Journal of Economic Behavior and Organization 21 (1993) 295–314. North-Holland

A game-theoretic rendering of promises and threats*

Daniel B. Klein

University of California, Irvine, CA, USA

Brendan O'Flaherty

Columbia University, New York, NY, USA

Received July 1991, final version received January 1992

The paper formalizes several of the ideas about commitment set out by Thomas Schelling in *The Strategy of Conflict*. Using a game-theoretic framework we formalize and interpret 'promise' and 'threat' as different species of commitment. We also distinguish the 'pure promise', the 'pure threat', and the 'hybrid'. Using examples we discuss the friendliness of promises and the unfriendliness of threats.

Game theorists know an incredible threat when they see one. When the union says to management, 'We'll strike if you don't meet our demands,' and the payoffs in the subgame as originally written suggest 'Don't strike' rather than 'strike', the game theorist will call it an 'incredible threat'. He may observe that the threat equilibrium is not subgame perfect. But Nash equilibria that fail subgame perfection do not always involve a threat, and not every threat is part of such an equilibrium. The game theorist has no general characterization of threats.

Furthermore, the adjective 'incredible' is a bit mysterious. In finite noncooperative games, when a strategy announcement is 'credible', by the usual usage, it is not a threat. Consider the following example: 'If you [the ordinary Joe] attack me [Superman], I will defend myself.' The claim certainly seems credible, but would you call it a threat? Thomas Schelling would not; his term is 'warning' (1960, p. 123, see also p. 35).

Promises are still more problematic. The term 'promise' is used less frequently than 'threat' because promises do not fit anywhere in Nash

Correspondence to: Daniel B. Klein, Department of Economics, University of California, Irvine, CA 92717, USA.

*For valuable discussion the authors wish to thank Tyler Cowen, Ami Glazer, Joseph Harrington, Greg Kavka, Randy Kroszner, Gary Ramey, Stergios Skaperdas, John Van Huyck and an anonymous JEBO referee.

solution concepts. Occasionally game theorists speak of 'incredible promises' in infinite games [e.g., Kreps (1990, pp. 50–53)]. But the term remains mysterious. In a finite game what would a credible promise look like? Consider the following example: 'If you bake me my favourite chocolate cake, I will eat it.' Again, credibility is not in doubt, but one would hardly call this a promise.[1]

In this paper we offer general characterizations of 'promise' and 'threat', as well as of 'pure promise' and 'pure threat', and we need to work outside the standard noncooperative vocabulary to do so. Promises and threats are portrayed as different species of plans announced by a 'ruler' who is able to convey commitments before the play of the game. Elsewhere we have developed a vocabulary to describe and analyze commitment (1992, 1993). Here we use this vocabulary to arrive at general formulations of promise and threat. These formulations conform closely to Thomas Schelling's notions of commitment, promise, and threat as exposited in *The Strategy of Conflict*. In our formulation the question of 'credibility' is sidestepped; 'credibility' plays only a secondary role. Regarding the beliefs of the public – the set of players other than the ruler – center-stage is taken by 'credulity'.

People may make pre-play utterances for several reasons. Three stand out: (a) information disclosure ('I love chocolate cake'), (b) cheap-talk coordination goals ('Meet met at the ballet'), and (c) strategic influence (in Schelling's sense). The subject here is (c). As Schelling said about a strategic move, 'it must involve some notion of commitment – real or fake – if it is to be anything' (p. 127).

A declaration (also called 'announcement' or 'plan') can have strategic influence if the public interprets *the act of declaration* as *altering* the ruler's payoffs. Specifically, the ruler conveys selective subtractions from her own endnode payoffs. This is Schelling's notion of commitment and is further explored below. Thus, in this paper we designate one player as 'the ruler' who conveys a commitment to one of her strategies in an extensive form game (the 'reference game'). The public believes these commitments and responds in a passive fashion.

Part 1 of the paper is 'Interpretations'. Part 2 is 'On the Friendliness of Promises and the Unfriendliness of Threats.' Part 3 is 'Formal Characterizations.'

1. Interpretations

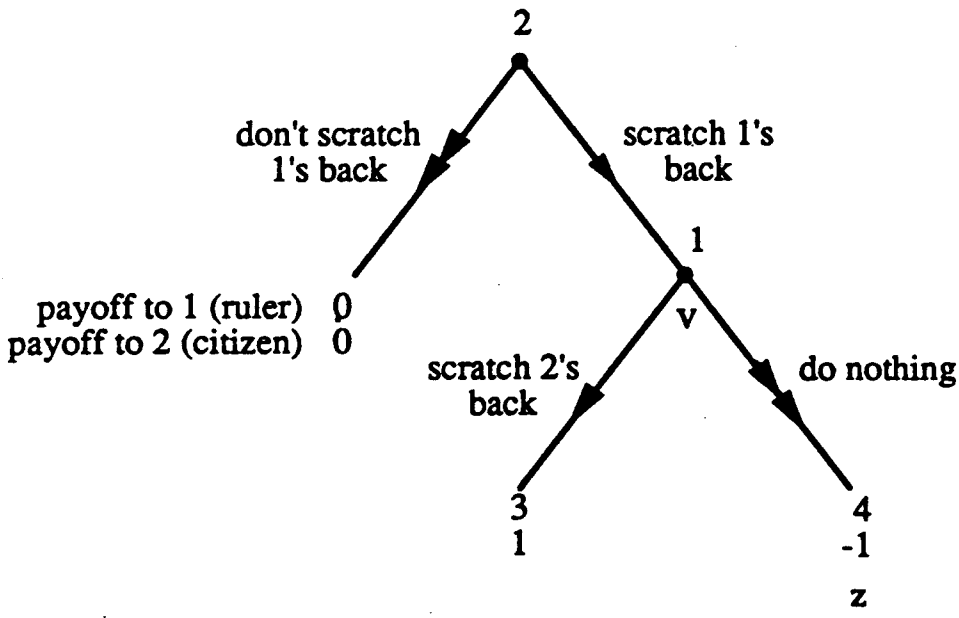
Canonical examples of promise and threat

'If you scratch my back I'll scratch your back,' announces the ruler in fig. 1. (Throughout this paper Player 1 is the ruler.) She had better make the

[1]Promises and threats are credible in knife-edge situations, as in Nalebuff and Shubik (1988).

Pure Promise

"If you scratch my back I'll scratch your back."



Single arrows show the promise and the response made by a credulous citizen. The plan is time inconsistent, because at v player 1 (the ruler) would like to switch to (do nothing).

