B.J.Pol.S. **28,** 355–387
Printed in the United Kingdom
© *1998 Cambridge University Press*

A Game Theoretic Analysis of the Politics of Taking It in Turns

HUGH WARD*

Taking it in turns is an endemic feature of political and social life. For instance, in a democratic system parties take turns in office according to the outcome of elections. Despite its empirical importance, taking it in turns has received virtually no attention from rational choice theorists. First, a general definition of taking it in turns is given. Secondly, the conditions for taking it in turns to be stable and efficient in Prisoners' Dilemma, Chicken, and Co-ordination supergames are discussed. The key result is that, where stable adherence to a taking it in turns pattern demands that players' continued adherence is conditional on the past adherence of others, players will stick with the pattern if they get their turn often enough. The precise conditions are shown to depend on the payoff structure and the discount rate. Potential problems are identified with initiating stable taking it in turns because of the existence of multiple equilibria and players' desire to get a stable taking it in turns pattern that gives them a high expected payoff. The paradigmatic version of this problem is who goes first when two players alternate between getting a high and a low payoff. The applicability of the results is illustrated by considering the questions of democratic stability and transitions to democracy.

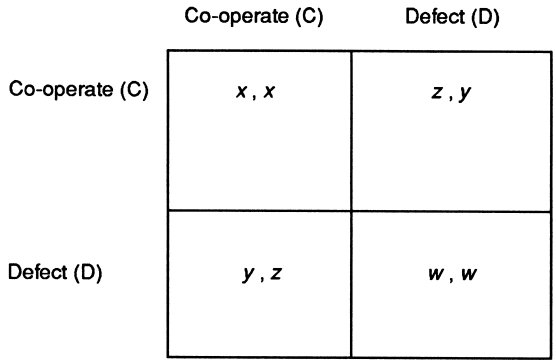
In 1958, after a period of intense social conflict, the intervention of the military, and the failure of attempts at inter-party compromise, the internally divided Conservative party and the Liberal party in Colombia signed a National Front agreement. Among other measures this stipulated that the presidency would alternate between the two parties. Up until the last minute there was considerable contention over whom the parties should nominate and which party's nominee should go forward as the single candidate in the first election. Despite the difficulties in deciding which party should go first, the arrangement remained in place until 1974.[1]

In such situations each electoral term can be thought of as a round of a game extending through time. The game matrix and strategies are shown in Figure 1. It is reasonable to assume that the parties preferred holding the presidency alone to a joint presidency. Thus $y > x$. However, power sharing was probably preferable to the political and social chaos which would have ensued if both sides fought for the office. Thus $x > w$. Suppose that the worst outcome for each

* Department of Government, University of Essex. I would very much like to thank Albert Weale, Iain McLean, the members of the PSA Rational Choice Group and the anonymous referees of this article for their comments.

[1] J. Hartlyn, *The Politics of Coalition Rule in Colombia* (Cambridge: Cambridge University Press, 1988), pp. 63–5; J. Peeler, 'Elite Settlement and Democratic Consolidation: Colombia, Costa Rica and Venezuela', in J. Higley and R. Gunther (eds), *Elites and Democratic Consolidation in Latin America and Southern Europe* (Cambridge: Cambridge University Press, 1992), pp. 94–5.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press



C is the action of power sharing if the other party chooses C and allowing the other party to take office if it chooses D. D is the action of contesting who should hold office if the other party chooses D and taking office if the other party chooses C.

Fig. 1. Payoffs in the stage-game

party in a particular round was to see its rival hold the presidency rather than contest the outcome – that is $w > z$. Under these assumptions the parties repeatedly played a symmetric two strategy, two player Prisoner's Dilemma game or (2×2) PD.[2]

In alternating the presidency the parties were taking it in turns between free-riding (playing D while the other played C) and being taken for a sucker (playing C while the other played D). In this pattern, which I will call taking it in turns to free-ride, it might seem rational for the sucker to defect rather than co-operate, challenging whether it is the other player's turn to free-ride. However, the temptation to break the pattern may be offset by the threat that this would trigger a breakdown, both sides switching to always playing D. In the models developed below this idea is formalized by making each side's strategy conditional, so that their continued adherence to the pattern is contingent on the past adherence of the other side. Even if taking it in turns to free-ride was potentially stable in Colombia, chaos could still have been the result. Each player can get first turn in taking it in turns to free-ride. Each party wanted to hold the presidency first because of uncertainty over whether the pact would hold in subsequent rounds. The danger was that commitment to going first – to playing D in the first game – might have triggered social chaos. Attempts at power sharing in rotation sometimes fail for this reason. In 1996 the leaders of two centre-right parties in Turkey considered rotating the

[2] Another interpretation is that the two sides played Chicken, the worst outcome being social chaos. I assume that the parties were unsure whether the game would continue at the end of each round. Although payoffs may not have been symmetric, this part of my argument only relies on ordinal utilities.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

premiership as the basis of a coalition government to overcome a political crisis occasioned by a hung election. However, the insistence of Mrs Çiller, the leader of the True Path Party, that she should hold office first lead to the collapse of the talks with the Motherland Party.[3]

In a representative democracy it is the will of the electorate that decides who should hold office, not just deals struck between parties like that over the Colombian presidency. It is fundamental that parties should accept electoral defeat, even though they may have the means at their disposal to prevent others assuming state power.[4] It is a party's turn when it is in office. With the exception of national emergencies, some parties are always out of office. If a party is to accept defeat, the belief that their opponents will, in their turn, hand over power – and that they will do so repeatedly according to the will of the electorate – must more than compensate for the benefits which could be gained from challenging the election result. Future election results are necessarily unpredictable. So democratic stability must be founded on probabilistic beliefs about how the 'electoral dice' will roll in future.[5]

In everyday life it is common for people to roll dice or to use some other random device to decide whose turn it is. In stable democracies parties take turns in office according to the outcome of an analogous device. While parties can influence the electoral odds by campaigning and in numerous other ways, no party fully determines the outcome of election and parties are never completely certain about who will win. It will greatly aid a unified exposition in what is to follow to see all forms of taking it turns as governed by a device which tells players whose turn it is at each moment of time. Devices need not be random. For instance, the alternation of the presidency in Colombia can be seen as governed by a notional deterministic device that decided which party would get first go and then told the parties to change control of the presidency at every election. The key to a stable pattern of taking it in turns is often that players' strategies conform to the operation of the device so long as everyone has conformed in the past. Given their strategies are of this form, the device co-ordinates players' actions at every point in time so that they take it in turns.

Dialogue is an elemental political form in which those involved take it in turns to make their point. The device makes it a player's turn when he is asked to reply or has an especially pertinent point to make. Failure of collective action is common when there is cacophony as players speak out of turn. Probably one reason is that the implicit or explicit rules about whose turn it is to speak have unclear implications. Although imperfect devices which do not always clearly indicate whose turn it is to all concerned are important, I assume here that devices are perfect. In dialogues one person typically speaks more often and for

[3] *Financial Times* 27 February 1996.

[4] R. A. Dahl, *Polyarchy: Participation and Opposition* (New Haven, Conn.: Yale University Press, 1971), p. 20.

[5] A. Przeworski, *Democracy and the Market: Political and Economic Reform in Eastern Europe and Latin America* (Cambridge: Cambridge University Press, 1991), chap. 1.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

longer than others, so the general definition of taking it in turns must also allow some player to get more, or longer, turns.

The definition needs to cover both two-person and *n*-person forms of taking it in turns. The importance of the *n*-person case is easily illustrated. Rather than dividing the pork into small pieces, politicians and their constituents may prefer to get a large slice every so often, taking it turns with those in different areas to get meaningful political payoffs. The device in operation here is undoubtedly complex, but is likely to give more senior and more powerful politicians more turns as well as reflecting the needs of legislators to bring home the pork so as to ensure re-election. The most onerous and the most sought after tasks are often rotated among more than two bureaucrats, politicians or nations. The device may be deterministic and known by all in advance, as is the case for the rotation of the presidency in the European Union, or not, as in the case of rotation of the office of prime minister among the parties of the Swiss grand coalition. *N*-person taking it in turns may also be governed by the random device of a lottery, where each has an equal chance of a turn in every round, as in the case of rotation of citizens in public office in the democratic period of ancient Athens or the introduction of most Private Members' Bills in the British House of Commons.

In terms of the matrix form of Figure 1, the stage game is (2×2) Chicken if $y > x > z > w$ and (2×2) Co-ordination if $y > z > x > w$. Mixed-motive games, where there is a tension between the beneficial results of social co-operation and self-interest, are metaphors for characteristic forms of collective action failure: PD for failure due to the dominance of the strategy of free-riding; Chicken for failure due to pre-commitment to free-riding, and Co-ordination for failure due to strategy choices that do not mesh together in mutually beneficial ways. My definition of taking it in turns is applicable to the repeated versions of these games, among others. As I show below, taking it in turns to free-ride, along with other taking it in turns patterns, is a potentially efficient solution for these games when they are repeated through time. In each case there is a danger of collective action failing because of the desire by both sides to get first go, or to institute another taking it in turns pattern that favours their interests. This danger is endemic in social life, the use of pre-commitment tactics to get a favourable pattern leading to the failure to co-ordinate on one pattern and to mutual recriminations.

Social conventions like 'first come first served' may solve the problem of who gets first go on an everyday basis.[6] States play an important role in stabilizing such conventions, which sometimes become enshrined in law and ultimately backed by legal sanctions. Rules which determine whose turn it is to have right of way at sea or on land, which first developed as informal conventions and then acquired legal standing, are standard examples. Yet hierarchical co-ordination

[6] R. Sugden, *The Economics of Rights, Cooperation, and Welfare* (Oxford: Blackwell, 1986), chap. 3 especially. Sugden makes use of the idea of evolutionarily stable equilibria, drawn from the application of game-theory to biology, to show how stable conventions may evolve through trial and error learning or selection when large numbers of randomly chosen pairs of players play games.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

through the use of state power is clearly not capable of resolving all the questions about whose turn it is or of lending stability to all forms of taking it in turns. Indeed, the state often works best in interaction with informal institutions and communal networks, and in an indirect way.[7] For example, within the mutual ship-insurance societies common in nineteenth-century Britain a group of owners undertook partially to compensate other members for the loss of a vessel or a cargo, solving their co-ordination problem over how the costs should be split by contributing in proportion to the value that their own vessels were insured for.[8] Of course, if you ever lost a vessel, you would be compensated in turn. This was a lower-cost alternative to buying insurance on the London markets for some owners, and it was vital to the maintenance of the fleet of small sailing ships right up to the First World War. Like other reciprocity practices, it was probably strengthened by being embedded in strong community structures which provided major social and material incentives to stick to the arrangement: the societies were located in a single, small port community; members were often related and typically had multiple, overlapping business linkages.[9] However, the role of the state was important, too. First, it gave legal status to the mutuals. Secondly, the Merchant Shipping Acts encouraged and formalized the division of ownership of any single ship among many shareholders, reducing individual risks and further interlocking society members into parallel reciprocity practices.

As in the case of other mutual insurance funds, a participant in mutualist ship insurance might never actually make a claim against the fund. We still recognize the practice as being an example of taking it in turns because each participant always has a positive probability of making a claim at any point in time. It is important that any formal definition of taking it in turns allows for this sort of possibility, since restricting the definition to cases where a player must get a turn with certainty at some stage runs against common usage.

Remarkably little theoretical attention has been given to taking it in turns. Rapoport and Chammah definitionally exclude the possibility that taking it in turns to free-ride is Pareto-optimal in repeated PD games.[10] Similarly Axelrod places restrictions on the one-shot payoff matrices he considers in order to prevent taking it in turns to free-ride being Pareto-optimal. His reasoning is that

[7] On the role of formal and legal structures in bolstering older, informal practices surrounding the rotation of access to irrigation water in arid parts of Spain, see E. Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge: Cambridge University Press, 1990), pp. 69–82. For a discussion of heirarchical, negotiated and mixed solutions to co-ordination problems, see F. W. Scharpf, 'Games Real Actors Could Play: Positive and Negative Co-ordination in Embedded Negotiation', *Journal of Theoretical Politics*, 6 (1994), 27–53.

[8] My account draws heavily on B. Greenhill, *The Merchant Schooners* (London: Conway Press, 1988), especially pp. 117–18 and pp. 126–7.

[9] On the role of community in bolstering reciprocity practices, see M. J. Taylor, *Community, Anarchy and Liberty* (Cambridge: Cambridge University Press, 1982).

[10] A. Rapoport and A. Chammah, *Prisoner's Dilemma* (Ann Arbor: University of Michigan Press, 1965).

