Games and Economic Behavior 69 (2010) 127–137

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

A commitment folk theorem ☆

Adam Tauman Kalai [a,*], Ehud Kalai [b], Ehud Lehrer [c], Dov Samet [d]

[a] *Microsoft Research New England, Cambridge, MA, United States*
[b] *Kellogg School of Management, Northwestern University, Evanston, IL, United States*
[c] *School of Mathematical Sciences, Tel Aviv University and INSEAD, Israel*
[d] *Faculty of Management, Tel Aviv University, Israel*

ARTICLE INFO

Article history:
Received 19 March 2009
Available online 30 September 2009

JEL classification:
C70

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 establish the following commitment folk theorem: For any two-person strategic game G, there is a set of voluntary commitment devices $\mathfrak{D}$, such that the Nash equilibria of the game with commitments, $G^{\mathfrak{D}}$, span all the individually-rational correlated-strategy payoffs of G. In particular, in a decentralized manner the players may voluntarily 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 repetition in order to avoid conflicts (à la prisoner's dilemma) between cooperative and noncooperative solutions. The availability of a sufficiently rich set of voluntary commitment devices is enough to resolve all such conflicts. Also, such commitment devices can replace and strictly improve on the correlation devices of Aumann (1974, 1987), which are Pareto superior to simple Nash equilibria but fail to achieve full efficiency.

The goal of this paper is to map out the mathematical possibilities of commitment devices. For this reason the commitment devices we use are mathematical constructs, which may not be "natural" in real life applications. The determination of whether a device is natural, which may involve issues from psychology and other fields, would require a different type of analysis that is left for future work.

While the illustration of the above folk theorem requires nothing beyond elementary mathematics, it introduces two modeling innovations. First, it avoids pitfalls and circularities of conditional commitments by incorporating into the model a simple yet well-defined notion of commitment space. Second, in order to obtain the full generality, especially in games that

☆ This paper replaces "Meta-Games and Program Equilibrium." The research of Adam Tauman Kalai and Ehud Kalai is partially supported by the National Science Foundation grant no. SES-0527656 and by BSF grant no. 2004092.

* Corresponding author.
E-mail addresses: adum@microsoft.com (A.T. Kalai), kalai@kellogg.northwestern.edu (E. Kalai), lehrer@post.tau.ac.il (E. Lehrer), samet@tau.ac.il (D. Samet).

0899-8256/$ – see front matter
doi:10.1016/j.geb.2009.09.008

have no Pareto-efficient *pure-strategy* payoffs (unlike the prisoner's dilemma, for example) and pure max–min strategies, our commitment space uses 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, the equilibria of the commitment game give rise to the same set of payoffs as the repeated-game *folk* theorem. Second (and again in parallel to repeated games), results of the above type are known to many authors in different contexts. The earlier literature, however, only established partial folk theorems in special restricted contexts, while the current paper presents a complete folk theorem with no restrictions on the underlying games. Next, we 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, 1960). For example, when a player in a game delegates his play to an agent, with an 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) studies implications of strategic delegation in economic applications. Fershtman et al. (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 also display the use of more sophisticated, *conditional* commitment devices. For example, when placing an advertisement that states "we will sell X at a price of \$500, but will match any competitor's price," a retailer commits to a conditional pricing strategy. Such conditional commitment may be highly efficient. For example, in oligopoly pricing games match-the-competitors clauses make the monopolist price 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.

More recently, Tennenholtz (2004) introduced a sophisticated model of conditional commitments, called program equilibrium. In his model, two players simultaneously submit computer programs. The program of player A reads the program of player B and outputs a (mixed) strategy for A, and in a symmetric manner the program of player B outputs a strategy for B. The output strategies are then played, and the players receive the corresponding realized payoffs. Tennenholtz shows that program equilibria, which consist of pairs of programs that are best response to each other, lead to more efficient payoffs than what is possible under a Nash equilibrium of the unmodified game. However, these equilibria are short of reaching efficiency.[1]

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 advertisement 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 advertisements, because each of the prices should be \$50 lower than the other price.

Another example of the failure of conditioning to achieve a desirable outcome is the prisoner's 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 contradictions and circularities as above, he constructed hierarchical spaces in which higher levels of commitments are defined inductively over lower ones.[2] However, perhaps due to its complexity, Howard's model of conditional commitments did not lead to many applications.

