval $[0, 1]$, and the third part is a fall-back strategy in S_i. Let $M = |S|$ and let $[M]$ denote $\{1, 2, \ldots, M\}$.

[4]This is done for simplicity in the proof of the folk theorem. One can produce alternative proofs with more natural device spaces. Further discussion of related issues is offered at the concluding section of the paper.

We now describe a method for encoding any correlated strategy γ over S by a unique $x \in \Delta_M = \{x \in [0,1]^M | \sum_i x_i = 1\}$, the simplex of dimension $M-1$. The important property is that there is a function $f : \Delta_M \times [0,1] \to S$ such that the probability that $f(x,r) = s$ for a uniformly random $r \in [0,1]$ is the same as the probability assigned to s by γ. (There are several ways to achieve this, and any other method of achieving it would be satisfactory.) For completeness, we give one such encoding now. Any $x \in \Delta_M$ corresponds to a probability distribution over $[M]$ by choosing r uniformly from $[0,1]$ and the following map $g : \Delta_M \times [0,1] \to [M]$,

$$g(x,r) = \min\big\{j \in [M] \;\big|\; x_1 + x_2 + \ldots + x_j \geq r\big\}.$$

Finally, let $\pi : [M] \to S$ denote an arbitrary bijection from $[M]$ to S. The map π should be fixed and known in advance to all players. Hence, Δ_M gives a unique encoding of correlated strategies over S, where the correlated strategy corresponding to $x \in \Delta_M$ is chosen by picking r uniformly at random from $[0,1]$ and taking $f(x,r) = \pi(g(x,r))$.

We can now specify $\mathfrak{C} = (C\ (= C_1 \times C_2)\ , L)$. $C_i = (\Delta_M \cup \{\perp\}) \times [0,1] \times S_i$. The special symbol $\perp$ is necessary to make the game voluntary, and indicates that the player wants to play the fall-back strategy, and L is defined by,

$$L\big((x_1,r_1,s_1),(x_2,r_2,s_2)\big) = \begin{cases} f(x_1, r_1 + r_2 - \lfloor r_1 + r_2 \rfloor) & \text{if } x_1 = x_2 \text{ and } x_1 \neq \perp \\ (s_1, s_2) & \text{otherwise} \end{cases}$$

The expression $r_1 + r_2 - \lfloor r_1 + r_2 \rfloor$ above computes the fractional part of $r_1 + r_2$.

Now let γ be an individually rational correlated strategy of G. We will see that there is a mixed device equilibrium of $\mathfrak{C}$ with an outcome distribution that coincides

with the correlated strategy γ. Let x be the unique encoding of γ so that, for any $s \in S$, the probability that $f(x, r) = s$ is equal to the probability that γ assigns to s. Take the mixed device for each player μ_i that chooses (x_i, r_i, s_i) by taking $x_i = x$ (with probability 1), $r_i \in [0, 1]$ uniformly at random and, independently, s_i according to the mixed minmax strategy of player i.

To see that $\mu = (\mu_1, \mu_2)$ has the desired properties, notice first that for any r_i chosen by player i, the equilibrium strategy of the opponent induces the distribution γ on S. In other words, player i cannot gain by deviating from the uniform distribution on his r_i's. Moreover, deviating by submitting a vector $x_i' \neq x$, makes him face the minmax distribution of his opponent, which can only decrease his payoff.

The game is voluntary because player i has neutral strategy $(\perp, 0, s_i)$ for any strategy $s_i \in S_i$. □

The proof of the theorem above uses infinitely many commitment devices. Two finite folk theorems are presented in an appendix to this paper. One is an *approximately* complete folk theorem with a finite number of devices. The other shows that an *(exact)* complete folk theorem with a finite number of devices can only be obtained for a highly specialized class of games.

5. Comparison with earlier notions

5.1. **Comparison to correlated equilibria.** As it turns out, the set of commitment-equilibrium payoffs is significantly larger than that of correlated-equilibrium payoffs. For example, In the fight-or-relinquish game above the *only* correlated equilibrium payoffs are $(2, 2)$, whereas any payoffs in the convex hull of $\{\{(2, 2), (2, 8), (8, 2)\}\}$ (including $(5, 5)$) can be obtained at commitment equilibria.