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

if this pattern (which he calls mutual exploitation) is efficient, players can obviously get out of their dilemma in a repeated game by adhering to it.[11] In fact players who do not place a heavy enough weight on future payoffs will not adhere to this pattern, as I show below. The argument also ignores conflict over who gets first go.

Taylor shows that taking it in turns to free-ride may be an equilibrium of the infinitely repeated (2×2) PD.[12] However, like most others, he misses the importance of taking it in turns empirically and of the potential applications of formal modelling to such phenomena as democratic competition.[13] Also, there are many possible taking it in turns patterns beside taking it in turns to free-ride. The Folk Theorem on infinitely repeated games suggests that any pattern of play through time will be stable if enough weight is placed on future payoffs.[14] However, no one has made explicit the connection between the Folk Theorem and taking it in turns. Also the Folk Theorem does not tell us anything precise about the conditions under which a particular form of taking it in turns is stable.

Taylor makes the important point that where there are two equilibria in which players take it in turns to free-ride, each may attempt to get the one in which he goes first by using commitment tactics, generating an interaction similar to a Chicken game.[15] However, his analysis ignores the complexities which arise as the result of the plethora of equilibria which are likely to exist. Beside this, his argument about commitment is confined to taking it in turns to free-ride in the infinitely repeated (2×2) PD. However, I show that it is much more widely applicable than this.

In Section I, I give a definition of taking it in turns applicable to a very wide class of games. Section II is about taking it in turns in infinitely repeated (2×2) PD, (2×2) Chicken, and (2×2) Co-ordination games. In Section III I generalize the problem of who goes first, showing why failures of collective action can occur.

I. DEFINITION

Taking in turns is a pattern in which: each player can expect to get a turn every so often; it is always someone's turn in every round; and it cannot be everyone's

[11] R. Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984), p. 10.

[12] M. J. Taylor, *The Possibility of Cooperation* (Cambridge: Cambridge University Press, 1987), pp. 75–80.

[13] A notable exception is G. Tsebelis, *Nested Games: Rational Choice in Comparative Politics* (Berkeley: University of California Press, 1990), pp. 175–84, where the idea that institutions within consociational democracy help solidify a form of taking in turns in which each major pillar of society gets its way when the issue is crucial to its interests is explored.

[14] So called because no one can recall who first proved the result. For reviews of the literature, see H. Sabourian, 'Repeated Games: A Survey', in F. Hahn (ed.), *The Economics of Missing Markets, Information, and Games* (Oxford: Oxford University Press, 1989); and D. Fudenberg and J. Tirole, *Game Theory* (Cambridge, Mass.: MIT Press, 1992), chap. 5.

[15] Taylor, *The Possibility of Cooperation*, p. 80.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

turn at once. It is a players' turn when he gets his highest possible payoff under the pattern. Consider a repeated game where a fixed set of n players play the same stage-game at times $t = 1$, $t = 2$, and so on. The pattern of play, P, is the result of the operation in each round of a device Δ which tells each player which stage-game strategy to chose for that round. In this way Δ also determines the stage-game outcome in each round. Δ may operate probabilisticly in round t. Then it has a probability distribution over a set of possible strategy vectors for that round, at least two strategy vectors having a non-zero probability of being drawn. The limiting case is that in which the probability distribution puts all the probability weight on one strategy vector. Then Δ operates deterministicly in round t. A player conforms to Δ in round t if he chooses the strategy Δ tells him to. Thus, there is a well-defined set of possible stage-game outcomes $O_{\Delta,t}$ for each round t; $O_{\Delta,t}$ contains one member if Δ operates deterministicly in round t; $O_{\Delta,t}$ contains more than one outcome if Δ operates probabilisticly in round t. The set of stage-game outcomes which can occur in the whole pattern, P, is

$$\mathbf{Op} = \bigcup_{\forall t} O_{\Delta,t}.$$

Say that a player gets a turn in round t under Δ only if the outcome is one of her highest ranked outcomes in **Op** and that she misses a turn in round t under Δ only if the outcome is not one of her highest ranked outcomes in **Op**. In other words, a turn is seen here as a round in which a player especially benefits by getting the highest payoff she can receive under the pattern. Notice that the highest ranked outcome for a player in **Op** need not be the same as her highest ranked outcome in the stage-game, since **Op** may not include that outcome.

If P is a form of taking it in turns, not everyone gets a turn simultaneously, but it is always at least one player's turn:

(i) Under Δ in any round t, $t \geqslant 1$, there is at least one player getting a turn and at least one player not getting a turn.

Op must contain more than one outcome, or else it would be impossible to satisfy Condition (i). Moreover, for Condition (i) to hold, any outcome in **Op** must be one of the best outcomes in **Op** for at least one player and not one of the best outcomes in **Op** for at least one player.[16] An implication of this is that Δ may trace out a pattern which is a form of taking it in turns for some configurations of player-preferences but not for others.

Nothing that has been said so far rules out Δ telling players to choose mixed

[16] The definition here could be slightly relaxed to allow for taking it in turns in contexts where the same set of players participates in each round but the stage-game played may differ from round-to-round. Then there are still regularity requirements as a result of Condition 1: each stage-game must have at least one outcome such that it is among the best outcomes in **Op** for at least one player but is not among the best outcomes in **Op** for all players, or else Condition 1 cannot be satisfied. Given the outcomes of the stage-games are finite, this requirement means, loosely speaking, that the different stage-games cannot be too dissimilar.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

stage-game strategies; so that the stage-game outcome may be one only feasible under mixed strategies. However, in taking it in turns **Op** should only include outcomes of the stage-game which are feasible if all players choose pure stage-game strategies. To see why, consider what would happen if this condition were not satisfied. Suppose that we allowed **Op** to contain outcomes deriving from the choice of mixed stage-game strategies. In round t Δ tells at least one player to play a mixed stage-game strategy, i.e. at least one player is told to draw a pure stage-game strategy from a probability distribution given to him by Δ. As at least one player gets a turn in every round, there is some player for whom this 'mixed-strategy outcome' has an *expected* value at least as high if not higher than any that can occur under P. After players have drawn for pure strategies as directed by Δ, the *actual* outcome is one of those deriving from pure stage-game strategies, and the utility of his outcome to the player whose turn it is might be low. Yet it does not make intuitive sense to see a player as getting their turn if this happens. To make this graphic, suppose that a party leader whose party got the most votes in an election then had to undergo a lottery which determined who actually won, with the odds favouring his party rather than the opposition. If under this bizarre arrangement the opposition won the lottery despite the odds, it is hardly likely that the party leader would say his turn had occurred![17] We distinguish between a *high chance* of taking a turn and *actually getting* a turn. To mark this distinction:

(ii) In taking it in turns **Op** only contains stage-game outcomes feasible under the choice of pure stage-game strategies by the players.

Intuitively, we would not say that P is a form of taking it in turns if some player(s) never get a turn and/or some player(s) never miss a turn. However, we need to allow for devices which are probabilistic in some round(s). For P to be a form of taking it in turns: each player should have had a positive probability of having a turn, even if the actual outcomes drawn through time by Δ may leave that player without a turn; and each player should have had a positive probability of missing a turn, even if the actual outcomes drawn through time by Δ mean that she never misses a turn. To allow for this in repeated games with a finite number of rounds, r, say that a pattern is a form of taking it in turns only if:

(iii) For each player the probability that they get a turn in all r rounds is less than 1 and the probability that they miss a turn in all r rounds is less than 1.

In an infinitely repeated game there should be some finite future in which the player cannot be guaranteed to get a turn and some finite future in which the

[17] One of this *Journal's* referees insists that it makes intuitive sense for a player's turn to be a lottery ticket with a higher expected value than any other outcome in **Op**. If the definition did not exclude this possibility, this would make no difference to the conclusions about the stability of Pareto-efficient taking-it-in-turns patterns arrived at below. However, the taxonomy of classes of taking it in turns would no longer be a complete one. Moreover, any possible set of supergame payoffs would correspond to at least one form of taking it in turns, something that does not hold under my favoured definition.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

player has a non-zero chance of getting a turn. More precisely, in an infinitely repeated game for each t, $t \geqslant 1$ and each whole number n, $n \geqslant 0$ let: $g_i(t,n)$ be the probability that i always gets a turn in rounds $t+1$, $t+2$, $t+3$, …, $t+n$; and $m_i(t,n)$ be the probability that she always misses a turn in rounds $t+1$, $t+2$, $t+3$, …, $t+n$. Then a pattern is a form of taking it in turns only if:

(iii′) For each player i, for each value of t, $t \geqslant 1$, there exists a finite n such that $g_i(t,n) < 1$, and there exists a finite n' such that $m_i(t,n') < 1$.

Condition (iii′) rules out P being a form of taking it in turns if it gives some player or players a positive probability of getting a turn in some finite sequence of rounds but never allows them the possibility of getting a turn again. Such a pattern is, at most, an abridged form of taking it in turns. Notice that conditions (iii) and (iii′) cannot be fulfilled if some player ranks all the outcomes in **Op** equally. Then it is always their turn. Thus there is the additional restriction on the preference configuration that no player is indifferent between all outcomes in **Op**.

II. TAKING IT IN TURNS IN INFINITELY REPEATED PRISONER'S DILEMMA, CHICKEN, AND CO-ORDINATION GAMES

Discussion of repeated play in (2×2) PD and (2×2) Chicken games has been dominated by the attempt to establish under what conditions co-operation by both sides in each round is stable. In relation to Co-ordination games, the emphasis has been on getting players to stick to one of the equilibria through time. Whenever these practices are Pareto-efficient, though, some form of taking it in turns will also be efficient; and some forms of taking it in turns are efficient in (2×2) PD and (2×2) Chicken when co-operation by both sides in each round is not. There is, then, every reason to analyse taking it in turns if we seek efficient solutions to collective action dilemmas.

In order to conduct the analysis, it is necessary to know what forms of taking it in turns are possible, implying the construction of a taxonomy. The key result is that all forms of taking it in turns belong to one of a small number of classes. Moreover, constructing the taxonomy leads into a straightforward, visual discussion of the efficiency of whole classes of taking it in turns. A feasible resolution of a collective action dilemma through some form of taking it in turns also needs to be stable. The question of stability is complex, and the taxonomy is again of great help in sorting out the issue, not least because it allows simultaneous consideration to be given to the stability of whole classes of patterns. Some classes of taking it in turns in Chicken and all classes of taking it in turns in PD are only stable if conditional strategies are chosen, according to which players who do not adhere to the prescriptions of the device are punished. When this is the case, the key result is that stability requires players to get their turn 'frequently enough' and to be far-sighted enough. Although this sounds rather obvious, the meaning of 'frequently enough' actually takes some

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

unpacking. In contrast, all forms of taking it in turns in Co-ordination and some classes of taking it in turns in Chicken are stable even if the following hold: players' strategies would not punish others for deviation from the prescriptions of the device; one side gets her go with very low 'frequency'; and players are not at all far-sighted. This marks a crucial theoretical distinction, and provides important insights into the comparative viability of taking it in turns in different game structures and social contexts. After giving a full description of the repeated game, the questions of taxonomy, efficiency and stability are discussed in turn. I then give an illustrative application of some of the ideas to the question of democratic stability.

Description of the Games

In a supergame the same stage-game is played an infinite number of times at regular intervals, each play constituting a round.[18] The stage-game is either (2×2) PD, (2×2) Chicken, or (2×2) Co-ordination.[19] Since all-round co-operation is the best outcome for both players in both variants of (2×2). Assurance, taking it in turns is never Pareto-optimal. Because of this I do not discuss Assurance here. I stick to analysing games with symmetric payoff matrices because this makes the exposition more compact.[20]

Players know the actions taken by both sides in all previous rounds and make a simultaneous choice of action in any given round. Players are assumed to discount payoffs in round t, $t \geqslant 1$, of the game by a^t, where $1 > a > 0$. Thus, for example, the discounted value of a payoff of w in each round as seen from $t = 0$ before the first round of the game is

$$\lim_{t^* \to \infty} \sum_{t=1}^{t^*} (aw + a^2w + a^3w + \ldots + a^{t^*}w) = aw/(1 - a).$$

Binding agreements are impossible, hence the games are non-cooperative. I assume that players may use a device to co-ordinate play. For each round the device tells each player whether to choose C or D. The operation of any device being used is common knowledge. For brevity I often call games specified in this way the PD, Chicken, and Co-ordination super games.