The current paper goes beyond earlier models of commitments in several important ways. First, our model presents a simple yet powerful modeling approach that avoids the definability difficulties of conditional commitments. Second, each of our players has the explicit option not to commit and to play the game as is. And unlike the restricted models and partial folk theorems of the earlier literature, we present a general and complete folk theorem.

[1] Since we wrote this paper, we learned of an independent related study by Monderer and Tennenholtz (2006). In this paper the authors consider an extension of a game similar to our device, but focus on strong equilibria rather than a folk theorem.

[2] Klemperer and Meyer (1989) and Epstein and Peters (1999) present more advanced models that deal with additional economic issues.

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. In the newspaper advertisement game, for 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 result from advertisements of the opponent). This subtle difference is essential for well-defined outcomes.[3]

2. A model of commitment devices

In what follows we restrict ourselves to a fixed two-person finite strategic game, defined by a triple $G \equiv (N = \{1, 2\}, S = S_1 \times S_2, u = (u_1, u_2) : S \to \mathbb{R}^2)$.[4]

$N = \{1, 2\}$ is the set of players, each S_i is a non-empty finite set that describes 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 above, which is a particular 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, $u_i(s) = u_i(s_i, s_{-i}) \geqslant u_i(\bar{s}_i, s_{-i})$, for any alternative strategy $\bar{s}_i$ 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 $i \in N$, $u_i(\gamma) \geqslant \min_{\sigma_{-i}} \max_{\sigma_i} u_i(\sigma_1, \sigma_2)$. 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 *min–max 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 this device plays against the device d_{-i} of the opponent.

However, to ease the discussion we use a more compact representation for the response functions. The responses of the various devices of player i are aggregated into one *(grand) response function* $R_i : D_1 \times D_2 \to S_i$. So in effect, $R_i(d_i, d_{-i}) = r_{d_i}(d_{-i})$ describes the strategy chosen by the device d_i of player i in response to the possible device d_{-i} of the opponent. Put together, the two response functions describe a *joint response function* $R(d_1, d_2) = (R_1(d_1, d_2), R_2(d_1, d_2)) = (r_{d_1}(d_2), r_{d_2}(d_1))$, where $R(d_1, d_2)$ describes the pair of strategies selected by the devices when they respond to each other.

Note, however, that any function $R: D_1 \times D_2 \to S$ 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 $\mathfrak{D} \equiv (D = D_1 \times D_2, R: D \to S)$.

Each D_i is a non-empty 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 $G^{\mathfrak{D}}$).

The sets of devices D_1 and D_2 may be infinite. In that case we assume that they have a measurable structure, and that R is measurable with respect to the product structure. This enables us to define mixed strategies and expected payoffs in the game $G^{\mathfrak{D}}$.

[3] Insisting on well-defined outcomes is not just needed for proper mathematical modeling. In real life, players make costly efforts to hire agents, such as lawyers, to be sure that agreements lead to well-defined consequences.

[4] We restrict the model to two players for the sake of simplicity. We discuss the general case in Section 6.2. The extension of our results to games with infinite sets of strategies is straightforward.

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 $G^{\mathfrak{D}}$.[5]

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* (noncommittal) 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 $s_i, \bar{s}_i \in S_i$, $r_{d_{-i}}(s_i^{\mathfrak{D}}) = r_{d_{-i}}(\bar{s}_i^{\mathfrak{D}})$.

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

3. Examples

A trivial example of a voluntary commitment space is the game itself, with each $D_i = S_i$ and $G^{\mathfrak{D}} = G$. But all the examples discussed in the introduction, delegation to agents, newspaper advertisements, 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 the per-unit production cost is zero, 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 advertisements posted in Friday's newspaper. To fit into the formal model above, we let $D_i (= D_1 = D_2)$ describe all the advertisements that retailers are allowed to post. The newspaper (or some other legal entity) should verify that the advertisements lead to well-defined prices $R(d_1, d_2)$ (disallowing vague advertisement like "we will undercut opponents' prices by \$50," which fail to specify a response price to an identical competitor's advertisement).

Notice that the construction of a well-defined space of permissible advertisements is not as difficult as it may seem. For example, as we often see in newspapers, such advertisements are of the type: "We will meet any *posted* price of our competitors." Consistent with our model, a restriction to what is posted is a commitment to the text of the advertisement, as opposed to the price computed from the advertisement. More specifically, advertisement of this type may be viewed as devices that consist of a pair of items (p_i, h_i) that describe the posted price and the response rule. If retailers 1 and 2 place the advertisements $d_1 = (p_1, h_1)$ and $d_2 = (p_2, h_2)$ then the selling prices are $R(d_1, d_2) = (h_1(p_2), h_2(p_1))$.