Fig. 1

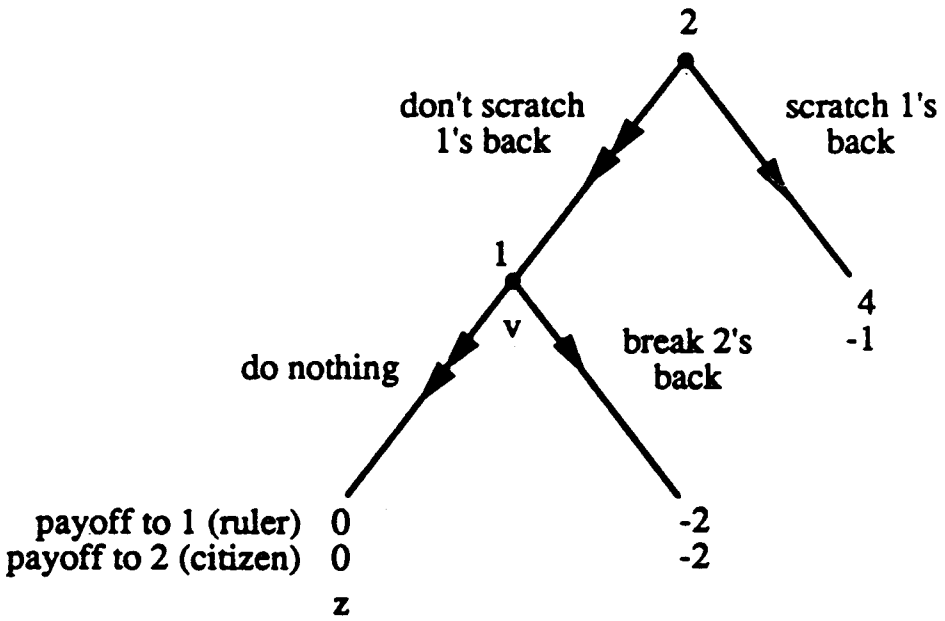
promise convincing, for otherwise Player 2 will regard it as hot air and therefore not scratch the ruler's back. Clearly the ruler values commitment conveyance – that is, the ability to get the public to believe the announcement – since without it her back goes unscratched. (Unless otherwise stated, we assume that the ruler has commitment conveyance.)

Fig. 1 is the canonical pure promise. Schelling offers many examples of it: 'The witness to a crime [Player 1] has a motive for unilateral promise if the criminal [Player 2] would kill to keep him from squealing. A nation [Player 1] known to be on the threshold of an absolutely potent surprise-attack weapon may have reason to forswear it unilaterally ... in order to forestall a desperate last-minute attempt by an enemy [Player 2] to strike first while he still has a chance' (pp. 132–133).

Fig. 2 depicts the second canonical way of getting your back scratched: the ruler says 'Scratch my back or else I'll break your back.' Again, the ruler needs to make the plan convincing, for otherwise Player 2 will not oblige.

Pure Threat

"Scratch my back or else I'll break your back."



Single arrows show the threat and the credulous response. The plan is sequentially irrational but time consistent.

Fig. 2

This plan is a threat. As Schelling says, 'the threat's efficacy depends on the *credulity* of the other party, and the threat is ineffectual unless the threatener can *rearrange or display* his own incentives *so as to* demonstrate that he would, *ex post*, have an incentive to carry it out' (our italics, p. 36).

Schelling pointed out the essential difference between the pure promise and the pure threat: 'The difference is that a promise is costly when it succeeds, and a threat is costly when it fails. A successful threat is one that is not carried out' whereas a successful (and genuine) promise is carried out (177). The promise in fig. 1 is time inconsistent: the ruler reaches a decision node where she would like to renege on the announcement if she could renege without penalty. The threat in fig. 2 is time consistent: the ruler never reaches a node where she would like to renege even if she could do so without

penalty. These distinctions are advanced also by Guiso and Terlizzese (1990).[2]

Schelling's notion of commitment as selective subtractions

Schelling says 'a commitment ... can usually be characterized in a fashion equivalent to the following: ...a player selectively reduces – visibly and irreversibly – some of *his own* payoffs' (his italics, p. 150, see also p. 129). In fig. 1 the ruler conveys a commitment to [scratch 2's back]. Following Schelling, we would say that she convinces Player 2 that she has subtracted at least one unit from her own payoff at terminal node *z*, thereby making the plan believed. In fig. 2, with optimal plan [break 2's back], the ruler convinces Player 2 that she has subtracted at least two units from her own payoff at *z*.

Schelling says, 'it is probably best to consider the threat and the promises to be names for different aspects of the same tactic of selective and conditional self-commitment' (p. 134). When, in common parlance, 'an agreement is struck,' or someone 'gives their word,' or 'makes a promise' there is some confidence that the act of declaration has real emotional or psychological meaning to the promise-maker. All these phrases are ways of saying that the promise-maker has subtracted utility from certain actions which otherwise she would have a temptation to take. God-fearing promise-makers will stake God's grace on their word, others 'cross their heart and hope to die,' others stake their business reputation and others their personal honor. Robert Frank (1988, p. 132) nicely explains the research finding on the 'multitude of statistically reliable clues to the emotions people feel.' Facial expressions, body language, and the pitch of the voice are often telltale clues to intentions. Sincere promise-making is difficult to mimic and therefore often effective in assuring contracts.

Similarly, a threat is often more than mere noises emanating from a larynx. In the film *War of the Roses*, Mr. Rose (Michael Douglas) say to Mrs. Rose (Kathleen Turner): 'You will never get that house. Do you understand? You will never get that house.' It is a threat of fighting to the bitter end. The very act of declaration is soaked with emotional content. The incident alters some of Mr. Rose's own payoffs and his actions exemplify commitment.[3] Social psychologist Robert Cialdini (1984, pp. 66–114) explains that people fancy themselves consistent beings. We have an emotional response that tells us to follow through on our declarations, if tested. *Talk is not always cheap.*[4]

[2]For a related discussion in a different mode, see Gauthier (1991).

[3]In this example 'the bitter end' turns out to be along the path. Mr. Rose's threat did not work and he carried it out. This failure may be ascribed to Mrs. Rose's less-than-complete credulity or to her preference for war, which was not recognized by Mr. Rose.

[4]See O'Flaherty (1985, pp. 28–31) for more discussion of commitment as selective subtraction.

But Schelling and we are using the speech-acts of promising and threatening as synecdoches for strategic conveyance of selective subtractions from one's own payoffs. Neither the technology of the subtraction, nor of they conveyance, needs to be the act of writing or speaking. Schelling says, '[c]ollision being about as mutual as anything can be[,] ... the maneuvers by which one conveys a threat of mutual damage to another driver ... are an instructive example of the kind of threat that is conveyed not by words but by actions' (12). Similarly in the case of promises, to make selective subtractions from her own payoffs, the ruler need not make heartfelt declarations; she may simply post a bond.