Taxonomy of Forms of Taking It in Turns

There are four outcomes of the stage-game deriving from the choice of pure stage-game strategies which could belong to **Op**. **Op** cannot contain all four

[18] Or equivalently the players do not know for certain when the last round is and place a fixed probability on the next round being played – see Fudenberg and Tirole, *Game Theory*, p. 148.

[19] Battle of the Sexes is like Co-ordination except that $x = w$. All the results stated here about Co-ordination also apply to Battle of the Sexes.

[20] The results are easy to extend to games with asymmetric payoff matrices. They can also easily be extended to mixed supergames such as Assurance/PD, Assurance/Chicken, and PD/Chicken.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

outcomes or any three out of the four outcomes, for this would allow an outcome to occur that is neither player's turn. Let CC be the stage-game outcome when both players chose C, and so on. Then if **Op** was {CC, DC, DD}, for instance, DD is one outcome possible under the pattern, but it is neither player's first preference from **Op**. If **Op** was {CC, DD}, *P* cannot be a form of taking it in turns because both player prefer CC to DD. **Op** must contain at least two outcomes if *P* is a form of taking it in turns. Thus **Op** must either be {CD, DC}, {CD, CC}, {DC, CC}, {CD, DD}, or {DC, DD}. In the PD supergame each of these sets is the basis for a class of taking it in turns patterns. For instance, the alternation of the presidency in Columbia (i.e taking it in turns to free-ride) belongs to the class {CD, DC}. In the Chicken supergame the last two cannot form the basis of taking it in turns, because DD is the worst outcome in the set for both players; so it would be nobody's turn should this outcome occur. In the Co-ordination supergame, only the outcome set {CD, DC} forms the basis for taking it turns, as CC is worse-all-round than CD or DC, and DD is the worst outcome of all for both players.

By a supergame payoff point I mean the co-ordinate in payoff space corresponding to the values of the players' payoff stream for the whole game discounted back to their values at $t = 0$. Figure 2 shows the possible supergame payoffs for the symmetric (2×2) Chicken, (2×2) PD, and (2×2) Co-ordination supergames. The point $(wa/(1-a), wa/(1-a))$, which I treat as the origin, *o*, is the supergame payoff point when each player defects in every round. When row gets *z* in each round and column gets his best payoff *y* in each round, the supergame payoff point is $(az/(1-a), ay/(1-a))$, or $p1$. At $(ay/(1-a), az/(1-a))$, or $p2$, row gets her best payoff *y* in each round and column gets *z* in each round. The supergame payoff point when both players co-operate in every round is the point $p3$, or $(ax/(1-a), ax/(1-a))$.

The convex hull of *o*, $p1$, $p2$, and $p3$ is the smallest convex set containing all four points. If mixed and correlated supergame strategies as well as pure-supergame strategies may be chosen, any point in the convex hull is a feasible supergame payoff point.[21] Supergame payoff point $p3$ is an interior point so long as $(y+z)/2 > x$ and is on the line $p1p2$ if $(y+z)/2 = x$.

I prove the following result, along with other propositions given in the text, in the Appendix:

Proposition 1: Let Δ be a device generating a taking it in turns pattern *P* in an infinitely repeated (2×2) PD, (2×2) Chicken, or (2×2) Co-ordination game. Noting that there can only be two possible outcomes o_1 and o_2 in **Op**, denote player *i*'s payoff under the first and second possible outcomes by $o_{1,i}$ and $o_{2,i}$ respectively. Then the supergame payoff point if both players conform to Δ in each round is a convex combination of the supergame payoff points $(o_{1,r}.a/(1-a), o_{1,c}.a/(1-a))$ and $(o_{2,r}.a/(1-a), o_{2,c}.a/(1-a))$ with non-zero weight on each outcome, where *r* denotes row and *c* denotes column.

[21] See Fudenberg and Tirole, *Game Theory*, pp. 151–2. Space constraints preclude me from showing how the summations of supergame payoff streams were carried out.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

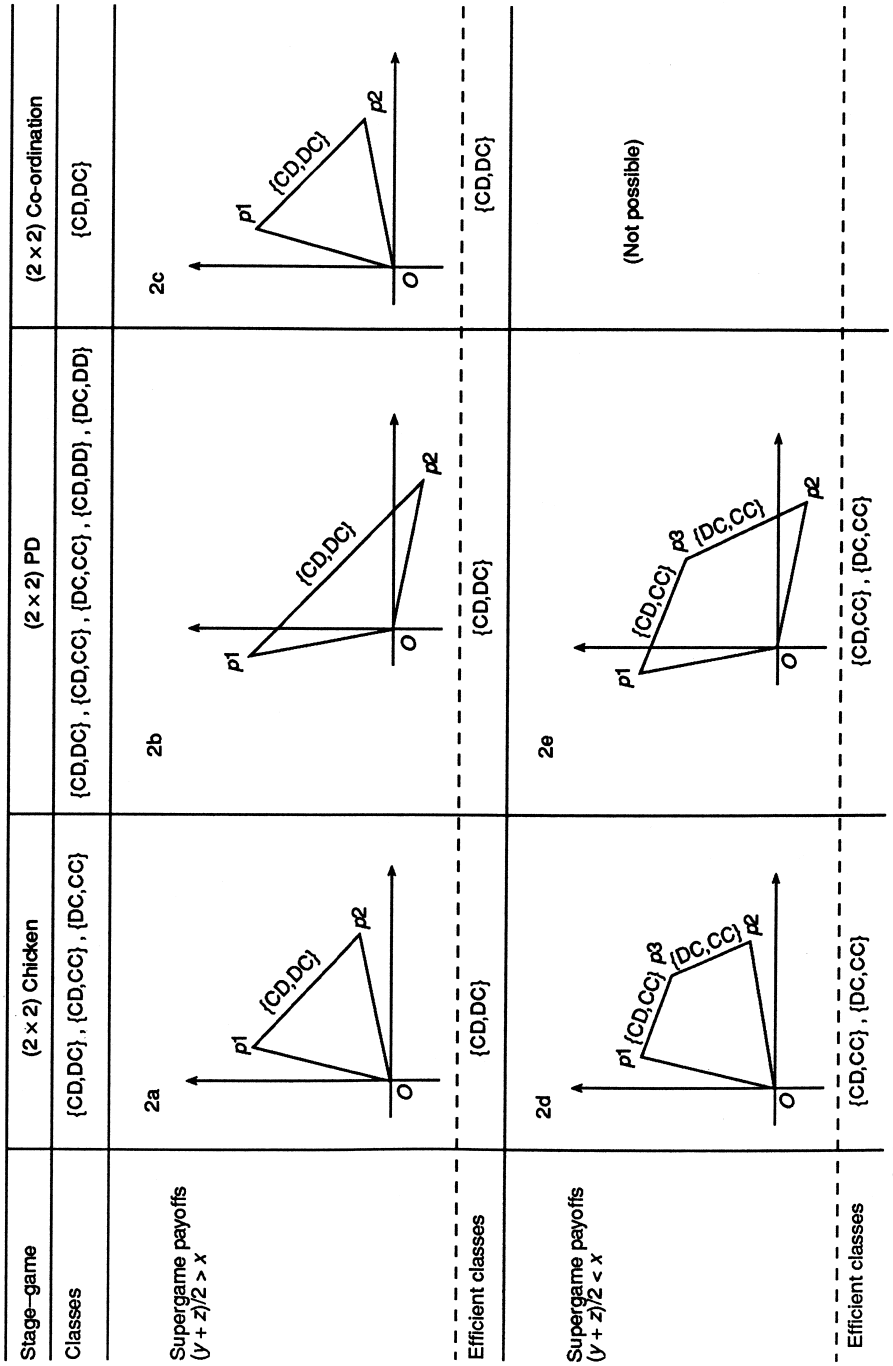


Fig. 2. Taxonomy of taking it in turns

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

If **Op** = {CD, DC}, in both possible stage-game outcomes under P one player free-rides and the other is a sucker, and it is a player's turn when he free-rides. Proposition 1 implies that any member of this class generates a supergame payoff point somewhere on the line $p1p2$, excluding its endpoints. I indicate this by writing {CD, DC} alongside the line $p1p2$ in Figures 2a, 2b and 2c. Beside taking it in turns to free-ride, other members of this class include the patterns in which: (a) row gets m goes, column gets n goes, $m > n$, then the pattern is repeated; (b) a coin is tossed to decide who goes first in taking it in turns to free-ride; (c) in each round t, $t \geqslant 1$, Δ gives column a turn with fixed probability e, $0 < e < 1$, and row a turn with probability $(1 - e)$. In effect, row gets longer turns than column in example (a). The device in example (b) is only probabilistic in the first round. However, devices may be probabilistic throughout the repeated game, as example (c) shows. I call example (c) a simple fixed lottery belonging to the class {CD, DC}. It meets Condition (iii′) above because, in any round, the probability that a player gets a turn is less than 1 but greater than 0. As the final example shows, devices do not have to give a player the same chances of getting a turn in every round. (d) The two system-states of Δ are CD and CC. Then Δ is fully described once its state in round 1 is known and fixed transition probabilities, neither equal to 0 nor 1, for movement from its two possible states in round t, $t \geqslant 1$, to its two possible states in round $(t - 1)$ are given. Example (d) is a Markov-chain. Suppose the player currently taking her turn is in a statistically favourable position in the next round because the transition probabilities favour repetition of the same state. Then the pattern exhibits 'incumbency bias'. Starting in any round, the probability that a player will get a turn in the next round is less than 1 and greater than 0, no matter what the system state in that round, so that Condition (iii′) is satisfied.

The supergame payoff point under the simple fixed lottery belonging to the class {CD, DC} is $(a(ez + (1 - e)y)/(1 - a), a(ey(1 - e)z)/(1 - a))$. As e increases from 0 to 1, the supergame payoff point moves upwards from $p2$ to $p1$ along the line $p1p2$. covering every point on the line except the endpoints. Thus this class has infinitely many members. Any member has a supergame payoff point on the line $p1p2$; so its supergame payoff point corresponds to that of a simple fixed lottery for some value of e, $1 > e > 0$. Thus no matter how complex Δ is, the players could adopt a simple fixed lottery which has the same supergame payoff point. For instance, example (b) above is equivalent to the simple fixed lottery with $e = 0.5$.

A second class for the PD and Chicken supergames is based on **Op** = {CD, CC}. When the stage-game outcome is CD, it is column's turn, and when the stage-game outcome is CC, it is row's turn. By Proposition 1 the supergame payoff point of any member of this class is a point on the line $p1p3$, but not an endpoint. Again, there is an infinity of patterns in this class The supergame payoff point of any member is the same as that of a simple fixed lottery belonging to the class {CD, CC} for some value of e The class {DC, CC}, again only possible in the PD and Chicken supergames, has similar properties but the supergame payoff points all lie on the line $p2p3$, excluding the endpoints. In this class it is row's turn when the outcome is DC and column's turn when the outcome is CC. I indicate where the supergame payoffs points for these classes lie along the Pareto frontier by writing {CD, CC} alongside the line $p1p3$ and {DC, CC} along the line $p2$ $p3$ in Figures 2d and 2e.

In the PD supergame there are two other classes of taking it in turns {CD, DD} and {DC, DD}. By Proposition 1 the supergame payoff point for any member of

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

the class {CD, DD} is point on the line $op1$, but not an endpoint, the supergame payoff point for any member of the class {DC, DD} is a point on the line $op2$, but not an endpoint. No member of these classes is efficient, though. In Figure 2 the Pareto frontier is the north-eastern boundary of the convex hull: $p1$ $p2$ in the cases of Figures 2a, 2b and 2c; $p1p3p2$ in the case of Figures 2d and 2e. The classes {CD, DD} and {DC, DD} in the PD supergame do not contain efficient patterns because the supergame payoff point is never on the Pareto frontier.

The supergame payoff point $p3$ corresponds to both players playing C in every round of the supergame game. However, $p3$ is only an efficient outcome if it is at or to the north-east of the point half way along the line $p1p2$, i.e. if $ax/(1-a) \geqslant a(y+z)/2(1-a$ or $x \geqslant (y+z)/2$. If $x \geqslant (y+z)/2$ the Pareto frontier is $p1$, $p3$, $p2$. Hence, patterns in the classes {CD, CC} and {DC, CC} in the PD and Chicken supergames will also be efficient. Thus, in PD and Chicken supergame if it is efficient for players to choose strategies under which they both co-operate throughout the game, it is also efficient for them to adopt a taking it in turns pattern belong to either the class {CD, CC} or the class {DC, CC}. If $(y+z)/2 \geqslant x$, the Pareto frontier is the line $p1$, $p2$ and all members of the class {CD, DC} are efficient. If $(y+z)/2 > x$, co-operation by both players in every round and also the members of the classes {CD, CC} and {DC, CC } are inefficient. Figure 2 summarizes the taxonomy of classes of taking it in turns and states which classes are efficient.