Example 2 *(Divorce settlement).* Two players, he and she, are engaged in a divorce settlement. Each has two possible strategies: cooperative (c) and aggressive (a). The payoffs are the same as in the standard prisoner's dilemma.

But assume now that each player has the option of being represented in the game by a lawyer of her choice, 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, *tl*'s choose the strategy a. But *fl*'s choose the strategy c when they face an opponent of type *fl*, and choose the strategy a against all other opponents.

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

[5] In the main result, we employ a stronger notion, where we use the same device space $\mathfrak{D}$ for all the equilibria (rather than a different device space of each equilibrium).

		Pl 2			
		fl	*tl*	$c^{\mathfrak{D}}$	$a^{\mathfrak{D}}$
Pl 1	*fl*	c, c	a, a	a, c	a, a
	tl	a, a	a, a	a, c	a, a
	$c^{\mathfrak{D}}$	c, a	c, a	c, c	c, a
	$a^{\mathfrak{D}}$	a, a	a, a	a, c	a, a

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 being aggressive towards 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 unconditionally (no matter what device is chosen by the opponent). Moreover, his choice is private, as none of the devices of Player 2 (in choosing an action for Player 2) ever differentiates 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 prisoner's dilemma payoffs in the sixteen cells in the table, it is easy to see that for both players fl is a weakly dominant strategy and therefore the pair (*fl*, *fl*) is an equilibrium. In effect, this equilibrium employs a tit-for-tat type of strategy to get cooperation in this one-shot prisoner's dilemma game: a player deviating from *fl* causes the opponent's device to switch from c to a.

4. A commitment folk theorem

4.1. Technical subtleties

Unlike Example 2, where the construction of a cooperative commitment equilibrium is easy, the proof of a general folk theorem is more difficult, as explained in the next two examples.

Example 3. (Fight-or-relinquish)

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

Here, the min–max strategies guarantee each player a payoff of at least 2. But unlike in the standard prisoner's dilemma game, there is no pair of pure strategies that simultaneously yield each player a payoff greater than 2. Yet, a full folk theorem should have Nash equilibria generating every payoff profile in the convex hull of $\{(2,2),(2,8),(8,2)\}$, which is the set of all individually rational and feasible payoffs, for example $(5,5)$.

In the repeated-game folk theorem convex combinations of payoffs are achieved by the players alternating between the cells (*fight*, *relinquish*) and (*relinquish*, *fight*), and a trigger strategy will induce the correct incentives to do so. However, such alternations are impossible if the game is played only once.

A second difficulty to overcome 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

This game is similar to the fight-or-relinquish game above, where in any cooperative outcome one of the players relinquishes. 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 min–max payoffs, are generated by pure strategies. Thus, when triggering to punishment 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 observed by the opponent's device, which would know whether the player plans to trigger it with the pure-strategy *style A* or with the pure-strategy *style B*.

There is 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 punishing 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 alternations of an infinitely-repeated game by a 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.[6]

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) Nash equilibrium of the commitment game $G^{\mathfrak{C}}$.

Notice that in the above definition, there is one (universal) device space that is used by all the equilibria.

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 a space $\mathfrak{C}$ with a continuum of devices, to be used as strategies in the commitment game $G^{\mathfrak{C}}$. The strategies of a player i are 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 $\{1,2,\dots,M\}$. The first two parts of a strategy are points in Euclidean spaces and the third one is an element from a finite set. Each of these parts has a standard measurable structure. We endow the set of strategies with the induced measurable structure.

We now describe a method for encoding any correlated strategy γ over S by a unique $x \in \Delta_M = \{x \in [0,1]^M \mid \sum_i x_i = 1\}$, 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 $g : \Delta_M \times [0,1] \to [M]$,

$$g(x,r) = \min\{j \in [M] \mid x^1 + x^2 + \dots + x^j \geqslant r\},$$

for every $x = (x^1,\dots,x^n)$. 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 $\mathfrak{C} = (D = D_1 \times D_2, R)$. $D_i = (\Delta_M \cup \{\perp\}) \times [0,1] \times S_i$. The special symbol $\perp$ is necessary to make the game voluntary, and indicates that the player wants to play the fall-back strategy, and R is defined by,

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

The expression $r_1 + r_2 - \lfloor r_1 + r_2 \rfloor$ above computes the fractional part of $r_1 + r_2$. It is easy to see that R is measurable.

