Trust and Cooperation Through Agent-specific Punishments

Fiona McGillivray and Alastair Smith

We have no quarrel with the people of Yugoslavia. . . . Our actions are directed against the repressive policy of the Yugoslav leadership.

President William Clinton, 24 March 1999

The language of Realpolitik focuses on the pronoun *them,* referring to a nation and its people as a whole, single entity. Yet contemporary declarations of foreign policy increasingly rely on the pronoun *him* or *her.* U.S. leaders often claim that the United States is "friends" with the people of a country and only has a quarrel with its leadership.[1] Clinton's statement is intended to encourage Serbia's people to oppose, even oust, the Yugoslav leadership. It implies that once the Yugoslav leadership is removed, cooperative relations will resume.

We explore how focusing policy against a leader, the agent of the people, rather than against the principal (the nation and its people) that he or she represents, affects the interactions between nations. Although the physical implementation of either form of policy is the same, we argue that, above and beyond the rhetorical effect, the identity of the target creates substantive differences in how nations interact. Adopting policies that specify individuals, rather than the nation as a whole, as the target of punishment can bolster trust, reduce the fragility of cooperation, and prevent festering relations by providing a mechanism for restoring cooperation.

Extant theories of cooperation suggest that nations maintain trust and cooperation by threatening to punish exploitative behavior through the removal of future cooperation. This threat of punishment prevents exploitation because a nation that defects today forgoes the benefits of cooperation tomorrow. While maintaining the intuition of this approach, our key insight derives from examining the effect of agent-specific

The ideas in this article were inspired by a conversation with Helen Milner—our thanks to her. We also thank David Cameron, Matt Gabel, Geoffrey Garrett, Pauline Jones Luong, Bruce Russett, Andy Sobel, Alan Stam, Andrew Stigler, and several anonymous reviewers for their comments. We presented an earlier version of this article at the 1998 Peace Science Society annual meeting and at the Yale University International Relations Reading Group; our thanks to the participants for their comments.

1. Woodrow Wilson's message to Congress on 2 April 1917 included a similar theme: "We have no quarrel with the German people." Cited in Russett 1993, 3.

International Organization 54, 4, Autumn 2000, pp. 809–824

punishments directed at a specific leader rather than at the nation as a whole. In particular, we look at strategies where punishments continue only as long as the incumbent leader remains in power. Although the concept of agent-specific punishment is germane to a wide range of interactions, we develop our formal model within the strict confines of international cooperation.[2]

International Cooperation

Scholars typically use the prisoners' dilemma game as a metaphor for international cooperation.[3] Nations behaving myopically can never cooperate in this game; the incentives to defect and exploit the other party dominate. Yet, despite the temptation to defect, nations can cooperate through conditioning future cooperation on current behavior. By threatening to withdraw future cooperation as punishment for exploitive behavior, nations make their partners trustworthy. Provided that the long-term benefits of cooperation outweigh the short-term gains from exploiting a partner, conditional punishment strategies make cooperation possible. The grim trigger and tit-for-tat strategies are well-known examples. In the grim trigger strategy a nation cooperates with its partner only if it has never been exploited in the past. Against this strategy a nation can exploit its partner's trust once, but it does so at the expense of forgoing all future cooperation. Unfortunately these mechanisms poorly account for the patterns of interactions between states, and scholars increasingly note the role of domestic institutions in shaping these patterns.[4]

A long-standing practice of U.S. leaders is to claim they are "friends" with the people of a country and only have a quarrel with the country's leadership. As testimony to this fact, few people believe that the current U.S. policy of targeting Iraq will continue once its incumbent leader, Saddam Hussein, leaves office. Statements by President Clinton at the start of Operation Allied Force in Kosovo also explicitly point to the agent-specific nature of U.S. policy; for example, "I cannot emphasize too strongly, the United States and our NATO allies have no quarrel with the Serbian people."[5]

Anecdotal evidence suggests that this phenomenon is not restricted to democratic states. For example, following the Gulf War, the Gulf states severed economic ties with Jordan because of its refusal to enforce sanctions on, and to some extent its passive support for, Iraq. On the death of King Hussein (7 February 1999), these states quickly stepped in with offers of monetary and trade support for the new regime.[6] Yet the accession of Hussein's son, Abdullah, suggests few, if any, substan-

2. Guisinger and Smith 1999 and Smith 1999 develop the logic in the context of international crises.

3. See Axelrod 1984; Axelrod and Keohane 1986; Fearon 1998; Gourevitch 1996; Milner 1992; and Pahre 1994.

4. See Gaubatz 1996; Gowa 1994; Leeds 1999; Mansfield, Milner, and Rosendorff 1998; Martin 1993; McGillivray 1999; Milner 1997; Milner and Rosendorff 1997; Oneal and Russett 1997; Remmer 1998; Rheinhardt 1996; Russett et al. 1998; Schultz and Weingast 1998; and Verdier 1998.