Given a game G, there are some important epistemological differences between the devices used to amend G. Aumann's (1974, 1987) correlation device outputs, prior to the start of the game, a vector of individual private messages according to a commonly known probability distribution.[5] The players proceed to play G after learning their private messages. Once a player received a signal he has no way to affect the other players' strategies.[6]

In the commitment setting, the game is amended with a commonly known space of commitment devices (no probability distributions). The players may choose individual commitment devices from this space. However, due to the conditioning, by changing his own commitment a player may change the other players' strategies.

5.2. **Comparison to delegation.** The delegation folk theorem presented in Judd, Fershtman and Kalai (1991) starts with a game G and states that the payoffs of any pure strategy profile of G that Pareto dominates some pure strategy Nash equilibrium of G, can be obtained at a Nash equilibrium of the game with delegation.

From a technical point of view, the proof of their delegation folk theorem is relatively easy, since it bypasses the two technical difficulties discussed prior to the proof of our folk theorem. Not surprisingly, the applications of their delegation folk theorem are severely limited, as can be seen by the two examples above.

In the fight-or-relinquish game, the only "new" equilibrium that may be deduced from this delegation folk theorem is when both players $fight$. Thus, unlike our commitment folk theorem (that enables payoffs like $(5, 5)$), their delegation theorem

[5]To ease the discussion, we use the notion of common knowledge carelessly. As readers familiar with the literature are aware, less than full common knowledge suffices in statements made here.
[6]To generate the probability distribution of a correlation device one needs an external impartial mediator, or, alternatively, use a system of devices that produces signaling that induce the desired correlated distribution over the game outcome (see, Barany (1992), Lehrer (1996), Lehrer and Sorin (1997), Ben-Porath (1998), Gossner (1998), and Urbano and Vila (2002)).

offers no Pareto improvements. In the game with fighting options, which has no pure strategy equilibria, their delegation folk theorem is vacuous.

5.3. **Comparison with program equilibrium.** Tennenholtz (2004) presents a *partial* folk theorem using program equilibria: The program equilibrium payoffs of a game G consist of all the individually-rational payoff pairs that can be obtained through *independent* (not correlated) mixed strategies of G. Applying the result of Tennenholtz to the modified-prisoners'-dilemma game above the largest symmetric program-equilibrium payoffs are $(3\frac{1}{8}, 3\frac{1}{8})$, short of the efficient payoffs $(5, 5)$ that can be obtained at a commitment equilibrium (as defined in this paper).

Tennenholtz's programs may be viewed as commitment devices, but there are important differences between the formal models. A commitment device, as defined in this paper, outputs a pure strategy for a player. A program, in Tennenholtz's model, outputs a mixed strategy for a player. Thus, for better or worse, Tennenholtz's programs are more sophisticated and offer more flexibility than our commitment devices.

Given this added flexibility, one may expect Tennenholtz to get a larger, rather than the obtained smaller, set of equilibrium payoffs. But this is explained by another important difference. Tennenholtz's analysis is restricted to the payoffs obtained through the use of *pure*-strategy program equilibria, while our model allows for *mixed*-strategy commitment equilibria.

There are pros and cons regarding the differences in the timing of randomization. Since Tennenholtz allows his devices to output mixed strategies, it is easier in his model to trigger punishing when the minmax strategy is not pure (recall the second difficulty we mention prior to the proof of the commitment folk theorem).

But there are advantages to allowing the mixing to be done ex-ante. First, it is necessary for a *full* folk theorem. But also from a conceptual view point, ex-ante randomization, done in a player's mind prior to a choice of a strategy, may be less demanding than having to construct devices that can randomize.

6. Additional remarks

6.1. **On natural commitment devices and implementation.** Cooperation obtained through commitment devices is useful in real-life situations. But its applicability depends largely on the availability of devices that are *natural* for such situations. We have discussed examples of natural devices: newspaper ads, delegates, computer programs, etc. But our proof of a general theorem relies on devices that are not natural for most real-life situations.

A formal investigation of natural commitment devices is a conceptual challenge that would require considerations beyond the scope of this paper.