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 min–max 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 min–max distribution of his opponent, which can only decrease his payoff.

The game is voluntary because player i has a 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 Appendix A of 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.

[6] 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 in the concluding section of the paper.

5. Comparison with earlier notions

5.1. Comparison with meta strategies

Earlier attempts to deal with sophisticated conditional commitments (without the use of well-defined commitment device spaces) lead to highly complex 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 prisoner's 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's level t rules.

The model of the current paper offers a simpler and 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.

5.2. 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 $(2,2)$, (because fight is a strongly dominant strategy), whereas any payoffs in the convex hull of $\{\{(2,2),(2,8),(8,2)\}\}$ (including $(5,5)$) can be obtained at commitment equilibrium.

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.[7] The players proceed to play G after learning their private messages. Once a player receives a signal he has no way of affecting the other players' strategies.[8]

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.

5.3. Comparison to delegation

The delegation folk theorem presented in Fershtman et al. (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. For example, in the game with fighting options above, which has no pure strategy equilibria, their delegation folk theorem is vacuous.

5.4. 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 game with fighting options, 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, whereas 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 set of equilibrium payoffs, rather than the smaller set in fact obtained. 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.

---

[7] We use the notion of common knowledge loosely here. As readers familiar with the literature are aware, less than full common knowledge suffices in the statements made here.

[8] To generate the probability distribution of a correlation device one needs an external impartial mediator, or, alternatively, a system of devices that produces signals 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).

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 min–max 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 to obtain 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 randomize.

6. Additional remarks

6.1. On natural commitment devices and implementation

We already commented on the need to study *natural* commitment devices. A related question is, where do commitment devices come from? Is there an outside entity (other than the players of the game) able to construct commitment spaces for the players, or are commitment devices something the players generate on their own? Under the former case, the study of commitment may be viewed as a subarea of the implementation literature, see the survey of Jackson (2001). And under the latter case, the study of commitments may lead to deeper issues of the evolution of language and vocabulary that players may come up with in order to communicate their commitments.

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.[9]

6.2. Extensions to n players

For the sake of simplicity we deal here with the two-player case. However, the same technique shows that any feasible, individually rational payoff (when punishing may be correlated between the players) can be obtained as a commitment equilibrium. The idea is that in a case of deviation, the punishing players jointly control the action used and thereby correlate the punishment.

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 of communication.

Example 5. A treasure is buried in one of three locations L_1, L_2 and L_3 with equal probability. It takes two to dig for it. Player 1 lives in L_1 and if the treasure is buried there she knows it. Likewise, player 2, who live in L_2, knows if the treasure is buried there. The players cannot communicate, but they can move simultaneously to any location. If they happen to meet at the location of the treasure they dig for it.

If each player stays in her location when the treasure is there, and moves to the location of the other player otherwise, then they will meet at the location of the treasure whenever it is buried in either L_1 or L_2, and thus, the chance they will dig the treasure it 2/3. To see that they cannot improve on 2/3, note that the chances of a player to reach the location of the treasure, let alone meeting there the other player, cannot be higher than 2/3. Of course, if the players are allowed to communicate, then they dig for the treasure with probability 1.

Using this game we show that commitment devices can serve as effective communication devices in games with differential information.

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 L_i no matter what device is used by the opponent. The device f_i chooses the location L_{-i} against the device s_{-i} of the opponent, but chooses L_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.

When dealing with commitments in Bayesian games, there are several modeling alternatives. For example, are the individual commitments done before or after the private information is revealed.

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

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

of a game. On the other hand, Fershtman and Kalai (1993) argued 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 (possibly infinite) number of contracts 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 will 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 natural and effective commitment devices of the type discussed in this paper.

But from the game theoretic perspective, 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 the battle of the sexes) game described below:

		Pl. 2	
		insist	*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 that guarantees a 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 will play (*insist*, *yield*), and if it shows T, they will play (*yield*, *insist*).