Interpreting commitment conveyance

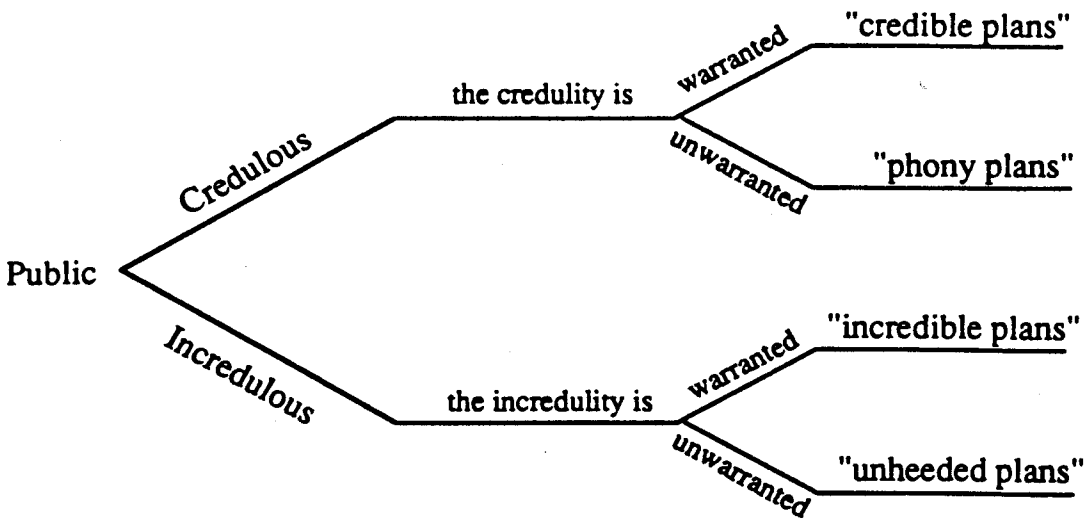
In the Schelling passage that began the previous subsection he says that the player making payoff subtractions must do so 'visibly and irreversibly.' Actually, 'apparently' is good enough and sometimes better. The fundamental requirement is that the public *believes* that the ruler has committed herself, not that she actually has. When someone actually does commit herself it is probably because that was the only way to make the others believe she has.

To talk about strategic commitment we need only the primary credulity of the public. Whether the public's belief in the ruler's announcement is warranted is a secondary matter. If the public is credulous we say that the ruler has commitment conveyance.

The credulity of the public and the genuineness of the plans give rise to four sorts of plan contexts, as shown in fig. 3. When the public is credulous, as assumed throughout this paper, the credulity can be either warranted (giving rise to 'credible plans') or unwarranted (giving rise to 'phony plans'). One might also be interested in the incredulous public, in which case the lack of credulity can again be warranted (giving rise to 'incredible plans') or unwarranted (giving rise to 'unheeded plans'). An interesting area for future investigation would be partial credulity.

'Credibility' in our analysis takes a backseat to 'credulity'. Our use of 'credibility' is different from the usual usage ('the union's threat to strike is incredible [subgame imperfect]'). 'Credibility' in our sense addresses the believability of pre-play selective subtractions (commitments), whereas the usual usage addresses the rationality of striking in the subgame as originally written.

If the ruler has phony commitment conveyance she will renege on the plan if the plan is time inconsistent. In fig. 1, if the ruler has phony commitment conveyance, she announces 'I'll scratch your back if you scratch my back,' but once her back is scratched she reverts to not scratching Player 2's back. In this case she benefits not only from being able to induce Player 2 to scratch her back but also from not having to deliver on the promise. In



Interpretations of Commitment Conveyance: When the public is credulous the ruler enjoys commitment conveyance. The ruler's commitments may be either genuine or phony.

Fig. 3

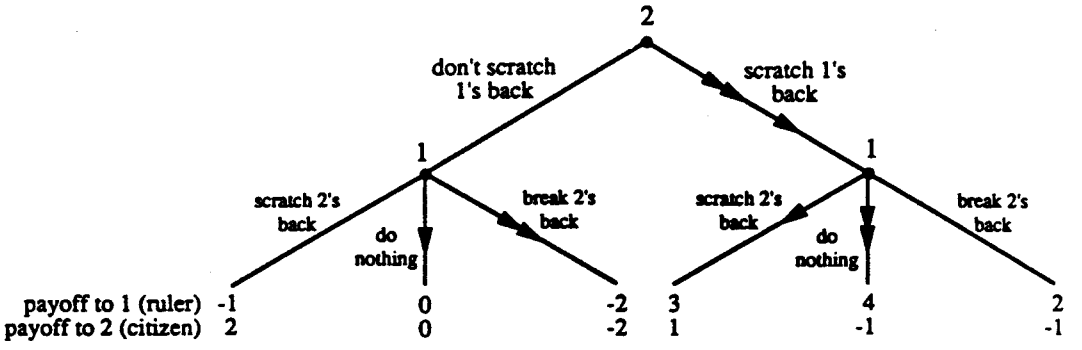
Fischer's model (1980) of savings-taxation in a representative-individual economy, phony commitment conveyance is the case of the 'benevolent dissembling government.'

Promises and threats are not strategically equivalent

Sometimes people say that there is no essential differences between a promise and a threat, that a promise is just a rephrasing of a threat, and vice versa. For the game in fig. 1, 'If you scratch my back I'll scratch you back' can be equivalently stated as 'Scratch my back or else I won't scratch your back.' The 'threatened' action in this case – not scratching – is *not* costly to the ruler when the 'threat' fails. Thus this 'threat' fails Schelling's criterion for threat. The rephrasing is still a promise, only awkwardly expressed.

Similarly, for the game in fig. 2, 'Scratch my back or else I'll break your back' can be rephrased as 'Only if you scratch my back will I refrain from breaking your back.' Like the Godfather's offer that his associates can't refuse, this has the ring of a promise but the 'promised' action in this case – not breaking – is *not* costly when the 'promise' succeeds, again failing Schelling's criterion. The rephrasing is still a threat, only awkwardly expressed.

Scenarios in figure 1 and figure 2 combined.
Promises and threats are not strategically equivalent.



There are two distinct ways in which the ruler can get her back scratched.
Single arrows shows the promise. Double arrows shows the threat.