A related question is, where do commitment devices come from? Is there an outside entity (other that the players of the game) able to construct commitment spaces for the players, or are commitment devices something the players generate on their on? Under the former narrower case, the study of commitment may be viewed as a subarea of the implementation literature, see the survey of Jackson (2001).[7]

But under the latter and more general case, the study of natural commitment devices may lead to deeper issues about the evolution of language and vocabulary.

[7]Indeed, the paper of Monderer and Tennenholtz (2006) mentioned in out introductory footnote, takes an implementation viewpoint on these issues.

For example, what makes "match my competitors" ads natural? Could an evolution of language and communication devices lead to other, perhaps better, natural commitment devices.

The implementation literature raises another issue. In the commitment folk theorem we devise a complete commitment space, one that spans all the individually rational correlated payoffs in the game. But it may be desirable to construct (natural) partial commitment spaces that span more restricted sets of payoffs. For example, it may be desirable to generate only the Pareto efficient ones or even subsets of these, like ones consisting of "fair" outcomes.[8]

6.2. **Extensions to n-players.** When dealing with more than two players, repeated-game folk theorems bring about some modeling choices. For example, if player i deviates from the equilibrium, can the remaining players secretly correlate their future strategies in order to achieve a more effective punishment against him? Different answers to the above question lead to different equilibrium sets.

Similar related choices must be faced when dealing with commitment devices of more than 2 players. For example, in the two-player case studied in this paper we assume that every player's device can condition on (e.g., it sees) the device used by his opponent. When we deal with more players, are all devices fully visible to all the players' devices, or should we allow each coalition to have devices that are only visible to the devices of its own members?

What equilibrium payoffs can be obtained under a various visibility assumptions? Can the results of Aumann (1961) on Alpha and Beta cores in repeated games be reproduced in one shot games with commitment devices?

[8]Nash (1953), Raiffa (1953) and Kalai and Rosenthal (1978) suggest examples of such commitment spaces.

6.3. **Commitment in Bayesian games.** Restricting ourselves to complete information games, the folk theorem above shows that strategic inefficiencies may be removed by commitments. The following example shows that one may expect similar improvements with regards to informational inefficiencies. Specifically, commitments may be used as means for communication.

Example 5 (Hunting a hidden stag)**.** *Consider two players,* 1 *and* 2*, and three locations,* H_1, H_2*, and* H_3*. A prize is located at random in one of the three locations (with probability* $1/3$ *for each), and each player* i *, who is initially located at* H_i *, is told whether the prize is at* his location*, or not. Following this, in one simultaneous move, each player chooses one of the three locations. If both players choose the location with the prize they are paid one dollar each, otherwise zero. Assuming no communication, the highest achievable equilibrium payoff is* $2/3$ *each.*

When dealing with commitments in Bayesian games, there are several modeling choices. For example, are the individual commitments done before or after the private information is revealed. Assuming the latter, the example above illustrates that, from considerations of Pareto efficiency, commitment devices may serve as effective communication devices.

Consider a commitment space in which each player i has two devices, s_i (for stubborn) and f_i (for flexible). The device s_i chooses the location H_i no matter what device is used by the opponent. The device f_i chooses the location H_{-i} against the device s_{-i} of the opponent, but chooses H_3 against the device f_{-i} of the opponent. Consider the strategy profile where each player i chooses s_i when the prize is at his location and f_i otherwise. It is easy to see that this is an equilibrium that guarantees that they both show up at the right place, whichever one it is.

6.4. **Uncertain, partial, and dynamic commitment.** What can be achieved by devices that are not fully observable? This issue has been partially studied in the delegation literature. For example, Katz and Shapiro (1985) argued that unobserved delegation could not really change the equilibrium of a game. On the other hand Fershtman and Kalai (1997) have shown that under restriction to *perfect* Nash equilibrium, even unobserved delegation may drastically affect payoffs.

Another important direction is partial commitments. What if the commitment devices do not fully determine the strategies of their owners, but only restrict the play to subsets of strategies, to be completed in subsequent play by the real players?

It seems that a fully developed model of commitments should allow for the options above and more. It should be dynamic, with gradually increasing levels of commitments that are only partially observable.

6.5. **Contracts.** While technically speaking the commitment equilibria discussed in this paper are decentralized, they still require a high degree of coordination due to the large multiplicity of the equilibria. This is an important issue when dealing with the selection of contracts.