Stability

Suppose both players' supergame strategies are unconditionally to conform to Δ in a PD supergame. Allowing for the possible probabilistic nature of Δ there is a round in which with non-zero probability one of the players will be called on unconditionally to play C, as **Op** contains either CD or DC or both. If she altered her supergame strategy to play D instead of C in this contingency, keeping it the same in all other respects, a player's expected payoff must increase: her payoff in the contingency increases from z to w and her expected payoff is otherwise the same because her opponent's strategy is unconditional.

Conditional supergame strategies can make continued conformity to the device conditional on the past conformity of the other player. If the threatened penalties for non-conformity are great enough compared to the gains from breaking the pattern, the strategies are in equilibrium.[22] It is not enough to show that a form of taking it in turns is generated by an equilibrium supergame strategy pair, though. Suppose that no matter what the history of the game, neither player wishes unilaterally to depart from her supergame strategy. Then any threat built into the strategy will be credible, because incentives exist to carry it out if the other's history of play demands this. We should only consider Nash equilibria such that the supergame strategies are still in equilibrium for every possible proper subgame,

[22] The equilibria I discuss are a special case of correlated equilibria. In a correlated equilibrium players both observe the outcome of the same randomizing device which tells them which strategy to choose; and the strategies are in equilibrium (see, e.g., E. Rasmusen, *Games and Information* (Oxford: Blackwell, 1989), p. 75). Here the randomizing device is allowed to change from round-to-round and players are restricted, for reasons I have already discussed, to playing pure strategies in every round.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

including those where threats must be carried out. Such equilibria are said to be subgame perfect.[23]

D^{∞} is the unconditional supergame strategy of playing D in every round of the supergame. (D^{∞}, D^{∞}) is always an equilibrium of the PD supergame, because any player switching to another strategy would lose $(w - z)$ in any round in which he co-operated and gain nothing in other rounds. This illustrates a general principle: any supergame strategy pair is a subgame perfect equilibrium so long as the actions of each player in every round give rise to one of the equilibria of the stage-game.[24] This suggests a simple way to construct subgame perfect strategies in PD supergames: if in any round of the game t, $t \geqslant 1$, one or both players fails to conform to D, both players switch to the strategy D^{∞} starting in round $(t + 1)$.[25] Subgame perfection is ensured. If the path along which both players conform is ever departed from, their strategies in the resulting subgame are in equilibrium: (D^{∞}.D^{∞}) is an equilibrium of the supergame; and because the game is infinitely repeated, the subgame starting in round $(t + 1)$ is strategically equivalent to the whole game; hence the player's strategies are in equilibrium in the subgame starting after the pattern is broken.

A player's supergame strategy, S, embodies a grim threat in relation to Δ in a PD supergame if he: conforms to Δ in round 1; conforms to Δ in round t, $t > 1$, so long as both players conformed to Δ in all past rounds; otherwise always play D. In the PD supergame if the strategies S and T both embody the grim threat in relation to Δ, (S, T) is a subgame perfect equilibrium, if it is an equilibrium. For example, let the pattern be taking it in terms to free-ride with column getting first turn and I and I′ be strategies embodying grim threats in relation to a device giving column first go in this pattern. Then if (I, I′) is an equilibrium, it is subgame perfect.[26]

A's security level is the lowest value B can drive A's payoff down to, supposing that A always reacts rationally to what B is doing. In the PD supergame, the players' security levels are w per round, or $aw/(1 - a)$ for the whole supergame. The worst punishment that can be inflicted on a player failing to conform to a device is to drive her payoff down to w per round; hence the greatest possible incentives exist for conforming if the grim threat hangs over her. Actually the punishment phase during which an opponent plays D if you fail to conform need not last forever.[27] However,

[23] R. Selten, 'Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games', *International Journal of Game Theory*, 4 (1975), 25–55.

[24] Fudenberg and Tirole, *Game Theory*, p. 149.

[25] J. Friedman, 'A Non Cooperative Equilibrium for Supergames', *Review of Economic Studies*, 38 (1971), 1–12.

[26] Taylor, *The Possibility of Cooperation*, p. 77, shows that taking it in turns to free-ride can be stable when one player chooses tit-for-tat starting with C (denoted by B by Taylor) and the other chooses tit-for-tat starting with D (denoted by B′ by Taylor). However, if column, say, 'wobbles' by playing D rather than C in a round where the other plays D then returns to B′ playing D in the next round, it will only pay row to retaliate in the next round by playing D (rather than to re-initiate the taking-it-in-turns pattern by playing C) if $aw/(1 - a) \geqslant a(z + ay)/(1 - a^2)$. However, it is necessary for (B, B′) to be an equilibrium that it does not pay row to change to D^{∞}, which implies $a(z + ay)/(1 - a^2) \geqslant aw/(1 - a)$; so that for a subgame perfect equilibrium the condition $aw/(1 - a) = a(z + ay)/(1 - a^2)$ must hold – a highly unlikely occurrence. The threat implicit in B is not credible except along this knife-edge.

[27] The essential features for ensuring subgame perfect equilibria are that players tempted to make a short-term gain by departing from the pattern should be punished and that there should be some

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

if it is shorter than under the grim threat, the incentives to fail to conform will be larger. Thus strategies which embody the grim threat enable us to characterize the weakest condition for stability.

Consider taking it in turns to free-ride in Chicken and Co-ordination supergames. Suppose that row chooses the supergame strategy co-operate in every odd round and defect in every even round, denoted by H; and column chooses the supergame strategy defect in every odd round and co-operate in every even round, denoted by H′. The strategy pairs (H, H′) and (H′, H) give rise to taking it in turns to free-ride, column getting first turn with the pair (H, H′) and row getting the first turn with the pair (H′, H). (C, D) and (D, C) are the equilibria of the (2×2) Chicken and (2×2) Co-ordination stage-games. (H, H′) and (H′, H) must be subgame perfect supergame equilibria no matter what the players' discount rates and the particular values of the payoffs, because the players' actions in each stage-game correspond to a stage-game equilibrium. This argument generalizes to other members of the class {CD, DC} as Proposition 2.[28]

Proposition 2: Any form of taking it in turns belonging to the class {CD, DC} is always stable under unconditional strategies, no matter what the values of the payoffs and the discount rate, when the stage-game is (2×2) Chicken or (2×2) Co-ordination.

In contrast, in Chicken supergames, conditional strategies must be used to stabilize members of the classes {CD, CC} and {DC, CC}. Players may both be asked by Δ to play C in some round; and if one player's strategy was unconditional, the other would rationally change to D, increasing her payoff in the round concerned with no future consequences. If conditional strategies are to be used, the threats embodied in them must be credible. In the case of the Chicken supergame, (D^{∞}, D^{∞}) is not an equilibrium: if one side is playing D^{∞}, the best reply for the other side is C^{∞} the unconditional strategy of playing C in every round. However, (D^{∞}, C^{∞}) and (C^{∞}, D^{∞}) are always supergame equilibria, because (C, D) and (D, C) are equilibria of the (2×2) Chicken stage-game. Say that a player's supergame strategy S embodies a grim threat in relation to Δ in an infinitely repeated (2×2) Chicken if he: conforms to Δ in round 1; conforms to Δ in round t, $t > 1$, so long as both players conformed to Δ in all past rounds; always plays D in subsequent rounds if in the first round in which there was not conformity by both players, the other player failed to conform, always place C in subsequent rounds if in the first round in which both sides do not conform, only he failed to conform. If S and T both embody the

(F'note continued)

incentive for the prescribed schedule of punishment to be stuck to, even if this hurts the punisher. This does not require eternal punishment – see D. Fudenberg and E. Maskin, 'The Folk Theorem in Repeated Games with Discounting or with Incomplete Information', *Econometrica*, 54 (1986), 533–56.

[28] Proposition 2 holds even if it is common knowledge that the supergame only has a finite number of rounds.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

grim threat, (S, T) is a subgame perfect equilibrium if it is an equilibrium.[29] The player breaking the pattern has her payoff forced down to her security level of z per round. Again grim threats ensure subgame perfection in a way which clarifies the least restrictive conditions under which taking it in turns equilibria occur.

Suppose that I's supergame payoff under some strategy pair (S, T) was less than her security level for the whole supergame of $aw/(1-a)$ in the PD supergame or $az/(1-a)$ in the Chicken supergame. I could get at least w per round and, thus, at least $wa/(1-a)$ overall by changing to D^{∞} in the PD supergame and at least z per round and, thus, at least $az/(1-a)$ overall by changing to C^{∞} in the Chicken supergame. Thus (S, T) could not be a supergame equilibrium. The Folk Theorem implies that, if a is high enough and the supergame payoff point gives each player more than her security level, any form of taking it in turns may be generated by a pair of strategies in subgame perfect equilibrium in the PD and Chicken supergames.[30] The intuition is that if players are future orientated enough, they will forgo short-term gains from breaking the pattern rather than triggering a move to an equilibrium in which they are punished by obtaining their security level. This can be illustrated in relation to taking it in turns to free-ride:

Proposition 3: If the stage-game is (2×2) PD, (I, I′) is a subgame perfect supergame equilibrium if and only if neither player can strictly increase his payoff by changing to D^{∞}, given the other player's strategy remains fixed. In algebraic terms, taking it in turns to free-ride is stable under these strategies is and only if:

$$a(z + ay)/(1 - a^2) \geqslant aw/(1 - a).$$

As the payoff matrix is symmetric, if (I, I′) is an equilibrium, so is (I′, I). The players' supergame payoffs in taking it in turns to free-ride are $a(y + az)/(1 - a^2)$ for the player who goes first and $a(z + ay)/(1 - a^2)$ for the player who goes second. The player who goes first gets the higher payoff: because payoffs are discounted, it is strictly better to get your highest payoff, y, in the first and all odd numbered rounds and your lowest payoff, z, in the second and all even numbered rounds, than it is to go second. If the payoff matrix was symmetric, this was the condition for the stability of the presidential pact in Colombia.

There will often be costs in switching between C and D between rounds in a taking-it-in-turns pattern. These costs may deter players from taking it in turns.

[29] Under these strategies both players would play D forever if they simultaneously departed from the pattern, and their strategies in the subgame would not be in equilibrium. This does not matter for subgame perfection, though. In checking for subgame perfection in infinite supergames with discounting and perfect recall, we need only consider deviations from the equilibrium path made by one player in any round – see Fudenberg and Tirole, *Game Theory*, pp. 108–10.

[30] If the supergame is Chicken, Friedman's version of the Folk Theorem might restrict the set of feasible outcomes which can arise as equilibria. However, using more complex threats, and with mild restrictions on the feasible payoffs the result then applies to the infinite supergames of all stage-games, including Chicken. See Fudenberg and Maskin, 'The Folk Theorem in Repeated Games with Discounting or with Incomplete Information'.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

Suppose that every time a player switches between C in one round and D in the next, or vice versa, she incurs costs of σ. The condition for (I, I′) or (I′, I) to be stable, deriving from Proposition 3, now becomes:

$$a(z + ay)/(1 - a^2) - a^2\sigma/(1 - a) \geqslant aw/(1 - a).$$

The second term on the left-hand side of the inequality is the discounted cost of switching that is not incurred under the strategy D^∞. This condition is harder to satisfy the larger the value of σ, so such switching costs make taking it in turns to free-ride less likely. As seems intuitively plausible, it is generally the case that the conditions for other form of taking it in turns discussed in this section to be stable become harder to satisfy the larger the value of σ.[31] So taking it in turns is a less plausible solution to collective action dilemmas than all-round co-operation when both are efficient but there are large switching costs.

I will now show that there is generally an infinity of efficient taking-it-in-turns equilibria if there are any at all. Cheat1 is the strategy which continues to conform until the first round in which Δ demands that both sides play C, plays D in that round, then plays C thereafter. Cheat2 is the strategy which continues to conform until the first round in which Δ demands that the player concerned plays C and the other player plays D, plays D in that round, then always plays D in subsequent rounds. Cheat3 is the strategy which continues to conform until the first round in which Δ demands that both players play C, plays D in that round, then always plays D in subsequent rounds. Propositions 4, 5 and 6 are about simple lottery taking-it-in-turns patterns.

Proposition 4: Suppose that the stage-game is (2×2) Chicken, P is a simple lottery taking-it-in-turns pattern belonging to the class {CD, CC} or the class {DC, CC} and S and T embody grim threats in relation to the device Δ that generates P. Then (S, T) is a subgame perfect equilibrium if and only if the player who always co-operates cannot strictly increase his payoff by changing to Cheat1, given the other's strategy remains fixed.