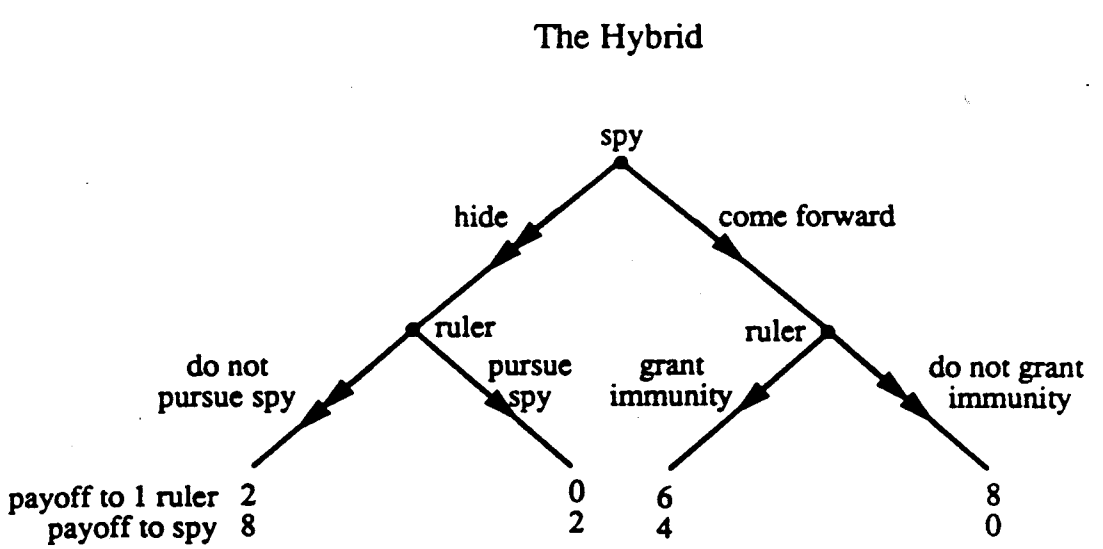
Fig. 4

Let's bring both back-scratching stories into a single figure. Fig. 4 encompasses both fig. 1 and fig. 2. The ruler has additively separable payoff function: 4 utils from having her back scratched, −1 util from having to scratch 2's back, and −2 utils from having to break 2's back. Player 2 has payoff function: 2 utils from having his back scratched, −1 util from having to scratch the ruler's back, and −2 utils from having his back broken. 'If you scratch my back I'll scratch your back' and its credulous response are shown by single arrows. 'Scratch my back or else I'll break your back' and its credulous response are shown by double arrows. These are two distinct ways by which Player 1 can induce Player 2 to scratch her back.

Hybrids

Schelling points out that '[o]ne cannot force spies, conspirators, or carriers of social diseases to reveal themselves solely by the threat of a relentless pursuit that spares no cost; one must also promise immunity to those that come forward' (p. 134). In fig. 5 the threat of pursuit combined with the promise of immunity induces the spy to reveal himself (single arrows). Just the threat or just the promise would not work.

In fig. 5 we can break down the ruler's plan into a local threat move (pursue) and a local promise move (immunity). In fig. 6 the matter is subtler. Consider the play shown by single arrows. The ruler is promising Player 3



The plan shown by single arrows is both a promise and a threat.

Fig. 5

that she will threaten Player 2 into going L. Notice that once Player 3 has responded to the promise by moving $[b]$ the ruler would like to renege on the commitment and reannounce $[f]$ at node v, inducing Player 2 to switch to $[R]$. The plan is time inconsistent. It is both a promise and a threat.

Because of cases like fig. 6 we use 'promise' and 'threat' as descriptions of *entire plans*, rather than as descriptions of local moves. The strategic relevance of a local move depends on what is specified elsewhere in the tree, so we need to keep an eye on the entire plan.

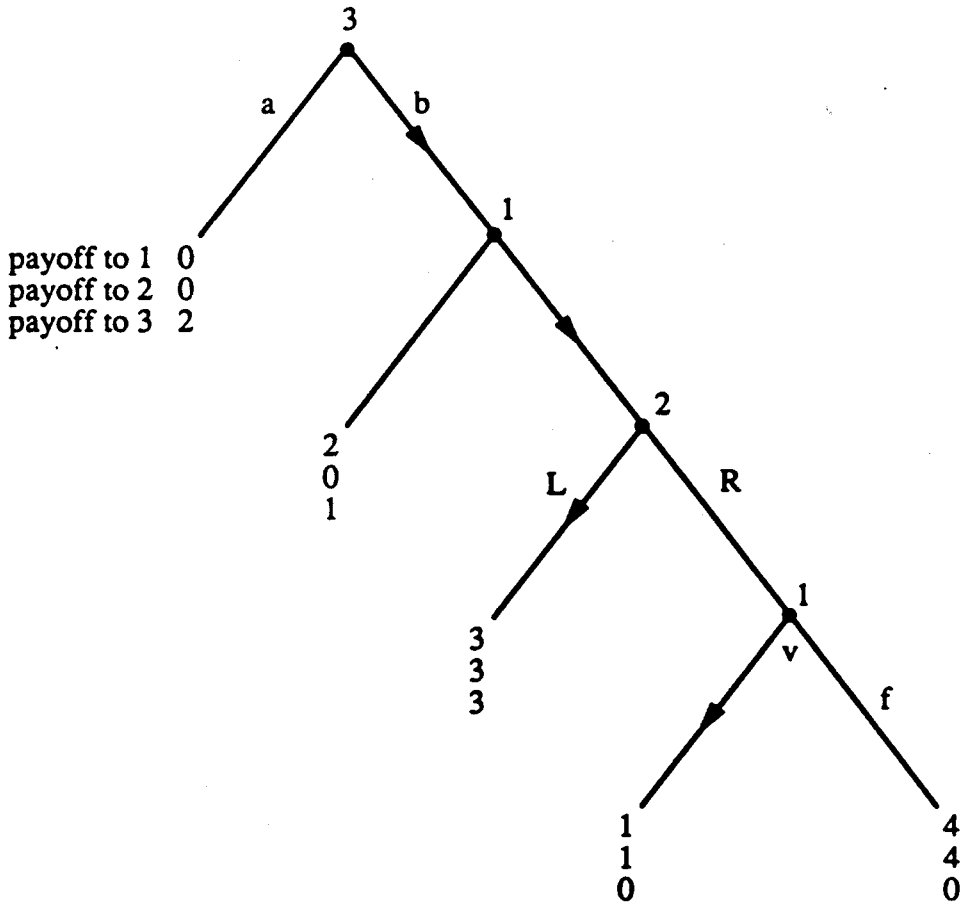
Because of hybrids it is useful to refine our notions of 'promise' and 'threat'. This refinement will be made rigorous in Part 3, but the intuition is fairly simple:

A promise is a plan that is time inconsistent, like 'If you scratch my back I'll scratch your back.' A threat is a plan that involves a strategic irrationality off the path of play, as in 'Scratch my back or else I'll break your back.' Notice that the hybrid play in fig. 5 is both a promise and a threat.

A pure promise is a promise that is not a threat. A pure threat is a threat that is not a promise. Thus the plan in fig. 5 is neither a pure promise nor a pure threat.

2. On the friendliness of promises and the unfriendliness of threats

In everyday language the term 'promise' has a friendly connotation.



The ruler is promising Player 3 that she will threaten Player 2.