First, to fit into our formal model, imagine a possible transaction between a seller and a buyer, conducted in a certain real-estate office. The real estate agent may have a large (possible infinite) number of contacts around, and each of the two players can choose to sign any of these contracts. But unless they both choose to sign the *same* contract, the transaction does not take effect. If there are positive gains from the transaction, there is a large multiplicity of (equilibrium) contracts that may be signed.

Without communication, it is hard to imagine that the parties will sign the same contract. But under nonbinding (cheap talk) communication, it is fairly likely that they would coordinate and sign the same contract (as we observe in real life situations). Thus, in situations where binding contracts are legal, contracts combined with cheap talk are a natural and effective commitment devices of the type discussed in this paper.

But from the game theoretic perspectives, the contracts described above are pure-strategy Nash equilibria. Thus they may not suffice for generating the full gains from cooperation as described in the commitment folk theorem.

To gain the full benefits, it may be desirable to mimic the ideas in the commitment folk theorem by allowing *strategic contracts*. These would incorporate the possibilities of jointly controlled lotteries into the contract agreement.

For a concrete example consider the (version of battle of the sexes) game described below.

		Pl. 2 *insist*	Pl. 2 *yield*
Pl. 1	*insist*	0,0	3,2
	yield	2,3	0,0

For a helpful interpretation, imagine that there is one precious indivisible item to be allocated to one of the two players (e.g., custody of a child). If one player *insist*s and the other one *yield*s, the item is allocated to the insisting player. In all other situations neither one of the players is allocated the item.

Can they sign a contract, regarding their chosen strategies, that guarantees the (fair) allocation of the item to one of them? An obvious solution is a randomizing contract. For example, this contract may stipulate that some impartial mediator

will flip a coin, if it shows H, they would play $(insist, yield)$, and if it shows T, they would play $(yield, insist)$.

But the use of the outside randomizing mediator may be avoided through the use of a strategic contract. For example, each player would submit a sealed envelope with an integer $s_i = 1$ or 2 together with a contract that states that if their integers match, they would play $(insist, yield)$ and if their integers mismatch they would play $(yield, insist)$. Under such a contract, submitting the integers 1 or 2 with equal probabilities guarantees each player the expected payoff of 2.5, no matter what integer the opponent submits.[9]

7. Appendix

7.1. **Finite number of devices.** The device space in the commitment folk theorem is infinite. It may be important to note that a finite version approximation of the above folk theorem can be made where correlated strategies have coefficients that are integer multiples of $1/n$, meaning that the probabilities assigned to the different strategies $s \in S$ are in the set $\left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}$. While this does not give a full-folk theorem, it is sufficient for many practical purposes and has the advantages of being finite.

[9]The simultaneous submission of sealed envelope with an *invisible* integer can be replaced by the submission of *observable messages*, under assumptions from the theory of cryptography, see Naor (1991). For example, each player may openly submit with the contract a large integer that is the product of two or of three prime numbers. The contact will condition, in the same manner as the one above, on matching or mismatching the number of factors of the two submitted integers. By current assumptions of cryptography, it is practically impossible for any player, other than the one submitting the number, to know whether the observed submitted integer has 2 or 3 factors. But it is trivial for the player who constructed the number to illustrate the answer to this question. So in effect, the observed submitted numbers still have "sealed" values of 2 or 3, until the players "open" them by revealing their factorizations.

Theorem 2 (Finite commitment-device folk-theorem)**.** *For the two player game G and any $n \geq 1$, there exists a* finite *voluntary commitment device space $\mathfrak{C}_n = (C_n, L_n)$ with a commitment game $G_n^{\mathfrak{C}}$ that has the following property. Every individually rational correlated strategy in the game G whose coefficients are in $\left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}$ can be obtained as a (mixed strategy) Nash equilibrium of the commitment game $G_n^{\mathfrak{C}}$. Moreover, the function L_n can be computed in time polynomial in* $\log(n)$.

The proof of the above theorem is nearly the same as that of Theorem 1. The only difference is that the correlated strategies (and simplex) are discretized to an accuracy of $1/n$ and the players choose $r_i \in \left\{\frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}$ uniformly at random. Such numbers are represented using $O(\log n)$ bits. The function L_n is straightforward to efficiently compute, i.e., compute in time polynomial in the input length.