But the use of an outside randomizing mediator may be avoided through the use of a strategic contract. For example, each player submits a sealed envelope with an integer $s_i = 1$ or 2 together with a contract that states that if their integers match, they will play (*insist*, *yield*) and if their integers mismatch they will 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.[10]

Acknowledgment

This work was supported in part by NSF award SES-0734780 and by BSF grant no. 2004092.

[10] The submission of a 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 contract 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.

Appendix A

A.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 $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$. While this does not give a full folk theorem, it is sufficient for many practical purposes and has the advantage of being finite.

Theorem 2 *(Finite commitment-device folk-theorem). For the two-player game G and any $n \geqslant 1$, 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 $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ 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 $r_i \in \{\frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ uniformly at random. Such numbers are represented using $O(\log n)$ bits. It is straightforward to efficiently compute the function L_n, 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 with rational number payoffs the following two conditions are equivalent:*

1. *There exists a finite voluntary 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 set of feasible and individually rational payoffs of G, denoted FIR is a rectangle with facets parallel to the axes.*

Proof. Assume that (2) holds. Suppose first that there are four payoffs in the game which are the extreme points of a rectangle that contains *FIS*. 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. Denote the players' strategies in this game by H and T and its joint response function (recall Definition 1) by R. This game is not a voluntary commitment game.

The strategies of a player i in the voluntary commitment game are pairs in $\{\perp, H, T\} \times S_i$, and the joint response function, L, is defined by,

$$L\big((x, s_1), (y, s_2)\big) = \begin{cases} R(s_1, s_2) & \text{if } x, y \in \{H, T\}, \\ (s_1, s_2) & \text{otherwise.} \end{cases}$$

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

Let P_i be the mixed strategy by which player i min–maxes player $-i$. Fix $0 \leqslant \alpha \leqslant 1$ and denote by σ_i^t the strategy of player i in the commitment game by which he chooses (H, s_i) with probability $\alpha P_i(s_i)$ and (T, s_i) with probability $(1-\alpha)P_i(s_i)$. Note that $(\sigma_1^\alpha, \sigma_2^\beta)$ is an equilibrium for every $\alpha, \beta \in [0, 1]$. Moreover, the payoffs corresponding to these equilibria cover all the set of feasible and individually rational payoffs.

The proof so far hinged on the existence of four payoffs in G which are the extreme points of a rectangle that contains *FIS*. If such points do not exist, we replace them by matrices that consist of payoffs that do exist. Since *FIS* is a rectangle whose facets are parallel to the axes, there are three payoffs that are extreme points of a triangle that contains *FIS*. Denote by A the northeastern point, by B the northwestern and by C the southwestern point. Thus, A and B are player 2 payoff equivalent, and C and B are player 1 payoff equivalent.

There are numbers $\gamma, \delta \in [0, 1]$ such that *FIS* is in the rectangle whose extreme points are (clockwise, starting on top-left) $\gamma A + (1-\gamma)B$, B, $\delta C + (1-\delta)B$ and $\delta C + (1-\delta)A$. Since the payoffs of G are rational,[11] γ, δ can be chosen to be rational, say $\frac{k}{n}$ and $\frac{\ell}{n}$, respectively, with k, ℓ, n being natural numbers. The 2×2 device game above is replaced by $2n \times 2n$ device game that consists of 4 $n \times n$ sub-matrices with the outcomes A, B and C.

The top-left matrix contains in each row and in each column k A's and $n-k$ B's (i.e., it is a Latin square). The top-right sub-matrix is constantly B, the bottom-right matrix contains in each row and in each column ℓ C's and $n-\ell$ B's, and finally, the bottom-left matrix contains in each row and each column ℓ C's and $n-\ell$ A's. If each player plays uniformly

[11] This is the only point where this assumption is used.

over each sub-matrix, the effect is exactly that of the 2×2 game above. The extension of this game to a voluntary game is similar to what we have shown above, and is therefore omitted.

As for the converse, assume (1) and that (contrary to (2)) one of the facets of G's feasible and individually rational 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 $\mathfrak{C}^{\sigma} = (D_1^{\sigma} \times D_2^{\sigma}, L)$ 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 $\mathfrak{C}' = (D_1' \times D_2', L)$ of $\mathfrak{C}$ where all payoffs of $G^{\mathfrak{C}'}$ are in F. By a positive affine transformation of the payoffs of the players, $G^{\mathfrak{C}'}$ can be transformed either to a zero-sum game or to a coordination game, say G.