5. 26 March 1999.

6. Neighbors Rally to Jordan, Easing Financial Fears, *New York Times*, 19 February 1999, A3.

tive policy changes. It appears that the clock is simply reset on inauguration of a new leader, and cooperation recommences. Such a pattern of cooperative behavior is poorly accounted for by extant unitary actor models of international cooperation.

We argue that the empirical repercussions of agent-specific policies depend on domestic political institutions. If foreign policies are directed against agents, then nations are punished only as long as their miscreant agents remain in office. The principals can end the punishment and restore cooperation simply by replacing the responsible agent. Hence, against leader-specific policies, citizens have an incentive to remove any wrongdoing agent. For the punisher, this is one of the attractions of agent-specific policies. The observed pattern of interaction between states depends on the ease of domestic removal. If replacing agents is difficult, then once cooperation falters, agent-specific punishment policies often lead to prolonged hostilities and periods of acrimonious or bitter relations between states. As NATO's Supreme Allied Commander in Europe, General Wesley Clark, stated, "it is a real political problem for the people of Yugoslavia because I think world leaders have made very clear that they don't see Yugoslavia really being readmitted into the European Community of nations or receiving the kinds of reconstruction that it really needs while he's [Milosevic] still in place as the President."[7]

In the case of Jordan, Hussein's failure to impose sanctions on Iraq meant the cessation of aid from its Arab neighbors for the eight years following the Gulf War. The earlier the king's removal, the earlier the resumption of international cooperation. In autocratic states, a leader's death is often the trigger for rehabilitating international relations.

In polities with accountable agents, instances of punishment are less common and shorter in duration. Once an agent has transgressed, punishment continues until that agent is replaced. The ease with which cooperation can be restarted provides principals with an impetus to remove agents whose actions lead to the breakdown in cooperation. Principals can escape the costs of punishment by replacing their agents. Hence, we expect accountable agents to be quickly removed and cooperation restored. As such, instances of punishment are likely to be short in duration. Punishment is also likely to be rare, since office-seeking agents want to avoid contingencies that lead to their removal. Agents that violate international norms not only expose their nation to the wrath of others but also lose their jobs in the process.

Domestically accountable agents are more "trustworthy" than their less accountable counterparts. When a nonaccountable agent subjects their nation to the rancor of other nations, the agent suffers no more than the nation as a whole.[8] An accountable agent, however, faces the additional distress of ouster. The risk of expulsion means that domestically accountable agents pay higher costs for incurring the ire of other nations. Therefore, they are less likely to do so and can be trusted to honor agree-

7. Comment to the BBC World Service, 20 July 1999.

8. In contrast, Pape, among others, claims that the anger of the international community often enhances the status of autocratic leaders. Pape 1997. While we would argue that public demonstrations of support are a consequence of a leader needing to work harder to maintain his or her position, the key fact is that autocrats typically remain in power and their nations remain ostracized.

ments to a greater extent, since not to do so endangers their domestic political careers. As such, polities with accountable agents are more trustworthy and achieve higher levels of cooperation.

In summary, agent-specific punishments suggest that changes in leadership affect the patterns of behavior between states. Furthermore, the accountability of agents influences the patterns of behavior we expect to see. Before developing our arguments formally, we outline the principal-agent relationship through which we model domestic politics. Although for our purposes here we focus predominantly on nations, the theory applies equally in other contexts where principals hire agents to make decisions on their behalf. Thus, our theory is as applicable to the interaction of firms as to the interaction of nations.

Domestic Political Arrangements

The key feature of domestic political arrangements is that the citizens are represented by a leader who determines policy. In the principal-agent context, the citizens are the principal and the leader is their agent. Throughout we use the terms *leader* and *agent* interchangeably. We assume that the leader is solely responsible for choosing and implementing policy. We recognize that in reality it is often difficult to identify precisely who is responsible for policy choice. However, we leave to future research the question of whether the responsible agent is an individual leader, a coalition cabinet, a political party, or an entire bureaucracy.[9] For the present we assume that there is a clearly identified leader, or agent, of the people.