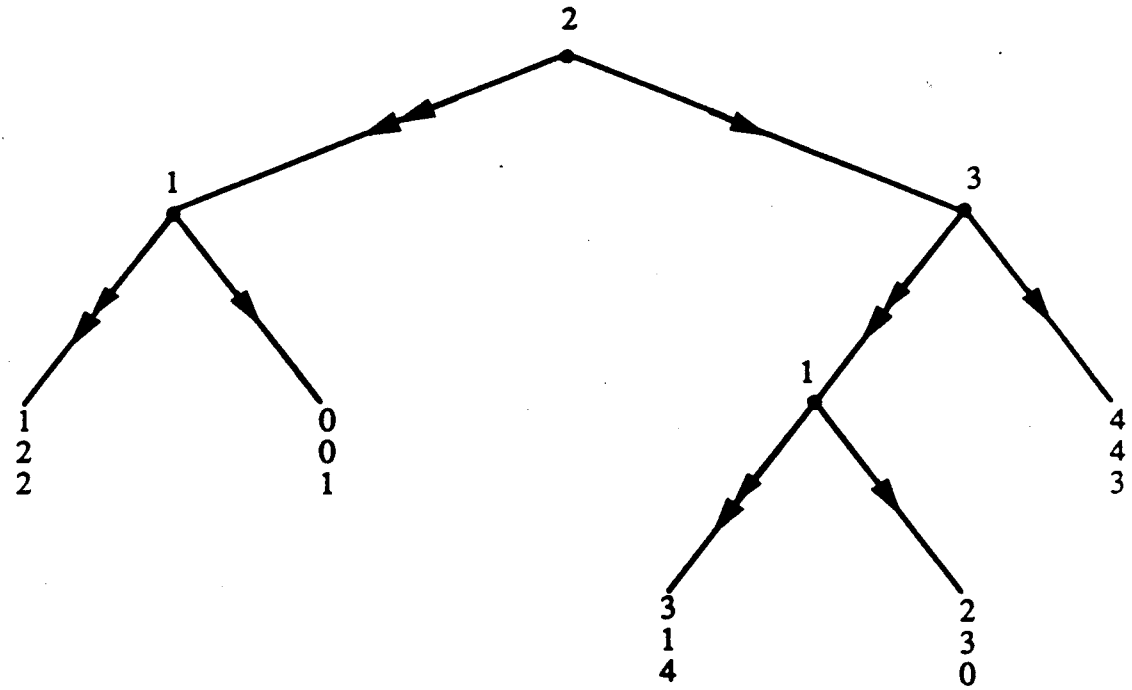
Fig. 6

Friends are in the habit of making each other promises, whereas enemies are in the habit of making each other threats.

Going back to the pure promise of fig. 1, now focus on the welfare of Player 2. Player 2 welcomes Player 1's commitment conveyance (let's assume it is genuine), since Player 2 get his back scratched and receives a payoff of 1 util. When Player 1 lacks commitment conveyance Player 2 receives a payoff of zero utils. The promise is friendly.

In contrast, the term 'threat' has an unfriendly connotation. In fig. 1 Player 2 curses Player 1's commitment conveyance. The threat compels Player 2 into unreciprocated scratching, yielding a payoff of -1 util. When Player 1 has no commitment conveyance Player 2 receives a payoff of zero utils. The threat is unfriendly.

Is every promise friendly? Is every threat unfriendly? The answer is no in each case.



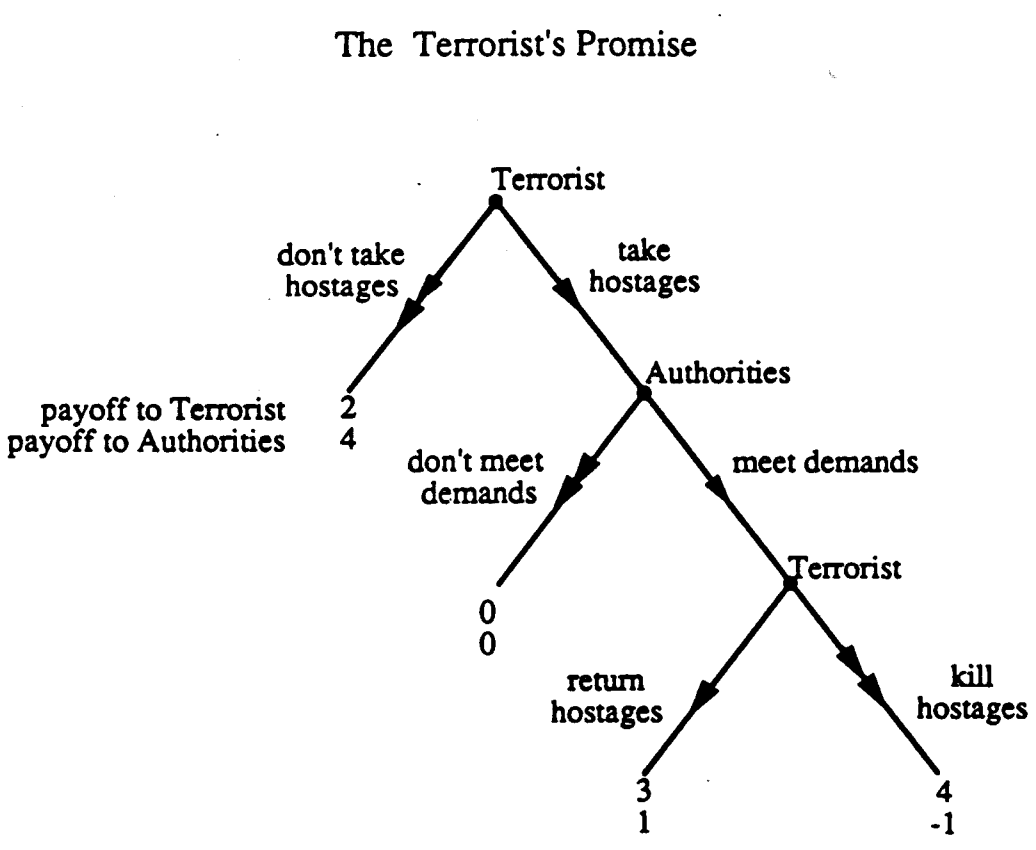
Payoffs are listed in the order: player 1 (ruler)
player 2
player 3

Single arrows is a pure threat and both citizens are better off than under sequential rationality (double arrows).

Fig. 7

Examples can easily show that hybrids can be either friendly or unfriendly. We need to restrict attention to pure promises and pure threats.