If G is a zero-sum game, it has only one equilibrium payoff. In particular, all full-support equilibria of G induce the same payoff. Since G is derived from $G^{\mathfrak{C}'}$ by positive affine transformations, any full-support equilibrium of G is a full-support equilibrium of $G^{\mathfrak{C}'}$.

Assume now that G is a coordination game. Denote by G_0 the zero-sum game derived from G by making player 2 a minimizer, rather than a maximizer. Any full-support equilibrium of G is an equilibrium of G_0 and since there is only one equilibrium payoff in G_0, there is only one full-support equilibrium in G.

Consequently, any $G^{\mathfrak{C}'}$ has only one full-support equilibrium payoff. Since there are finitely many subgames 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 strategy payoffs in F. This contradiction leads to the conclusion that if $G^{\mathfrak{C}}$ is finite, then all the facets *FIS* are parallel to the axes. $\square$

References

Aumann, R.J., 1974. Subjectivity and correlation in randomized strategies. J. Math. Econ. 1, 67–96.

Aumann, R.J., 1987. Correlated equilibrium as an expression of Bayesian rationality. Econometrica 55 (1), 1–18.

Aumann, R.J., Maschler, M., 1995. Repeated Games with Incomplete Information. MIT Press, Cambridge.

Barany, I., 1992. Fair distribution protocols or how players replace fortune. Math. Oper. Res. 17, 327–340.

Ben-Porath, E., 1998. Correlation without mediation: Expanding the set of equilibrium outcomes by cheap pre-play procedures. J. Econ. 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., Peters, M., 1999. A revelation principle for competing mechanisms. J. Econ. Theory 88, 119–161.

Fershtman, C., Judd, K., 1987. Equilibrium incentives in oligopoly. Amer. Econ. Rev. 77 (5), 927–940.

Fershtman, C., Judd, K., Kalai, E., 1991. Observable contracts: Strategic delegation and cooperation. Int. Econ. Rev. 32 (3), 551–559.

Fershtman, C., Kalai, E., 1993. Unobserved delegation. Int. Econ. Rev. 38 (4), 763–774.

Gossner, O., 1998. Secure protocols or how communication generates correlation. J. Econ. 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 Welfare 18 (4), 655–708.

Kalai, E., 1981. Preplay negotiations and the prisoner's dilemma. Math. Social Sci. 1 (4), 375–379.

Kalai, E., Rosenthal, R.W., 1978. Arbitration of two-party disputes under ignorance. Int. J. Game Theory 7 (2), 65–72.

Kalai, E., Samet, D., 1985. Unanimity games and Pareto optimality. Int. J. Game Theory 14 (1), 41–50.

Kalai, E., Satterthwaite, M., 1986. The kinked demand curve, facilitating practices and oligopolistic competition, DP 677. Center for Math. Studies in Econ. and Mgt. Science;
Published also in: Gilles, R.P., Ruys, P.H.M. (Eds.), 1986. Imperfection and Behavior in Economic Organizations. Kluwer Academic Publishers, Boston, pp. 15–38.

Katz, M.L., Shapiro, C., 1985. Network externalities, competition, and compatibility. Amer. Econ. Rev. 75 (3), 424–440.

Klemperer, P.D., Meyer, M.A., 1989. Supply function equilibria in oligopoly under uncertainty. Econometrica 57 (6), 1243–1277.

Lehrer, E., 1996. Mediated talk. Int. J. Game Theory 25, 177–188.

Lehrer, E., Sorin, S., 1997. One-shot public mediated talk. Games Econ. Behav. 20 (2), 131–148.

Monderer, D., Tennenholtz, M., 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. J. 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. In: Kuhn, H.W., Tucker, A.W. (Eds.), Annals of Mathematics Studies, vol. 28. Princeton Univ. Press, Princeton, pp. 361–387.

Salop, S.C., 1986. Practices that (credibly) facilitate oligopoly coordination. Analysis of Market Structure. MIT Press, Cambridge, pp. 265–290.

Schelling, T.C., 1956. An essay on bargaining. Amer. Econ. Rev. 46 (3), 281–306.

Schelling, T.C., 1960. The Strategy of Conflict. Harvard Univ. Press, Cambridge, MA.

Tennenholtz, M., 2004. Program equilibrium. Games Econ. Behav. 49, 363–373.

Urbano, A., Vila, J.E., 2002. Computational complexity and communication: Coordination in two-player games. Econometrica 70, 1893–1927.