We distinguish between two types of political systems according to the ease with which leaders can be replaced. In one type of system leaders survive in office despite disastrous performance; in the other type even successful leaders are in jeopardy.[10] Although in reality all leaders are accountable to at least some extent and all leaders possess some incumbency advantage, initially we separate regimes as accountable or unaccountable. The key features of our principal-agent setup are as follows:

1. Agents have the authority to make decisions. Each nation is represented by an agent (leader) who is solely responsible for policy decisions.
2. Although all agents are at least partially accountable to their principals, domestic political institutions determine the extent of this accountability. In particular, we assume domestic political arrangements determine the cost the principal must pay to remove an agent. When the cost is low, we label the leader accountable. When the cost is high, we label the leader unaccountable. (Later we derive the precise cost of removal that distinguishes between these two cases.)

9. Powell and Whitten 1993; Lijphart 1990; Roubini and Sachs 1989; and Laver and Schofield 1990.

10. The outcome of the Gulf War provides an example of this. Saddam Hussein survives in office despite military defeat and the continuance of crippling economic sanctions against Iraq. In contrast, the victorious George Bush was displaced by a slight slowdown in the U.S. economy. More generally, Bueno de Mesquita and Siverson show how regime type moderates the domestic consequences of international outcomes. Bueno de Mesquita and Siverson 1995.

		Nation 2	
		Cooperate, C	Defect, D
Nation 1	Cooperate, C	R, R	S, T
	Defect, D	T, S	P, P

Where $T > R > P > S$

FIGURE 1. *The prisoners' dilemma game*

Cooperation in the Prisoners' Dilemma

Cooperation between nations is most commonly discussed in terms of the prisoners' dilemma interaction. We stay within this genre. Given the profusion of explanations of the prisoners' dilemma game in the political science literature, we dispense with a lengthy introduction and assume the reader is familiar with the model. In the prisoners' dilemma (see Figure 1) each player chooses between two strategies, C (cooperate) and D (defect). The payoffs associated with the outcomes are such that $T > R > P > S$ and $R > (T + S)/2$, where T is the temptation payoff for exploiting a partner's cooperation, R is the reward for mutual cooperation, P is the punishment payoff from failing to cooperate, and S is the sucker's payoff from unilateral cooperation. The unique single-shot Nash equilibrium is both players defecting (D, D). Although the outcome of both cooperating, (C, C), is Pareto superior ($R > P$), both players have a dominant strategy to defect.

In the single-shot context, the prospects for cooperation are dismal, yet in the repeated-play context, cooperation is possible if nations condition future play on past behavior. Grim trigger is a common example of such a strategy. Under the grim trigger, nations initially cooperate (play C) and continue to do so provided that defection is never observed. If either player ever defects (play D), then nations refuse to cooperate (play D) in every subsequent period. Since the threat of punishment, the indefinite removal of cooperation, is the harshest possible penalty for noncooperation, the grim trigger strategy represents the limiting case.[11] If this mechanism is insufficient to support cooperation between two unitary actors, then no such mechanism exists.

Before integrating domestic politics into our explanation of international cooperation, we revisit the mathematics behind cooperation within the grim trigger strategy. Reiterating this material serves several purposes. First, it helps place our explanation

11. While grim trigger represents the limiting case, if nations are sufficiently patient, then there are many other equilibria that support cooperation, a result typically referred to as the folk theorem. See Fudenberg and Maskin 1986.

in context. Second, it provides a baseline prediction against which to compare our results. Third, our model draws extensively on the extant intuition and consequently shares a similar mathematical structure. Given this, by recapping the conventional argument we can develop within a familiar setting the mathematical tools required for our model.

Grim Trigger Strategy

In the grim trigger strategy each nation initially cooperates and continues to do so only as long as defection never occurs. Once defection occurs, nations defect in every subsequent period. The grim trigger strategy is a subgame perfect Nash equilibrium provided that $\delta \geq (T - R)/(T - P)$, where δ, the discount factor, measures the extent to which actors value rewards in the future relative to rewards today (that is, δ measures patience).[12] Under this strategy, nations start cooperating and cooperation is sustained as long as neither nation ever defects. Should defection ever occur, then both nations defect in every subsequent round and cooperation is ended forever. Hence, a single incidence of defection results in the outcome (D, D) in all subsequent rounds. It is this threat of permanently ending cooperation that prevents nations from exploiting each other in the short run.