Proposition 5: Suppose that the stage-game is (2×2) PD, P is a simple lottery taking-it-in-turns pattern belonging to the class {CD, DC}, and S and T embody grim threats in relation to the device Δ that generates P. Then (S, T) is a subgame perfect equilibrium if and only if the player who has the lower probability of getting his turn in each round cannot strictly increase his payoff by changing to D^∞, given the other's strategy remains fixed.

Proposition 6: Suppose that the stage-game is (2×2) PD, P is a simple lottery taking-it-in-turns pattern belong to the class {CD, CC} or the class {DC, CC}, and S and T embody grim threats in relation to the device Δ that generates P. Then (S,

[31] For instance, forms of taking it in turns belonging to the class {CD, DC} are always stable under unconditional strategies in Chicken and Co-ordination supergames without switching costs; but they are stable, if at all, only under conditional strategies when $(y - \sigma) < x$ and/or $(z - \sigma) < w$. The reason is that DC is no longer a stage-game equilibrium after any round where the outcome was CD and CD is no longer a stage-game equilibrium after any round where the outcome was DC.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

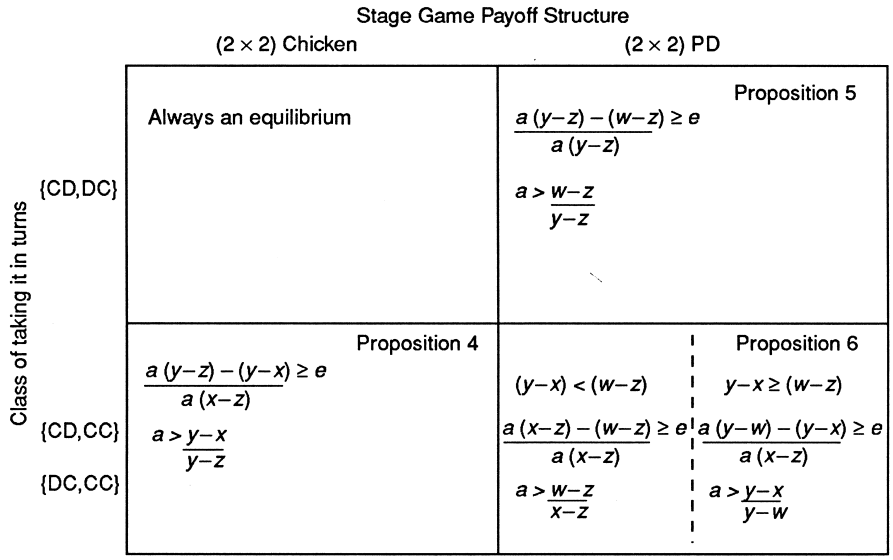


Fig. 3. Necessary and sufficient conditions for equilibrium in simple lottery forms of taking it in turns

T) is a subgame perfect equilibrium if and only if the player who always co-operates cannot strictly increase his payoff by changing to a Cheat2 if $(y - x) < (w - z)$ or Cheat3 if $(y - x) \geqslant (w - z)$.

The algebraic form of the necessary and sufficient conditions for equilibria are given in Figure 3. The cases in the upper left-hand cell are simple lotteries belonging to the class {CD, DC} in the Chicken supergame. Proposition 2 shows that the infinity of taking-it-in-turns patterns belonging to this class can always be generated by unconditional strategies in equilibrium.

In the other three cells in the table, e is the probability that for each round t, $t \geqslant 1$, the player referred to in the relevant proposition named in the cell does not get a turn in that round. For these three cells, if the pattern is to be stable, it must be the case that e is not too high. The intuition is that if a player does not get his turn often enough, the temptation to enjoy a once-off gain by breaking the pattern in some way will be compelling. The conditions in Propositions 4, 5 and 6 appear to apply only to one of the two players. The proofs demonstrate that if these conditions hold, it will not pay the other player to break the pattern either. The intuition is that the other player is more advantaged by the pattern and, hence, suffers larger opportunity costs if it is broken. Secondly, as suggested by the Folk Theorem, for stability a must be sufficiently high, or else the long-term opportunity costs of breaking the pattern will not loom large enough to deter players from doing so.

The key argument so far is that stability of taking it in simple lottery turns patterns demands that the frequency with which players get their turn is sufficiently high and a heavy enough weight is placed on the future. As Propositions 7 and 8 show, these arguments can be generalized.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

Proposition 7: Suppose that the stage-game is (2 × 2) PD, P is any taking-it-in-turns pattern belonging to the class {CD, DC}, and S and T embody grim threats in relation to the device Δ that generates P. Then there exists a simple lottery taking it in turns pattern P* belonging to the same class as P, such that if strategies S* and T* embody grim threats in relation to the associated device Δ* and if (S*, T*) is an equilibrium, then (S, T) is a subgame perfect equilibrium.

Proposition 8: Suppose that the stage-game is (2 × 2) PD or (2 × 2) Chicken, P is any taking-it-in-turns pattern belonging to the class {CD, CC} or the class {DC, CC}, and S and T embody grim threats in relation to the device Δ that generates P. Then there exists a simple lottery taking-it-in-turns pattern P*, belonging to the same class as P, such that if strategies S* and T* embody grim threats in relation to the associated device Δ* and if (S*, T*) is an equilibrium, then (S, T) is a subgame perfect equilibrium.

The extension of Proposition 1 (see the Appendix) shows that, no matter what round has been reached in the supergame, the discounted sum of the players' future payoffs is equal to that of some simple lottery, *e*, belonging to the same class. The calculation of *e* takes account of: how often the players get their turn in rounds in which Δ is deterministic, less weight being placed on a turn the further into the future it occurs; the probabilities with which players get their turn in rounds in which Δ is probabilistic, less weight being placed on the probability of getting a turn the further into the future the round is. Thus *e* can be thought of as a summary statistic measuring the 'frequency' with which a player gets her go under Δ from some given point onwards. Propositions 7 and 8 say that the stability of the patterns concerned is ensured if there is no round in which either player has an incentive to break some equivalent simple lottery pattern starting in that round, i.e. players always expect to get a go 'frequently enough' in future rounds.

We can summarize the overall argument informally as follows. Where conditional strategies are necessary to stabilize a form of taking it in turns, the pattern will generally be stable as long as each player expects to receive a turn 'frequently enough', irrespective of how many rounds of the supergame have occurred already. For simple lotteries, the frequency is *e*, the probability of getting a go in any round. For more complex patterns, the 'frequency' is a summary measure of the future pattern. However, forms of taking it in turns in the Co-ordination supergame and members of the class {CD, DC} in the Chicken supergame are always stable no matter how 'infrequently' players get their turn.

An Application to Democratic Stability

It is not possible fully to discuss applications of these results. Nevertheless, I will illustrate the possibilities of applying the results about the class {CD, DC}. Above, I introduced the idea that parties take it in turns in office. In an ideal-type majoritarian democracy the following conditions hold: single-party governments are the norm; two major parties take it in turns in office; turns in office

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

are determined by the outcome of periodic elections under plurality rule.[32] Assuming that each party has the social base seriously to challenge election outcomes in a way that could destroy democracy, the stage-game made up of one fixed term in office is the same as that shown in Figure 1. Parties infinitely repeat the stage-game and discount future payoffs by the discount factor a, $0 < a < 1$. Majoritarian democracy will be modelled as a simple lottery taking-it-in-turns pattern belonging to the class {CD, DC}, where e, $0 < e < 1$, is the shared subjective probability that row wins and column loses any future election.[33]

If the stage-game was (2×2) Chicken, the worst outcome for each party in any particular round occurs when both parties play D. Suppose both sides' strategy is unconditionally to conform to the electoral result in each round. According to Proposition 2, these strategies must always be in equilibrium, no matter what the values of e and a, or the payoffs. In this case we might say that democracy is unconditionally stable. In contrast, democracy is never unconditionally stable if the stage-game is a (2×2) PD. Then majoritarian democracy is only stable if parties' conformity to the current election outcomes is conditional upon the other side's past conformity.[34] Then the weakest conditions for the stability of democracy are given by Proposition 5 and in the upper right-hand cell of Figure 3. If the conditions are not satisfied, one of the parties will challenge the first election it loses, both parties playing D in that round and in all subsequent rounds.

An empirically testable implication of the model is that the perceived probability of winning future elections for the party seen as least likely to win must not fall too far if democracy is to be stable when the parties repeatedly play PD.[35] Then majoritarian democracy is only stable if the electoral system gives

[32] A. Lijphart, *Democracies: Patterns of Majoritarian and Consensus Government in Twenty-One Countries* (New Haven, Conn.: Yale University Press, 1984), chap. 1.

[33] Another option here would be model the process as a Markov chain over the system states CD and DC, so that allowance could be made for such effects as incumbency bias in defining transition probabilities. Arguments about convergence in two-party systems might suggest that $e = {}^1/_2$. However, the argument here is neither restricted to one-dimensional competition where voters have single-peaked preferences nor to pure office-seeking parties. There is no presupposition that all voters vote rationally, and differences in the size of blocks of identifiers may be related to different values of e.

[34] If both sides switch to D^∞, rounds are still defined by the period before an election would have been held under the constitution that prevailed before the breakdown of democracy.

[35] Whether we are dealing with a democracy if e is too low is an important normative issue. See K. Strom, 'Democracy as Political Competition', in G. Marks and L. Diamond, eds, *Reexamining Democracy: Essays in Honor of Seymour Martin Lipset* (Newbury Park, Calif.: Sage, 1992). The literature on democratic transitions suggests that moderate reformers must not have an overwhelming probability of winning future elections if stability is to be assured. See A. Przeworski, 'The Games of Transition', in S. Mainwaring, G. O'Donnell and J. S. Valenzuela, eds, *Issues in Democratic Consolidation* (Notre Dame, Ind.: University of Notre Dame Press, 1992), pp. xxx–xxx, at p. 121; G. O'Donnell and P. C. Schmitter, *Transitions From Authoritarian Rule: Tentative Conclusions about Uncertain Democracies* (Baltimore, Md.: The Johns Hopkins University Press, 1986), pp. 62–3. My conclusion here is similar to that reached by Przeworski – see *Democracy and the Market*, pp. 30–1.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

each party a reasonable chance of holding office alone, because parties will not adhere to electoral rules under which they have little or no chance of gaining office. Second parties must place enough weight on the future prospect of democracy breaking down to forgo the short-term gains from challenging the results of elections which they lose. The condition in the upper right-hand cell of Figure 3 becomes harder to satisfy, other things being equal, as *y* and *z* fall (because a party gets a lower payoff in the current and future rounds by sticking to the majoritarian pattern under which it sometimes gets *y* and sometimes gets *z*) and as *w* increases (because the party gets a higher payoff both in the current and future rounds by breaking the majoritarian pattern).[36]

III. COLLECTIVE ACTION FAILURE IN TAKING IT IN TURNS

The Folk Theorem is notorious because it suggests that there will generally be an infinity of subgame perfect supergame equilibria. This worries some game theorists because the predictive power of the approach is reduced, leading to attempts to develop further refinements of the equilibrium concept.[37] Yet multiple equilibria are an underlying feature of social reality which generate important social phenomena – in this case the potential for collective action failure. I first illustrate the problem using the paradigmatic case of who goes first in taking it in turns to free-ride in a (2×2) PD supergame. I then show that there is a general problem of the same sort with other patterns. Finally, I show how the analysis can be applied to transitions to democracy.

(f'note continued)
There are, however, considerable differences between my own analysis and Przeworski's: (i) his analysis does not establish game-theoretic conditions for stability, because the full range of alternative strategies to adhering to the current election results is not considered; (ii) he does not analyse the credibility of the threat strategies that in his model implicitly face those who consider not adhering to an election result; (iii) he does not consider the implications of the multiplicity of equilibria that may exist, including power-sharing and equilibria associated with different rules giving different probabilities of success, so his model does not capture the commitment problem discussed in the next section.

[36] The literature on transitions stresses that the constraints on the powers of government to reshape society, enshrined in broader social pacts, must not push the benefits of office holding, *y*, too low for moderate reformers, or else they will not play by the rules. See Przeworski, 'The Games of Transition', p. 127. Past experience of conflicts, which would arguably lead to lower values of *w*, seems to put pressure on elites to accept taking it in turns. See Burton, Higley and Gunther, 'Introduction: Elite Transformation and Domestic Regimes', p. 14; O'Donnell and Schmitter, *Transitions From Authoritarian Rule*, p. 31. In the model as $(w - z)$ – the opportunity cost of accepting an election result which leaves you without office – increases, stability becomes less likely. Convergence between the ideologies of elites should reduce $(w - z)$and such convergence is known to be one factor promoting stability. See Burton, Higley and Gunther, 'Introduction: Elite Transformation and Domestic Regimes', pp. 24–30.