The three-player S-game in fig. 7 shows a pure threat that is friendly all around – it is Pareto improving. Without commitment conveyance (double arrows) the payoffs are (1, 2, 2). With commitment conveyance the payoffs are (4, 4, 3). Klein (1990) provides a model of an N-person community living on a plateau around a flood plain and governed by a utilitarian ruler. Each citizen must decide whether to move to the plain. In that model the uncommitted (or 'discretionary') ruler will bail out flood victims with money from plateau dwellers, if there be any. Under this regime everyone moves to the plain, but when flooding occurs there is no one left on the plateau to tax, so no assistance is made. When the ruler has commitment conveyance and threatens to turn a blind eye on flood victims, everyone stays on the plateau and everyone is better off. The plan is a pure threat and it is friendly. Pure threats can be friendly where there is conflict (or a prisoner's dilemma)



Single arrows shows a pure promise that is quite unfriendly.

Fig. 8

among the citizens. [Hillier, Klein and O'Flaherty (1992) generalize this result.]

The terrorist's promise and the benefactor's threat

The friendliness of promises and the unfriendliness of threats can break down even in two-player *S*-games. Consider the Terrorist game of fig. 8, which is identical to an example given by Schelling (1989, p. 115). Terrorist is the ruler. Without commitment conveyance Terrorist is restricted to sequential rationality (double arrows). Even if Terrorist's demands are met he would have the incentive to kill the hostages to preserve his secrecy. Foreseeing this, the authorities would not meet his demands, so Terrorist realizes the futility of taking hostages, much to the relief of the authorities. If Terrorist has commitment conveyance he can promise to return the hostages once his demands are met, which would induce the authorities to meet his

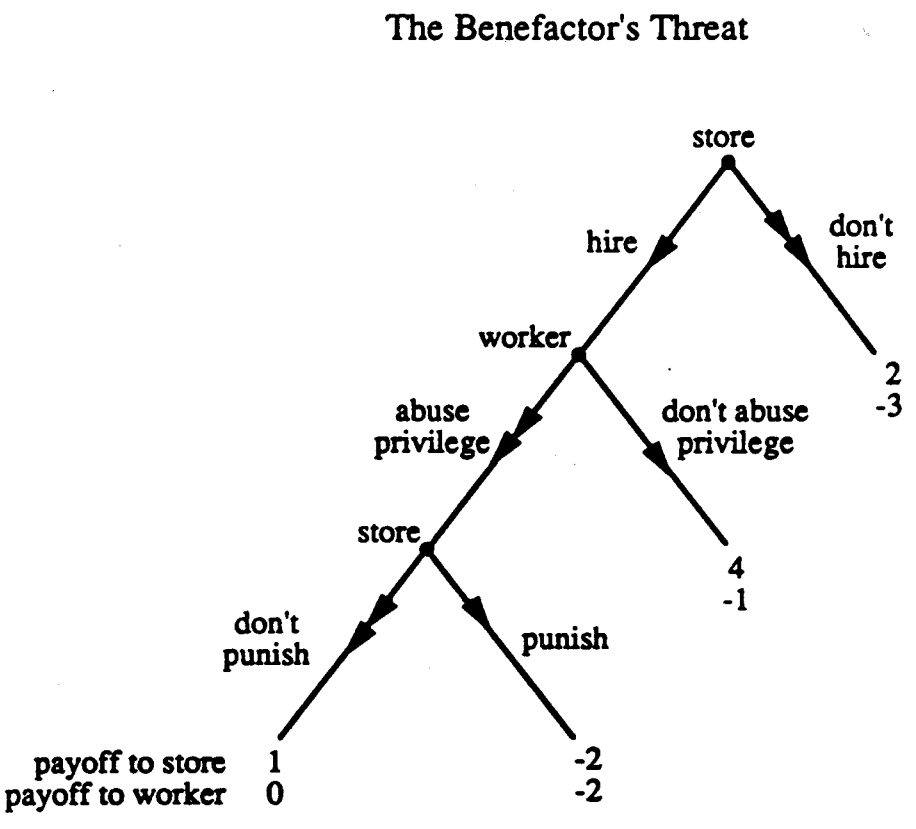


Fig. 9

demands, which would induce Terrorist to take hostages. The promise is the reduction of Terrorist's own payoff from 'kill hostages' by at least one unit – perhaps by publicly taking sacred religious vows, perhaps by taking only women and children, perhaps by demonstrating a growing affection for the hostages. The points is that, although the plan is a pure promise, it is unfriendly.

Fig. 9 shows the analog, the 'Benefactor's Threat,' again like an example given by Schelling (1989, p. 116). The jewelry store is the ruler. It can hire Player 2 as a security guard, but Player 2 can abuse the privilege. If he does, the reference game payoffs say that it is not worthwhile for the jewelry store to punish him. So under sequential rationality (double arrows) Player 2 is not hired. But when the jewelry store can successfully threaten to punish an abusive Player 2, Player 2 does not abuse the privilege of the job and the jewelry stores hires him. The plan is a pure threat, but it is friendly.

The Terrorist's Promise and the Benefactor's Threat are not obvious cases,

so it is not surprising that promises connote friendliness and threats connote unfriendliness. In Part 3 we refine pure promise even further, to exclude the Terrorist's Promise, and use the refinement for a Friendliness Theorem. In the case of pure threats the analogous refinement does not deliver an analogous theorem.

3. Formal characterizations

In this part, we provide formal vocabulary that permits us to (a) define 'promise' and 'threat', (b) define 'pure promise' and 'pure threat', (c) refine the notion of pure promise into 'simple pure promise,' and (d) prove that in two-player S-games simple pure promises are friendly.

Some basic notation

An S-game is a four tuple $\Sigma=(G, R, C, s)$. The 'S' is meant to honor both Stackelberg and Schelling. The reference game, G, is a finite extensive form game with $m+1$ players. The 'R' tells us which player in the G get to be the ruler, and the 's' tells us how the public responds to the ruler's announced plan. We assume that the public is completely credulous and respond with a determinate perfect equilibrium $s(\cdot)$ to the public game induced by the ruler's plan. The 'C' refers to restrictions on the set of plans the ruler can choose from. If the ruler faces no restrictions then she can choose any of her behavior strategies in G as her plan in Σ. C may limit this choice set.

Since player R is the ruler we assume that in G every information set belonging to R is a singleton and that at every node in G all previous play by the ruler has been publicly observed. These assumptions ensure that a subgame originates at every ruler node.

Let $h_j(b,d)$ denote the payoff of player j when the ruler uses plan b and the public uses behavior strategy m-tuple d. For any ruler node v let b_v denote the local strategy specified by b at v. Let $b|b'_v$ denote the ruler's behavior strategy that results if the local strategy assigned by b to node v is changes to b'_v while the local strategies assigned by b to other ruler nodes remain unchanged. For any citizen node w which is an information set, let $s_w(b)$ denote the local strategy specified by $s(b)$ at w. Let $s(b)|s_w(b')$ denote the public strategy m-tuple that results if the local strategy assigned by $s(b)$ to node w is changed to $s_w(b')$ while the local strategies assigned by $s(b)$ to other citizen information sets remain unchanged.