In some applications, a finite number of commitment devices may be sufficient to achieve a full folk theorem. It may be useful to know, however, that for the folk theorem with the generality above (one complete commitment space that achieves all equilibria of the game G) one needs infinitely many devices, unless the game is of a very narrow form. The following is a sketch of such a theorem and its proof.

Theorem 3. *For any two player game G the following two conditions are equivalent:*

1. There exists a finite device space $\mathfrak{C}$ in which every individually rational correlated strategy in G can be obtained as a Nash equilibrium of $G^{\mathfrak{C}}$,

2. The feasible payoffs set of G is a rectangle with facets parallel to the axes.

Proof. If (2) holds, then there are four payoffs in the game which are the extreme points of the feasible set. Thus, one can define a 2×2 device game in which each

player controls the payoff of the other and has no say over her own payoff. The equilibrium payoffs in this game are the entire feasible set of G.

As for the converse, assume (1) and that (contrary to (2)) one the facets of G's payoffs, say F, is not parallel to one of the axes. Since F is a facet of the feasible set, in order to obtain a (correlated) payoff in F, all the payoffs involved should be also in F.

Let $\sigma = (\sigma_1, \sigma_2)$ be any equilibrium of $G^{\mathfrak{C}}$ whose payoff is in F and let $\mathfrak{C}^{\sigma} = (D_1^{\sigma} \times D_2^{\sigma}, T)$ where each D_i^{σ} denotes the supports of σ_i. The payoffs of $G^{\mathfrak{C}^{\sigma}}$ are all in F. Moreover, σ induces a full-support equilibrium of $G^{\mathfrak{C}^{\sigma}}$.

Consider any subspace $\mathfrak{C}' = (D_1' \times D_2', T)$ of $\mathfrak{C}$ where all payoffs of $G^{\mathfrak{C}'}$ are in F. By a linear transformation of the payoffs of player 1, $G^{\mathfrak{C}'}$ can be transformed to a zero-sum game, say $G_0^{\mathfrak{C}'}$. As a zero-sum game $G_0^{\mathfrak{C}'}$ has only one equilibrium payoff. In particular, all full-support equilibria of $G_0^{\mathfrak{C}'}$ induce the same payoff.

Since $G_0^{\mathfrak{C}'}$ is derived from $G^{\mathfrak{C}'}$ by a linear transformation (of the payoffs of one of the players), any full-support equilibrium of $G_0^{\mathfrak{C}'}$ is a full-support equilibrium of $G^{\mathfrak{C}'}$. Consequently, any $G^{\mathfrak{C}'}$ has only one full-support equilibrium payoff. Since there are finitely many subgames $G^{\mathfrak{C}'}$ in $G^{\mathfrak{C}}$ with payoffs in F, and each has at most one full-support equilibrium payoff, there are only finitely many equilibrium payoffs of $G^{\mathfrak{C}}$ in F. Thus, the equilibrium payoffs of $G^{\mathfrak{C}}$ cannot cover all the correlated equilibrium strategies payoffs in F. This contradiction leads to the conclusion that if $G^{\mathfrak{C}}$ is finite, then all the facets of the feasible set of $\mathfrak{C}$ are parallel to the axes. $\square$

8. References

Aumann, R.J. (1961), "The core of a cooperative game without side payments," *Transactions of the American Mathematical Society,* **98**, 539-552.

Aumann, R.J. (1974), "Subjectivity and correlation in randomized strategies." *Journal of Mathematical Economics*, **1**, 67-96.

Aumann, R.J. (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, **55**(1), 1-18.

Aumann, R.J. and M. Maschler (1995), *Repeated Games with Incomplete Information*, MIT Press.

Barany, I. (1992), "Fair distribution protocols or how players replace fortune," *Mathematics of Operations Research*, **17**, 327-340.

Ben-Porath, E. (1998), "Correlation without mediation: expanding the set of equilibrium outcomes by cheap pre-play procedures,' *Journal of Economic Theory*, **80**, 108-122.

Blum, M. (1983), "Coin Flipping by Telephone: A Protocol for Solving Impossible Problems," SIGACT News, **15**(1), 23-27.