[37] J. C. Harsanyi and R. Selten, *A General Theory of Equilibrium Selection in Games* (Cambridge, Mass.: MIT Press, 1988).

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

Who Goes First in Taking It in Turns to Free-ride

Supposing that (I′, I) is an efficient equilibrium, there is still no reason in terms of the logic of pure game-theory to suppose that if row chose I′ column would choose I. If (I′, I) is an equilibrium, so is (I, I′). If column chose a strategy corresponding to an equilibrium, he might choose I′, too; or he might choose a strategy corresponding to any one of the myriad of other equilibria of the supergame. Without a common conjecture about which equilibrium will eventuate, there is no solution to players' problems of rational choice. Common conjectures could arise because players' attention is focused on certain prominent equilibria which correspond to relatively simple strategies, are known to be efficient, and are seen as fair.[38] Taking it in turns to free-ride is likely to be one prominent pattern. First, it can be based on simple-strategies such as I′ and I which are easy to understand and to implement. Secondly, it is easy to check whether (I, I′) and (I′, I) are stable and efficient: it need only be shown that taking it in turns to free-ride is better all round than both players co-operating throughout the supergame.[39] Thirdly, taking it in turns to free-ride is fairer than many other forms of taking it in turns because each side gets the higher payoff of y in turn and each side gets its turn with equal frequency.[40] Yet, this still leaves the question of who should get first turn.

The conventions, devices and institutions for settling this question are varied. Tossing a coin to see who goes first is a widely used method. Another form of co-ordination is queuing, those who have queued longest getting first turn. Alternatively, fairness may demand that those with the greatest need get first turn. Another possibility is that players higher up a relevant social hierarchy may get first turn. There may even be a market in which the player who bids most gets first turn.[41] If the players accept that one of them should get first turn, there is no collective action problem. However, co-ordination devices may give conflicting guidance.

Suppose that the players' conjecture is that either the equilibrium will be (I, I′) or it will be (I′, I), yet players still contend over who gets first turn. Then each might commit to I′ hoping that this would make the other rationally choose I. One possible picture is of a pre-game, played before the supergame, with just

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

[39] Taking it in turns to free-ride is better all round than both players co-operating in every round if it is at least as good for the player who goes second, or $a(z + ay)/(1 - a^2) \geqslant ax/(1 - a)$ which is equivalent to $a \geqslant (x - z)/(y - x)$. This condition is strong enough to ensure taking it in turns to free-ride is an equilibrium if $(x - z)/(y - x) \geqslant (w - z)/(y - w)$ or $x(y - z) \geqslant w(y \geqslant z)$, which is true since $x > w$.

[40] I conjecture that members of the classes {CD, CC} and {DC, CC} will be met with more rarely empirically than members of the class {CD, DC} because they are liable to be seen as unfair to the player who always has to co-operate, as the other player's payoff is always at least as high if not higher.

[41] A possibility discussed in relation to rotating credit associations in T. Besley, S. Coate and G. Lonry, 'The Economics of Rotating Savings and Credit Associations', *American Economic Review* 83 (1993), 792–810, p. 793.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

two stages. In the first stage players can commit themselves to playing I′ or else do nothing. Such signals push up the prior-probability that the other side places on you playing I′. However, players' commitments need not be irreversible, so that the announcement of a commitment need not push up the prior probability to 1. In the second stage players have to choose between I and I′ in ignorance of the other side's final choice of supergame strategy. If you have announced a commitment to I′ but you actually play I, you suffer a penalty.

Such a pre-game could be solved using standard methods to find plausible equilibria in extensive-form games of imperfect information. However, there is no need fully to formalize and to solve such a game to show that both sides may actually play I′. First there is the danger that both sides irreversibly commit to playing I′, or that it is so costly for them to back down on their commitment to I′ having announced it, that they must rationally choose I′. This is not the only possible danger, though. Suppose commitments are not binding. At the second stage the incentives to choose I′ depend on subjective beliefs (possibly as modified in the light of any announced commitments) about how likely the other player is finally to choose I′. Each player has a larger incentive to choose I′ the lower the final probability assigned to the other choosing I′. Also the greater the advantages of going first as compared with going second, which go up the more heavily future payoffs are discounted, the greater the incentives to play I′.[42] Depending on these factors, both players may choose I′. In any case if both sides play I′, D is their first-round action, so the grim threats are activated and then both sides always play D.

The General Problem

There will typically be an infinity of efficient taking-it-in-turns equilibria. Players' attention may be focused on some of these equilibria, but there will typically be more than one pattern under the spotlight. Because they lie along the Pareto frontier, the two players' preferences over the associated supergame payoff points are diametrically opposed: the higher the outcome is ranked by one player, the lower it will be ranked by the other. From among the possibilities being considered, a player might commit to a supergame strategy corresponding to a device giving him his turn with high 'frequency'. If players make binding commitments, there is prone to be collective action failure. The form taken by this differs between the cases of the PD, Chicken, and Co-ordination supergames, as discussed below. Even when commitments are not binding players need not rationally choose strategies that 'match' in such a way that

[42] Let p be the probability row places on column choosing I and $(1-p)$ be the probability row places on column choosing I′. Then row would choose I′ over I if

$$pa\frac{(y+az)}{(1-a^2)}+(1-p)\frac{aw}{(1-a)}>pa\frac{(x+aw)}{(1-a)}+(1-p)\frac{a(z+ay)}{(1-a^2)}.$$

It is easily shown that row has lower incentives to choose I′ as p and a increase.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

threats are not triggered, as we will also see shortly. Trigger strategies need not force players' payoffs down to their security level forever like the grim strategies in PD supergames do. Also players may be able to talk their way out of being locked into supergame payoff points corresponding to security level payoffs. Nevertheless, there may be at least temporary inefficiency.

In the PD supergame if the two players choose supergame strategies corresponding to different taking-it-in-turns patterns, eventually there is a probability greater than zero that there will be a round in which the grim threat made by one or both players is activated as the other side does not carry out the required action in the last round. If the players choose strategies corresponding to different deterministic devices, the probability of collective action failure will be 1: in the first round in which the two devices make conflicting demands, threats will be triggered. The result is that players become locked into playing D forever, which is inefficient. If the devices are probabilistic in each round, players may escape from collective action failure. Consider the case where each player chooses a strategy corresponding to a fixed lottery but the probabilities of the players getting their turns are not the same. It is possible that the devices will continue to make the same demands on players, but the probability that they will continue to do so goes to 0 as the number of rounds goes to infinity.

The Chicken and Co-ordination supergames are different. In the Chicken supergame we need to consider two cases. First, if the players choose strategies corresponding to different patterns belonging to one of the classes {CD, CC} or {DC, CC} and the players' strategies embody grim threats, the triggering of the threat of one of the players leads the game either to the equilibrium (C^{∞}, D^{∞}) or to the equilibrium (D^{∞}, C^{∞}). These equilibria are efficient; so if threats are triggered there is no collective action failure. Secondly, if the players commit to different patterns belonging to the class {CD, DC} and their strategies are not conditional, there are two possible sources of inefficiency: (i) both players may be irreversibly committed to playing D in some round or rounds, which is worse all round than one playing C and the other D or both playing C; (ii) both players may be irreversibly committed to playing C in some rounds in conditions in which it would be better to adopt a pattern where the outcome was CD in some of those rounds and DC in others, because $(y + z)/2 > x$. The problem in Co-ordination supergames is similar. Players may commit to different patterns belonging to the class {CD, DC}. The danger is that either both will play C in a given round or both will play D in a given round. The outcomes CC and DD are both inefficient compared to the two outcomes where one side free-rides, CD and DC.

Commitment Problems in Transitions to Democracy

Transitions to democracy have often been pictured as a game played between the softliners in the authoritarian power block and moderate reformers among

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

the opposition.[43] Hardliners in the authoritarian power block and revolutionaries within the opposition can be politically side-lined so long as some agreement can be reached between softliners and reformers about: the timing of the transition; any limits on the scope of democratic contention, such as items like national security that will remain off the agenda; and the electoral system. The first point of contention is timing. Authoritarian regimes are typically 'regimes that practice dictatorship and repression in the present while promising democracy and freedom in the future'.[44] Softliners generally prefer to delay transition, partly in order to buy off hardliners; while reformers will want transition as soon as possible, partly to reduce support for revolutionaries. Secondly, softliners want electoral systems that favour the chances of parties representing the interests of the authoritarian bloc and its allies; whereas reformers want rules favouring their social constituencies.[45]

The process can be thought of in terms of the ideas about commitment and the effects it triggers, developed above. The normative appeal of democracy, internal pressure or pressure from other nations leads the softliners and reformers to focus upon 'Democracy Now' versus 'Delayed Democracy'. For simplicity I assume that the softliners (row) and the reformers (column) have a one-to-one fit with the parties that will play the democratic game after the transition, although things are often much more complex than this. Suppose that the stage-game is (2×2) PD. 'Delayed Democracy' is the taking-it-in-turns pattern favoured by the softliners where: they play D for n rounds, $n \geqslant 1$; the reformers play C for n rounds; then the democratic simple lottery with probability e belonging to the class {CD, DC} is instituted. Under this pattern the softliners remain in power for n rounds irrespective of their popular support. After this there is a transition and the softliners win power with probability e. 'Democracy Now' is the taking-it-in-turns pattern favoured by the reformers, corresponding to the simple lottery with probability e' belonging to the class {CD, DC}. Here the democratic lottery starts immediately, softliners being expected to win elections held under an alternative set of rules with probability e'. Consistent with the reformers' interest in an electoral system that favours their party, $e > e'$. Proposition 5 proves that both these patterns will be a subgame-perfect equilibria under grim threats if each side's perceived probability of winning future elections when they start to take place is high

[43] Compare J. M. Colomer, 'Transitions By Agreement: Modelling the Spanish Way', *American Political Science Review*, 85 (1991), 1283–302; Przeworski, *Democracy and the Market*, chap. 1; Przeworski, 'The Games of Transition'; G. Marks, 'Rational Sources of Chaos in Democratic Transitions', in Marks and Diamond, eds, *Reexamining Democracy*; Y. Cohen, *Radicals, Reformers and Reactionaries: The Prisoner's Dilemma and the Collapse of Democracy in Latin America* (Chicago: University of Chicago Press, 1994); J. M. Colomer and M. Pascual, 'The Polish Games of Transition', *Communist and Post Communist Studies*, 27 (1994), 275–94.

[44] O'Donnell and Schmitter, *Transitions from Authoritarian Rule*, p. 15.

[45] See M. Burton, R. Gunther and J. Higley, 'Elites and Democratic Consolidation in Latin America and Southern Europe: An Overview', in Higley and Gunther, eds, *Elites and Democratic Consolidation in Latin America and Southern Europe*, p. 334; Przeworski, 'The Games of Transition', pp. 116–26.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

enough and each side weighs future payoffs heavily enough. That is to say given the threats of breakdown of their relationship that are in the air, given a suitable electoral system, reformers may be willing to put up with a delayed transition *and* softliners may be willing to put up with an immediate transition.

Suppose that the reformers choose a grim strategy corresponding to 'Democracy Now' in the belief that the softliners are not actually fully committed to 'Delayed Democracy'. Meanwhile the softliners choose a grim strategy corresponding to 'Delayed Democracy' in the belief that the reformers have been scared into accepting delay by the use of commitment tactics. If the reformers believe that they would have won an election in the first n rounds under fairer rules and procedures giving them a probability of e' rather than zero of winning, they play D thereafter, and the softliners follows suit, triggering a failure to make a transition that would favour both sides. One way of understanding why transitions fail to occur is, then, that commitments to 'Delayed Democracy' and 'Democracy Now' trigger conflicts between softliners and reformers, often provoking a wider social conflict in which the hardliners and the military weigh in to suppress revolutionaries who emerge from the sidelines when the attempt at pact-making fails.[46]

Even if parties and the social blocs they represent have settled into a democratic pattern, there will be multiple opportunities for them to use commitment tactics to try and get to a more favourable equilibrium. For example, there is always the possibility that democracy will be imperilled as a party commits to remaining in power – typically with military backing – whatever the result of the election. Also parties may commit to a favourable change in the electoral system giving them a greater chance of control or changes in rules which alter the expected payoffs in their favour should the other side achieve power. Such tactics can be destabilizing. For instance, attempts to change the rules by the right at or after the election of a left of centre president were significant in the processes leading to the breakdown of democracy in Brazil in the 1960s and in Chile in the 1970s.