For any node t at which a subgame originates, let $G(t)$ denote the subgame whose origin is ruler node t. Let $b(t)$ denote the behavior strategy induced by plan b on $G(t)$. Denote player j's payoff function on $G(t)$ as $h_{jt}(\cdot)$. Let $s(b(t))$ denote the public's behaviour strategy m-tuple induced on $G(t)$ by $s(b)$.

Sequential rationality

Both the pure promise of fig. 1 and the pure threat of fig. 2 specify moves that would be irrational in subgames of the reference game. We say that the plans of both examples entail sequentially irrational moves. [Our usage of the term 'sequential rationality' is distinct from the usage of Kreps and Wilson (1982)]. Formally, we say that a move b_v at node v is *sequentially rational under* b iff for every local strategy b'_v at v

$$h_{Rv}(b(v), s(b(v))) \geqq h_{Rv}(b(v)|b'_v, s(b(v)|b'_v)). \tag{1}$$

Relation (1) says that what b specifies at node v is a best choice, where 'best' is defined locally. We say that a plan b is *sequentially rational* iff, for every ruler node v, b_v is sequentially rational under b. To comprehend this definition of a sequentially rational plan, think about applying (1) backward through the game. Finally, let $f_v(b)$ denote a move at v that his sequentially rational under b.

Using the idea of sequential rationality we get a straightforward standard for whether the ruler values commitment conveyance. We know that in the back-scratching examples the ruler values commitment conveyance. Without commitment conveyance the ruler is restricted to sequentially rational plans. Therefore, when the ruler has commitment conveyance and at least one plan b that yields to her a payoff higher than a sequentially rational plan b' (that is, $h_R(b, s(b)) > h_R(b', s(b'))$), then the ruler benefits from having commitment conveyance.[5] When the ruler is positively benefitted by commitment conveyance we say that she faces 'commitment dominance.'

Time consistency

The pure threat of fig. 2 shows that commitment dominance does not imply time inconsistency. Time consistency addresses the desirability of deviations from the original plan only along the path of play. Fig. 1 shows the simplest case of time inconsistency. Intuitively, a ruler faces time inconsistency iff along the path of her best plan she reaches a node where she would like to dissolve the original plan (or undo the apparent selective subtractions it implies) and revert to a subplan that is sequentially rational in that subgame. For play of a game in which reversion actually occurs (which implies that the original commitment was phony), there is a concatenation at the reversion point. The public is startled at the reversion point.

[5]Here we are assuming that for any two sequentially rational plans b and b', $h_{Rv}(b(v), s(b(v))) = h_{Rv}(b'(v), s(b'(v)))$ for every ruler node v. This condition is called 'SR-equivalence' in: Klein and O'Flaherty (1993), where it is described at greater length. SR-equivalence fails only when ties in payoffs exist.

Formally, the *concatenated behavior strategy for the ruler* $\kappa(b, v, b')$ is the behaviour strategy that results from behavior strategy b if the strategy on $G(v)$ is changes to $b'(v)$ while the local strategies assigned by b to other nodes remain unchanged. Similarly, the *concatenated behavior strategy m-tuple for the public* $\gamma(d, v, d')$ is the behavior strategy m-tuple that results from behavior strategy m-tuple d if the strategy m-tuple on $G(v)$ is changed to $d'(v)$ which the local strategies assigned by d to other citizen nodes remain unchanged.

A plan b is *time inconsistent* iff there exists a ruler node v and some sequentially rational plan b' such that

$$h_R(\kappa(b, v, b'), \gamma(s(b), v, s(b'))) > h_R(b, s(b)). \tag{2}$$

Built into relation (2) is the along-the-path feature of time inconsistency: if node v is not along the original path of play, the reversion at v will not contribute to making the LHS of (2) greater than the RHS. Also built into the definition is the idea of the public being startled at v. Although they play according to $s(b)$ only at citizens nodes that do not succeed v, they play at those nodes under the belief that b will hold *for the entire game.* This definition of time inconsistency is the natural and faithful game-theoretic version of what writers mean by the term [e.g., Kydland and Prescott (1977), Tesfatsion (1986)].[6]

Promises and threats

A plan b is a *promise* iff

(*P1*) $h_R(b, s(b)) > h_R(b', s(b'))$, where b' is any sequentially rational plan, and
(*P2*) b is time inconsistent.

A plan b is a *pure promise* iff it is a promise and

(*P3*) For every ruler node v off the path of $(b, s(b)), b_v$ is sequentially rationa under b.

Condition (*P1*) says that the plan performs better than sequentiall rational plans. Since we assume that sequential rationality is always availab to the ruler, condition (*P1*) is a bare minimum for the ruler to have a interest in plan b. Condition (*P2*) is the salient features of promises: tim inconsistency [see Guiso and Terlizzese (1990)]. Conditions (*P3*), in th

[6]The only distinctive feature of our definition of time inconsistency is that at a reached no the ruler (with phony commitment conveyance) can reannounce only a sequentially ratio subplan. Alternatively one may wish to permit her to reannounce convincingly any subplan, a to fool the public repeatedly. The issue of the proper choice set at the point of reversion h scarcely arisen in the time inconsistency literature.

definition of pure promise, ensures that the pure promise is not also a threat, since a threat entails self-damaging behavior somewhere off the path:

A plan b is a *threat* iff

(*T1*) $h_R(b, s(b)) > h_R(b', s(b'))$, where b' is any sequentially rational plan, and
(*T2*) there is at least one ruler node v off the path of $(b, s(b))$ such that $h_R(b|f_v(b), s(b|f_v(b))) < h_R(b, s(b))$.

A plan b is a *pure threat* iff it is a threat and

(*T3*) b is time consistent.

Condition (*T1*), like condition (*P1*), ensures that b is of interest. Condition (*T2*) is the salient feature of threats: the presence of locally bad moves in strategically good places in the tree. Condition (*T2*) says that the move b specifies at v not only is sequentially irrational under b, but also that this move is strategically beneficial. Condition (*T3*), in the definition of pure threat, ensures that the pure threat is not also a promise, since a promise is time inconsistent.

These definitions sustain the looser discussion given in Parts 1 and 2 of this paper. With these definitions we can speak more precisely about various species of commitment (promise, threat, hybrids, pure promise, and pure threat). In the appendix we further restrict the idea of promise to permit a theorem that says that the restricted promise is always friendly. The refinement is rather contrived; the exercise shows how far we must go to arrive at a theorem that conforms to our intuition that promises are friendly. We do not refine the idea of threat in an analogous fashion because the analogous claim is not true.

Concluding remark

Following the conceptual schema of Thomas Schelling we have given formulations of 'promise' and 'threat'. These notions are interpreted as species of commitment. Commitment is the strategic displaying of self-penalizing contingent actions. We believe that the Schelling vision of 'promise' and 'threat' conforms fairly well to what people mean by those terms in non-academic discourse and we hope that our formulations advance the usage of those terms in academic discourse.