Epstein, L. and M. Peters (1999), "A Revelation Principle for Competing Mechanisms,"*Journal of Economic Theory*, **88**, 119-161.

Fershtman, C., and K. Judd (1987), "Equilibrium Incentives in Oligopoly," *American Economic Review*, **77**(5), 927-940.

Fershtman, C., K. Judd and E. Kalai (1991), "Observable Contracts: Strategic Delegation and Cooperation," *International Economic Review*, **32**(3), 551-59.

Fershtman, C. and E. Kalai (1993), "Unobserved Delegation," *International Economic Review*, **38**(4), 763-74.

Fudenberg, D. and E. Maskin (1986), "Folk Theorem for Repeated Games with Discounting or with Incomplete Information," *Econometrica*, **54**(3), 533-554.

Gossner, O (1998), "Secure Protocols or How Communication Generates Correlation," *Journal of Economic Theory*, **83**(1), 69-89.

Howard, N. (1971), *Paradoxes of Rationality: Theory of Metagames and Political Behavior*, The MIT Press, Cambridge.

Jackson, M.O. (2001), "A Crash Course in Implementation Theory," *Social Choice and Welfare*, **18**(4), 655-708.

Kalai, E. and R. W. Rosenthal (1978), "Arbitration of Two-Party Disputes under Ignorance," *International Journal of Game Theory*, 7(2), 65-72

Kalai, E. and M. Satterthwaite (1986), "The Kinked Demand Curve, Facilitating Practices and Oligopolistic Competition," DP 677, Center for Math Studies in Econ and Mgt Science, published also in Imperfection and Behavior in Economic Organizations, R. P. Gilles and P. H. M. Ruys (eds.), Kluwer Academic Publishers, 1994, 15-38.

Kalai, E. (1981), "Preplay Negotiations and the Prisoner's Dilemma," *Mathematical Social Sciences*, **1**(4), 375-379.

Kalai, E. and D. Samet (1985), "Unanimity Games and Pareto Optimality," International Journal of Game Theory, **14**(1), 41-50.

Katz, M. L., and C. Shapiro (1985), "Network Externalities, Competition, and Compatibility," *The American Economic Review*, **75**(3), 424-440.

Klemperer, P.D., and M.A. Meyer (1989), "Supply Function Equilibria in Oligopoly under Uncertainty," *Econometrica*, **57**(6), 1243-1277.

Lehrer, E. (1996), “Mediated talk," *International Journal of Game Theory*, **25**, 177-188.

Lehrer, E. and S. Sorin (1997), “One-Shot Public Mediated Talk," *Games and Economic Behavior*, **20**(2), 131-148.

McAfee R.P. (1984), "Effective Computability in Economic Decisions," University of Western Ontario working paper.

Monderer, D. and M. Tennenholtz (2006), "Strong Mediated Equilibria," Discussion Paper in the Department of Industrial Engineering, Technion, Israel Institute of Technology.

Naor, M. (1991), “Bit Commitment Using Pseudo-Randomness,” *Journal of Cryptology*, **4**(22), 151-158.

Nash, J. (1953), "Two-Person Cooperative Games," *Econometrica* 21,128-140.

Raiffa, H. (1953), "Arbitration Schemes for Generalized Two-Person Games," Annals of Mathematics Studies 28, ed. by Kuhn and Tucker, Princeton, 361-87.

Salop, S.C. (1986), "Practices that (Credibly) Facilitate Oligopoly Coordination," Analysis of Market Structure, Cambridge: MIT Press, 265-290.

Schelling, T.C. (1956), “An Essay on Bargaining," *The American Economic Review*, **46**(3), 281-306.

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

Tennenholtz, M. (2004), “Program Equilibrium”, *Games and Economic Behavior*, **49**, 363-373.

Urbano, A. and J.E. Vila (2002), "Computational complexity and communication: coordination in two-player games,' *Econometrica*, **70**, 1893-1927.

College of Computing, Georgia Institute of Technology

E-mail address: atk@cc.gatech.edu

Kellogg School of Management, Northwestern University

E-mail address: kalai@kellogg.northwestern.edu

School of Mathematical Sciences, Tel Aviv University and INSEAD

E-mail address: lehrer@tau.ac.il

Faculty of Management, Tel Aviv University

E-mail address: samet@tau.ac.il