IV. CONCLUSION

I have shown that taking it in turns is an important political phenomenon. A general definition of it can be proposed that fits several empirically important cases, and game theory gives empirically plausible insights into questions of stability, efficiency and the use of tactics to get a favourable pattern. Much remains to be done, though. Many of the plausible empirical examples of taking it in turns in politics involve more than two players. This suggests that the

[46] Compare this possibility with the account of failures of transition given in Przeworski, *Democracy and the Market*, chap. 2. I do not claim that this model is a fully plausible account of failed transitions, but I would argue that tensions over the timing of transition and its form are one significant cause of conflict.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

analysis should be extended to the *n*-person cases covered by my general definition. Commitment tactics are not the only reason why taking it in turns may not get going. Problems may also arise because of distrust. As Thomas Hobbes noted, in the absence of some mechanism to make them binding, agreements in which A performs first, then B performs second are problematic: B may have no intention of keeping her part of the bargain, despite her *declared* preferences, and even though she may be punished by A withdrawing from reciprocity practices in the future.[47] In such situations B may gain by pretending that she actually intends to take it in turns, inducing A to start to play the pattern in order to get a turn, with no intention of giving A a turn in future rounds. A might be inclined to give B a turn because of his fear of the grim outcome. But he might incline away from this because of distrust of B's intentions. Information about others' payoffs and discount rates is often very incomplete in reality, and an assurance problem may arise because of this, leading to an inefficient outcome.[48] Future research should focus on modelling the consequences of distrust and ways of credibly signalling good intent. I have shown that some forms of taking it in turns are efficient whenever co-operation by both players throughout the game is not; even when it is efficient for both players to co-operate all the time, there will be taking it in turns patterns which are also efficient. Considerable effort has now been expended on discussing institutional designs which might foster conditional co-operation by both players throughout the game.[49] It is crucial that we now consider institutional designs to help stabilize taking it in turns.

APPENDIX

Proof of Proposition

Denote row and column's supergame payoffs under P by R and C respectively. If *x*, *y*, *z*, and *w* are not all strictly positive, as utility functions are only defined up to a positive affine transformation, we can transform them so that they are. Then

$$R = \lim_{t \to \infty} [a(e_{1,1} \cdot o_{1,r} + (1 - e_{1,1}) \cdot o_{2,r}) + a^2(e_{1,2} \cdot o_{1,r} + (1 - e_{1,2}) \cdot o_{2,r}) + \ldots + a^t(e_{1,t} \cdot o_{1,r} + (1 - e_{1,t}) \cdot o_{2,r}) + \ldots]$$

where $e_{1,t}$ is the probability that Δ draws the outcome o_1 in round t and $(1 - e_{1,t})$ is the probability that Δ draws the outcome o_2 in round t. Notice that if Δ operates deterministically giving the outcome o_1 in round t, $e_{1,t} = 1$; if Δ operates deterministicly giving the outcome o_2 in round t, $e_{1,t} = 0$; any compound lottery over o_1 and o_2 determining the outcome in round t can be reduced to a lottery of the assumed form. Row's payoff series can be partitioned into the following sub-series:

[47] T. Hobbes, *Leviathan*, edited by C. B. Macpherson (Harmondsworth, Middx.: Penguin, 1968), p. 196, pp. 204–5.

[48] H. Ward, 'Testing the Waters: Taking Risks to Gain Reassurance in Public Goods Games', *Journal of Conflict Resolution*, 33 (1989), 274–308, p. 281.

[49] For reviews of the relevant literature and related topics, see H. Nurmi, 'Problems in the Theory of Institutional Design: An Overview', *Journal of Theoretical Politics*, 5 (1993), 523–40; R. E. Goodin, ed., *The Theory of Institutional Design* (Cambridge: Cambridge University Press, 1995).

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

$$o_{1,r} \cdot [ae_{1,1} + \ldots + a^t e_{1,t} + \ldots] \text{ and}$$
$$o_{2,r} \cdot [a(1 - e_{1,1}) + \ldots + a^t(1 - e_{1,t}) + \ldots].$$

The sequence formed by the partial sums of the first subseries up to the t-th term is increasing and bounded above by $ao_{1,r}/(1 - a)$ as $o_{1,r}$ and $o_{2,r}$ are strictly positive and $e_{1,t} \leqslant 1$ for all t. Hence the sequence has a limit, say $o_{1,r}$S1. Similarly the sequence formed by the partial sums of the second subseries has a limit, say $o_{2,r}$S2. As the subseries have limits and R is the limit of their sum, $R = o_{1,r} \cdot \text{S1} + o_{2,r}\text{S2}$. Now S1 + S2 is equal to

$$\lim_{t \to \infty} [a(e_{1,1} + 1 - e_{1,1}) + \ldots + a^t(e_{1,t} + 1 - e_{1,t}) + \ldots] =$$
$$\lim_{t \to \infty} [a + \ldots + a^t + \ldots] = a/(1 - a)$$

Thus we can write

$$R = (o_{1,r},\ a/(1 - a) \cdot \text{W1}) + (o_{2,r} \cdot a/(1 - a) \cdot \text{W2}),$$

where $\text{W1} = (1 - a)\text{S1}/a$, $\text{W2} = (1 - a)\text{S2}/a$ and $\text{W1} + \text{W2} = (1 - a)\ (\text{S1} + \text{S2})/a = 1$.

Similarly:

$$\text{C} = \lim_{t \to \infty} [a(e_{1,1} \cdot o_{1,c} + (1 - e_{1,1}) \cdot o_{2,c}) + a^2(e_{1,2} \cdot o_{1,c} + (1 - e_{1,2}) \cdot o_{2,c}) + \ldots$$
$$+ a^t(e_{1,t} \cdot o_{1,c} + (1 - e_{1,t})o_{2,c}) + \ldots],$$

and

$$o_{1,c} \cdot [ae_{1,1} + \ \ldots + a^1 e_{1,t} + \ldots] = o_{1,c} \cdot \text{S1}$$
$$o_{2,c} \cdot [a(1 - e_{1,1}) + \ \ldots \ + a^t(1 - e_{1,t}) + \ldots] = o_{2,c} \cdot \text{S2},$$

so that

$$C = (o_{1,c} \cdot a/1 - a) \cdot \text{W1}) + (o_{2,c} \cdot a/(1 - a) \cdot \text{W2}).$$

W1 cannot be equal to zero if P is a taking it in turns pattern because o_1 is the turn of one of the two players, and part (iii′) of the definition of taking it in turns requires that there is a non-zero probability of each player getting her turn in the supergame as a whole.

Extension of Proposition 1

It is easy to see that the above argument can be repeated starting from any round in the supergame. Let R(t) and $C(t)$ be the players' supergame payoffs under P from the t-th round onwards, $t > 1$. Then

$$\text{R}(t) = (o_{1,r} \cdot a/(1 - a \cdot \text{W1}(t)) + (o_{2,r} \cdot a/(1 - a) \cdot \text{W2}(t))$$

and

$$\text{C}(t) = (o_{1,c} \cdot a/(1 - a) \cdot \text{W1}(t)) + (o_{2,c} \cdot a/(1 - a) \cdot \text{W2}(t)),$$

for weights W1(t) and W2(t) summing to 1. The proof is the same as that of Proposition 1, but summations start in round t rather than round 1.

Lemma 1. Suppose P is a taking-it-in-turns pattern belonging to one of the classes {CD, DC}, {CD, CC} or {DC, CC} and that the supergame strategies pair S and T embody grim threats in relation to the device Δ that generates P. Consider an alternative strategy for one of the players, say row. Suppose this alternative strategy S′ is not equivalent to S when played against T and does not always conform to Δ. Then if the stage-game is (2×2) Chicken, S′ can only be a best response to T if it fails to conform by playing D in a round in which Δ implies both players play C; and if the stage-game is (2×2) PD, S′ can only be a best response to T if it fails to conform in a round in which Δ implies row should play C.

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

So long as S and S′ continue to lead to the same action being chosen by row, column is not provoked to retaliate with its grim strategy of T. So column continues to conform to Δ. Hence row is the first to fail to conform. Failure to conform may mean: (i) row plays C when Δ says that he should play D and column should play C; (ii) row plays D when Δ says that both players should play C; (iii) row plays D when Δ says that he should play C and column should play D.

Whether the stage-game is (2×2) PD or (2×2) Chicken, it is never rational for row to fail to conform in the first way. To see this for the case where the stage-game is (2×2) PD consider a strategy S″ which: plays exactly the same as S′ in any rounds before S′ fails to conform for the first time; plays D in the round, t', in which S′ fails to conform; and plays D in each subsequent round. S″ must have a higher payoff than S′: for rounds $t < t'$ if any, S″ and S′ get the same payoff; in round t', S″ gets y while S′ gets x; for rounds $t > t'$, S′ gets at most w while S″ never gets less then w. To see this for the case where the stage-game is (2×2) Chicken consider the strategy S* which plays exactly the same way as S′ in any rounds before S′ first fails to conform, plays D in round t' in which S′ first fails to conform; then plays Cheat1 starting in the next round, where Cheat1 is as defined in the text. Then S* must have a higher payoff than S′: for rounds $t < t'$ if any, S* and S′ have the same payoff; in round t', S* gets y whereas S′ gets x; for $t > t'$, S′ gets at most z while S* never gets less than z in the same round.

If the stage-game is (2×2) Chicken, it is irrational for row to fail to conform in the third way under a strategy S′. Consider the strategy S^+ which plays the same way as S′ up to round t', plays C in round t' and then plays Cheat1. S^+ and S′ get the same payoff up to round t'; S^+ gets the higher payoff of z and S′ gets the lower payoff of w in round t'; after round t', S^+ never gets less than z and S′ never gets more than z in any round. □

Lemma 2. Let J's supergame strategy be T. Suppose for rounds $t > 1$, J's action is conditional on *I*'s action in round $(t-1)$. Subgame $g(C,t)$ is reached if I plays C in round $(t-1)$ and subgame $g(D,t)$ is reached if I plays D in round $t-1$. Now consider some $t' > t$. Subgame $g(C,t')$ is reached if I plays C in round $t'-1$ and subgame $g(D,t')$ is reached if I plays D in round $t'-1$. As the game is infinite, $g(C,t)$, $g(C,t')$, $g(D,t)$, and $g(D,t')$ all have the same extensive form, and we can talk of equivalent nodes of any two subgames. Now suppose that the continuation of J's strategy T in $g(C,t)$ implies the same action at each node of $g(C,t)$ where J acts as the continuation of T in $g(C,t')$ implies at the equivalent node of $g(C,t')$. Further, suppose that the continuation of strategy T in $g(D,t)$ implies the same action at each possible node of $g(D,t)$ where J acts as the continuation of T in $g(D,t')$ implies at the equivalent node of $g(D,t')$. Then I's best action is a in round $t'-1$ if and only if it is a in round $t-1$.

Choosing C at $t-1$ or $t'-1$ has the same consequences for J's actions and, given the discount rate is constant, for I's continuation payoffs from any given subgame strategy in the subgames $g(C,t)$ and $g(C,t')$. The same statement holds with C replaced by D throughout. Therefore the decisions at times t and t' are strategically equivalent. □

Proof of Proposition 3

Let I″ be an alternative strategy to I which is not the same as I when played against I′ and fails always to conform to Δ. Lemma 1 shows that it can only be rational for row to fail to conform in a round where he should play C and column should play D. Suppose that under I″ it is planned to conform in the first round. Then I″ can only be a rational response to I′ if it always conforms, so that it is equivalent to I. Let round $t > 1$, be another odd round when row could fail to conform by playing D when Δ says he should play C and column should play D. In the subgames $g(D, 2)$ and $g(D, t+1)$ column always plays D at each corresponding pair of nodes as he retaliates for row failing to conform. If row is to conform in the first round and in round t he must plan to play C in both these rounds. In the subgames $g(C, 2)$ and $g(C, t+1)$ for any equivalent pair of nodes which is on the path through the game tree associated with Δ, column plays C in even rounds and D in odd rounds. At any equivalent pair of nodes off the path associated with Δ, column always plays D, as I′ retaliates. Hence by Lemma 2 if it is rational to play C in the first round, it is rational to play C in round t. If I″ breaks the pattern and is a best response

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

to I′, it must fail to conform in the first round. As I′ embodies the grim threat, column always plays D after this round and row's best reply to this is D in every subsequent round. Hence if I′ fails always to conform to D and is a rational response to I′, it must play D in the first and all subsequent rounds, which is equivalent to D^{∞}. Similarly: by Lemma 1 it can only be rational for column to fail to conform in an even round; by Lemma 2, if it is rational to fail to conform at all, this must be done in the second round; and column must play D in each round after the pattern is broken; so that the only alternative to I′ which might be a rational response to I is D^{∞}. It is sufficient for (I, I′) to be an equilibrium that neither player can strictly increase her payoff by changing to D^{∞} given the other player's strategy remains the same. That this is a necessary condition for equilibrium follows by definition of the Nash equilibrium.