But we should remark that those terms, especially 'promise', do have other meanings in common parlance. When a mother tells her four-year old daughter during swimming lessons, 'I won't let your head go underwater, I promise,' we might hesitate to call this a strategic commitment. Hirshleifer (1987) explores 'promise' and 'threat' more as expressions of benevolence and malevolence than as strategic actions. In the swimming lesson example 'promise' might be seen as a term used to convey one's benevolence. Yet, if

this is the purpose, why isn't the first declaration ('I won't let your head go underwater') sufficient? Why does she add, 'I promise'? Perhaps the answer is this: since the child may not know if mother is feeling benevolent and even a nonbenevolent mother might costlessly utter the first declaration, the mother has to take an action that a nonbenevolent mother would not take, even though she really is feeling benevolent. Promising is the customary signifier of a pact, a pact that, if broken, warrants retaliation by the disappointed party. The mother, then, is covering all the bases – some of which, despite her benevolence, involve *as-if* selective subtractions *à la* Thomas Schelling.[7]

Appendix: Simple pure promises and a friendliness theorem

The Terrorist's Promise (fig. 8) showed that pure promises need not be friendly, and the Benefactor's Threat (fig. 9) showed that pure threats need not be unfriendly. Let us now restrict attention to two-player S-games, since examples easily show that with three or more players general result regarding friendliness are prohibitively costly.[8]

We can rule out the Terrorist's Promise by adding another condition: A plan b in a two-player S-game is a *simple pure promise* iff it is a pure promise and

(*P4*) at least one of Player 2's nodes w is reached under both $(b, s(b))$ and $(b', s(b'))$, for every sequentially rational b', and $s_w(b) \neq s_w(b')$, for every sequentially rational b'.

The Terrorist's Promise is not a simple pure promise because there is no Player 2 node reached under sequential rationality (double arrows), so (*P4*) fails.

The theorem proves that simple pure promises are friendly. We do not develop an analogous 'simple pure threat' because the analogous unfriendliness clain is not true.[9]

Theorem. Let Σ be a two-player S-game with Player 1 the ruler. Assume that Σ has perfect information[10] *and that the ruler has a unique sequential rational plan, b'. If b is a simple pure promise and b is a pure strategy in Σ*

[7]Schelling (1989, pp. 115–116) makes the same points as we do here.

[8]The S-game in fig. 2, for example, can be altered such that the actions shown affect welfare of a nonparticipating third player. Obviously the threat may benefit this third player.

[9]Consider using (P4) to define 'simple pure threat.' In this case, we can alter fig. (Benefactor's Threat) in the following way: add a worker node in front of the game, with 'out' leading to payoffs 4 for store and −2 for worker, and 'in' leading into the game as drawn. Benefactor's Threat would then be a simple pure threat but still be friendly.

[10]It is a small step to perfect information given that Σ is a two-player game and information restrictions we have already placed on Σ.

reference game, then Player 2 is at least as well off under b and he is under b'; that is, $h_2(b, s(b)) \geqq h_2(b', s(b'))$.

Proof. Consider Player 2's decision under b at a node w referred to in condition (*P4*). Condition (*P3*) and the uniqueness of sequential rationality tell us that if Player 2 moves $s_w(b')$, he moves into a subgame in which b' and $s(b')$ are being played. In other terms,

$$h_{2w}(b(w), s(b) \mid s_w(b')) = h_{2w}(b'(w), s(b'(w))). \tag{3}$$

Since w is along the path of $(b', s(b'))$ and b is a pure strategy, we know that

$$h_{2w}(b'(w), s(b'(w))) = h_2(b', s(b')). \tag{4}$$

By the assumption of perfect equilibrium play by the public in the game induced by a ruler plan we know

$$h_{2w}(b(w), s(b(w))) \geqq h_{2w}(b(w), s(b) \mid s_w(b')). \tag{5}$$

By the assumption of b being pure, we know that

$$h_{2w}(b(w), s(b(w))) = h_2(b, s(b)). \tag{6}$$

On the LHS of (5) we can substitute the RHS of (6), and on the RHS of (5) we can substitute the RHS of (4) (via (3)), yielding

$$h_2(b, s(b)) \geqq h_2(b', s(b)). \qquad \text{Q.E.D.}$$

The theorem establishes a relationship between the pure promise and the welfare of Player 2. That several qualifications are needed to make the theorem work indicates the fragility of the friendliness of promises. As remarked above, the unfriendliness of threats is even more fragile.

References

Cialdini, Robert B., 1984, Influence: How and why people agree to things (William Morrow, New York).

Frank, Robert, 1988, Passions within reason: The strategic role of the emotions (W.W. Norton, New York).

Guiso, Lulgi and Daniele Terlizzese, 1990, Time consistency and subgame perfection: The difference between promises and threats, Banca d'Italia discussion paper, no. 138.

Gauthier, David, 1991, Commitment and choice: An essay on the rationality of plans, ms., (University of Pittsburgh).

Hillier, Brian, Daniel Klein and Brendan O'Flaherty, 1992, Policy commitment and welfare gains, ms.

Fischer, Stanley, 1980, Dynamic inconsistency, cooperation and the benevolent dissembling government, Journal of Economic Dynamics and Control 2, 93–107.

Hirshleifer, Jack, 1987, On the emotions as guarantors of threats and promises, in: John Dupre, ed., The latest on the best: Essays on evolution and optimality (MIT Press, Cambridge, MA) 307–326.

Klein, Daniel B., 1990, The microfoundations of rules versus discretion, Constitutional Political Economy, Fall, 1–19.
Klein, Daniel B. and Brendan O'Flaherty, 1992, Commitment and time consistency: A game-theoretic discussion, ms.
Klein, Daniel B. and Brendan O'Flaherty, 1993, Time consistency and related concepts, ms.
Kreps, David M., 1990, Game theory and economic modeling (Oxford University Press, New York).
Kreps, David M. and Robert Wilson, 1982, Sequential equilibria, Econometrica 50, 863–894.
Kydland, Finn and Edward Prescott, 1977, Rules rather than discretion: The inconsistency of optimal plans, Journal of Political Economy 85, 473–493.
Nalebuff, Barry and Martin Shubik, 1988, Revenge and rational play, Discussion paper no. 138 (Princeton University, NJ).
O'Flaherty, Brendan, 1985, Rational commitment: A foundation for macroeconomics (Duke University Press, Durham, NC).
Schelling, Thomas C., 1960, The strategy of conflict (Harvard University Press, Cambridge, MA).
Schelling, Thomas C., 1989, Promises, Negotiation Journal, 113–118.
Tesfatsion, Leigh, 1986, Time inconsistency of benevolent government economies, Journal of Public Economics, 25–52.