Row cannot strictly increase his payoff by changing to D^{∞} if $a(z + ay)/(1 - a^2) \geqslant aw/(1 - a)$. As the first-round payoffs from I′ and D^{∞} are both ay, column cannot strictly increase his payoff by changing from I′ to D^{∞} if $a^2(z + ay)/(1 - a^2) \geqslant a^2w/(1 - a)$ which is the same as the condition on row's payoffs. □

Proof of Proposition 4

Suppose that row is the player who always co-operates. Let S′ be a strategy which is not equivalent to S when played against T and fails always to conform to Δ. Lemma 1 shows that S′ can only rationally fail to conform in a round in which Δ tells both players to play C. Suppose that under S′ it is planned to conform in the first round, t, in which Δ tells both players to play C. Then S′ can only be a rational response to T if it always conforms, so that it is equivalent to S. Let round t', $t' > t$, be another round in which row could fail to conform by playing D when Δ tells both players to play C. In the subgames $g(D, t + 1)$ and $g(D, t' + 1)$ column always plays D at each corresponding pair of nodes as he retaliates for row breaking the pattern. If row is to conform at t and t' he must plan to play C in both these rounds. In the subgames $g(C, t + 1)$ and $g(C, t' + 1)$ for any equivalent pair of nodes which is on the path through the game associated with Δ, column plays C (hence D) with the same probability, because P is a simple fixed lottery. At any equivalent pair of nodes off the path associated with Δ, J always plays D, because his strategy T retaliates for the breaking of the pattern. Hence by Lemma 2 if it is rational to play C at t it is rational to play C at t'. Hence if S′ ever fails to conform and is a best response to T, it must do so the first time Δ tells both players to play C. As T embodies the grim threat in relation to Δ, column always plays D after row fails to conform and row's best reply to this is C in every round after the pattern is broken. Hence if S′ is a rational response to T which ever fails to conform it must conform to Δ until the first time it tells both players play C; play D in that round; and C in all subsequent rounds. This is equivalent to Cheat1. The same argument applies to column. Hence it is sufficient for (S, T) to be an equilibrium that neither player can strictly increase his payoff by changing to Cheat1 given the other player's strategy remains the same. The necessity of this condition follows from the definition of Nash equilibria.

Let e, $0 < e < 1$, be the probability that Δ draws CD in any round. Up to the first time t when Δ tells both players to play C, S and Cheat1 have the same payoff; in round t, S gets x and Cheat1 gets y; in subsequent rounds the expected payoff of S is $ez + (1 - e)x$ per round and the payoff of Cheat1 is z per round. It does not pay row to switch to Cheat1 if $a^t x + a^{t+1}(ez + (1 - e)x)/(1 - a) \geqslant a^t y + a^{t+1}z/(1 - a)$ or $x + a(ez + (1 - e)x)/(1 - a) \geqslant y + az/(1 - a)$. Similarly, it does not pay column to switch to Cheat1 if $x + a(ey + (1 - e)x)/(1 - a) \geqslant y + az/(1 - a)$. As $z < y$, if the condition on row's payoffs is satisfied, so is the condition on column's payoffs. □

Proof of Proposition 5

The logic of the proof is similar to that of Proposition 4 and, for brevity, a sketch will be given. By Lemma 1, it can only be rational for either player to fail to conform by playing D when

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

Δ tells him to play C and the other player to play D. If a player does not fail to conform the first time that Δ calls on him to play C, by Lemma 2, it will never be rational to do so. Thus any alternative response to the other player's strategy which is rational must play D up until Δ says that you should play C, plays D in that round, and D thereafter, so that it is equivalent to D^{∞}. Thus, it is sufficient for (S, T) to be an equilibrium that neither player can strictly increase his payoff by changing to D^{∞}. The necessity follows from the definition of a Nash equilibrium.

Let e and $(1-e)$, $0 < e < 1$, be the probabilities that Δ draws CD and DC, respectively. Up to the first time, t, at which Δ tells row to play C, S and D^{∞} have the same payoff; in round t, S gets z and D^{∞} gets w; in subsequent rounds the expected payoff of S is $ez + (1-e)y$ per round and the payoff of D^{∞} is w per round. It does not pay row to switch to D^{∞} if $z + a(ez + (1-e)y)/(1-a) \geqslant w + aw/(1-a)$. Similarly, it does not pay column to switch from T to D^{∞} if $z + a(ey + (1-e)z)/(1-a) \geqslant w + aw/(1-a)$ The condition on row's payoffs is at least as hard to satisfy as that for column so long as $e \geqslant \frac{1}{2}$. □

Proof of Proposition 6

Again a sketch will suffice. Let the player who always plays C be row. By Lemma 1, it can only be rational for column to fail to conform by playing D when Δ says that both players should play C. By Lemma 2, if column does not break the pattern the first time Δ calls on both players to play C, it will never be rational for him to break the pattern. Thus any alternative response to row's strategy which is rational must play D up until Δ tells column to play C, play D in that round, and D thereafter, so that it is equivalent to D^{∞}. T is at least as good if not better than D^{∞} against S if $x + a(ey + (1-e)x)/(1-a) \geqslant y + aw/(1-a)$ … C(i). By Lemma 1, it can only be rational for row to fail to conform by playing D when Δ says that both players should play C or by playing D when Δ says that he should play C and column should play D. If row has not failed to conform by the time that it has called on him to play in both these ways, by Lemma 2 it will never be rational for him to fail to conform. Thus it is sufficient for S to be at least as good if not better than any other strategy against T that the following three conditions should be satisfied: (i) S is at least as good if not better against T than Cheat2 which plays like S up to the first time that Δ tells row to play C and column to play D, plays D in that round, and all subsequent rounds, or $z + a(ez + (1-e)x)/(1-a) \geqslant w + wa/(1-a)$ … C(ii); (ii) S is at least as good if not better against T than Cheat3 which plays like S up to the first time that Δ tells both players to play C, plays D in that round and in all subsequent rounds, or $x + a(ez + (1-e)x)/(1-a) \geqslant y + wa/(1-a)$ … C(iii); (iii) S is at least as good if not better against T than a strategy which plays D in the first round – i.e the first time that D tells row to play C and column to play D or both players to play C – then plays D^{∞} or $ez + (1-e)x + a(ez + (1-e)x)/(1-a) \geqslant (ew + (1-e)y) + aw/(1-a)$ … C(iv).

By multiplying both sides of C(ii) by e and both sides of C(iii) by $(1-e)$ then adding the right-hand sides of the resulting conditions then adding the left-hand sides of the resulting conditions, it is easily seen that C(iv) is true if C(ii) and C(iii) are both true. C(iii) implies C(i) as $z < y$. C(ii) implies C(iii) if $(y-x) < (w-z)$ and c(iii) implies c(ii) if $(y-x) > (w-z)$. Thus if C(ii) holds and $(y-x) < (w-z)$ or if c(iii) holds and $(y-x) \geqslant (w-z)$, all four conditions hold. □

Proof of Proposition 7

By Lemma 1 if it is rational to fail to conform to Δ at all, it will be in a round when it tells you to play C and the other player to play D. Moreover, after first failing to conform, to be rational you must play D in every round. By the extension of Proposition 1 the value of row and column's payoffs if P continues from round $t+1$ onwards, considered from the perspective of round t, $t \geqslant 1$, can be written as $a(W1(t)z + W2(t)y)/(1-a)$ and $a(W1(t)y + W2(t)z)/(1-a)$ respectively for weights $W1(t)$ and $W2(t)$ summing to 1. Thus it will pay row to conform in round t if Δ tells him to play C and column to play D long as $z + a(W1(t)z + W2(t)x)/$

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press

$(1-a) \geqslant w + aw/(1-a)$ … C(i). It will pay column to conform in round t if Δ tells him to play C and row to play D so long as $z + a(\mathrm{W1}(t)y + \mathrm{W2}(t)z)/(1-a) \geqslant w + aw/(1-a)$ … C(ii).

It is sufficient for (S, T) to be an equilibrium that C(i) and C(ii) are satisfied for all t. For each t define W(t) as W1(t) if W1(t) $\geqslant$ W2(t) and $W2(t)$ if W2(t) $>$ W1(t). Then if $z + a(\mathrm{W}(t)z + (1 - \mathrm{W}(t))y)/(1-a) \geqslant w + aw/(1-a)$ … C(iii) holds, so will c(i) and c(ii). Now let t^* be the value of t at which W(t) is maximal. Note that W(t) < 1, as in an infinite supergame each player has a probability of greater than 0 of having a turn in future rounds, no matter how many rounds have already been played, according to part (iii′) of the definition of taking it in turns. Also, by construction W(t^*) $\geqslant {}^1/_2$. If C(iii) holds for $t = t^*$, it holds for all t. C(iii) for $t = t^*$ is necessary and sufficient for the simple lottery P* belonging to the class {CD, DC} with $e = \mathrm{W}(t^*)$ to be an equilibrium under strategies embodying grim threats, as shown in Proposition 5. □

Proof of Proposition 8

Let row be the player who always co-operates. By Lemma 1, if the stage-game is (2 × 2) Chicken, if it is rational for row to fail to conform at all it must be in a round in which D tells both sides to play C. Also, after first failing to conform he must always play C. By Lemma 1, if the stage-game is (2 × 2) PD, it may be rational for row to fail to conform in a round in which either D tells both players to play C or row to play C and column to play D. Either way, row must play D after first failing to conform. By the extension of Proposition 1 the value of row and column's payoffs if P continues from round $t-1$ onwards, considered from the perspective of round t, $t \geqslant 1$, can be written as $a(\mathrm{W1}(t)z + \mathrm{W2}(t)x)/(1-a)$ and $a(\mathrm{W1}(t)y + \mathrm{W2}(t)x)/(1-a)$, respectively, for weights W1(t) and W2(t) summing to 1. Thus it will pay row to conform in round t if Δ tells both players to play C so long as $x + a(\mathrm{W1}(t)z + \mathrm{W2}(t)x)/(1-a) \geqslant y + as/(1-a)$ … C(i), where s is row and column's security level, i.e. w if the stage-game is (2 × 2) PD and z if the stage-game is (2 × 2) Chicken. It will pay row to conform in round t if Δ tells him to play C and column to play D so long as $z + a(\mathrm{W1}(t)z + \mathrm{W2}(t)x)/(1-a) \geqslant w + as/(1-a)$ … C(ii). By Lemma 1, if it is rational for column to fail to conform at all it will be by playing D in a round in which Δ tells both players to play C. Also, after this column must always play D if the stage-game is (2 × 2) PD and C if the stage-game is (2 × 2) Chicken. It will pay column to conform in round t if D tells both players to play C so long as $x + a(\mathrm{W1}(t)y + \mathrm{W2}(t)x)/(1-a) \geqslant y + as/(1-a)$ … C(iii).

If C(i) is satisfied so is C(iii), as $z < y$. It is easily shown that C(i) implies C(ii) if $(y-x) > (w-z)$ and C(ii) implies C(i) if $(y-x) \leqslant (w-z)$. Note that, if the stage-game is (2 × 2) Chicken $(y-x) > (w-z)$ as $z > w$; C(ii) is always satisfied as $s = z$ and $z > w$. Let t^* be the value of t for which W1(t) is maximal. Call C(i) the critical condition if the stage-game is (2 × 2) Chicken or (2 × 2) PD with $(y-x) > (w-z)$ and C(ii) the critical condition if the stage-game is (2 × 2) PD and $(w-z) > (y-x)$ Then if the critical condition is satisfied for $t = t^*$, C(i), C(ii) and C(iii) are satisfied for all t and (S, T) is an equilibrium. Also, if the critical condition is satisfied for $t = t^*$, a simple lottery of the same class with $e = \mathrm{W1}(t^*)$ is stable under grim threat strategies, according to Propositions 4 and 6. □

https://doi.org/10.1017/S0007123400580004 Published online by Cambridge University Press