A brief analysis of the mathematics illustrates the tradeoff between the long-run gains from cooperation and the short-run incentives for exploitation. If both nations play the grim trigger strategy, then they both start cooperating and continue to do so. In each period, nation 1 receives a payoff of R,[13] the current value of which is

$$R + \delta R + \delta^2 \mathrm{R} + \delta^3 \mathrm{R} + \ldots = \sum_{t=0}^{\infty} R\delta^t = R/(1 - \delta).$$

This is the expected value of playing the grim trigger strategy given that the other nation also plays this strategy. Yet, in the short term, nation 1 gets a higher payoff by defecting. Unfortunately, the long-term consequences of doing so are that cooperation never occurs again. Formally, if nation 1 defects, then it receives T today, and, since cooperation is permanently terminated, it receives P in every subsequent period, the current value of which is $T + \delta P + \delta^2 P + \delta^3 P + \ldots = T + \delta P/(1 - \delta)$. If the former payoff remains larger than the latter (a condition that implies $\delta \geq (T - R)/(T - P)$), against the grim trigger strategy nations can do no better than also play grim trigger. Hence, provided that nations are suitably patient, grim trigger allows them to cooperate. Using a standard example of $T = 4$, $R = 3$, $P = 2$, and $S = 1$, cooperation is only possible if $\delta \geq (4 - 3)/(4 - 2) = 1/2$.

12. Technically δ is a number between 0 and 1. Intuitively, $\delta = 0.7$ means that a risk-neutral individual is willing to take seventy cents on the dollar to receive a reward today rather than wait until the next period to receive the full dollar.

13. A useful mathematical result is that the infinite sum $x + \delta x + \delta^2 x + \ldots = x/(1 - \delta)$.

If nations are less patient, then there is no way to maintain cooperation. Yet the condition $\delta \geq (T - R)/(T - P)$ represents the limit of cooperation. Many scholars suggest that the grim trigger strategy is inappropriate because it requires perfect information and no noise.[14] In the real world, difficulties in determining exactly what happened can lead to interpretative mistakes that lead to the end of cooperation.[15] Thus, in practice cooperation is more fragile that the optimistic limiting case suggests. Next we explore how agent-specific punishments reduce the fragility of cooperation and provide a mechanism for restoring cooperation.

Agent-specific Punishments, Domestic Institutions, and Cooperation

We use a simple model of domestic politics in which the leader acts as an agent of the people (the principals). Consistent with most of the literature, we ignore any distributional consequences of international cooperation and assume that all the actors within a state have identical preferences over the outcomes of the prisoners' dilemma.[16] In addition to the payoffs they receive from the international interaction, agents receive a payoff of Ψ for each period that they keep their jobs.[17] At the end of each period, principals can choose to keep the incumbent agent or replace the agent with a new leader at a cost of K.[18] The following is the setup for the modified stage game:

1. The current agents play prisoners' dilemma.
2. The principals in each nation observe the outcome of the agents' choices and then decide whether they want to retain their respective agents or replace them at a cost of K.

In the final stage (no. 2) the ease with which the citizens can remove an incumbent leader depends on domestic political institutions. The democratic process provides voters with a relatively low-cost means of replacing leaders. Yet ousting authoritarian leaders is more costly, often requiring social unrest and possibly even civil war.[19]

14. See, for example, Bendor 1993; Wu and Axelrod 1995; and Signorino 1996.

15. Although the prisoners' dilemma captures the general incentives of nations, the exact interaction might vary in each period. To indicate the consequences of this, suppose that as a result of an exogenous shock the temptation payoff is particularly high for one period. This shock places much greater strain on the condition $\delta(T-R)/(T-P)$. If, for example, nation 1's temptation reward for this particular period is 10, then the maintenance of cooperation requires nation 1 to be really patient, specifically, $\delta \geq (10 - 3)/(10 - 2) = 0.875$.

16. Notable exceptions that model the redistributional effects of trade agreements include Mansfield 1998; Mansfield, Milner, and Rosendorff 1998; Milner 1997; and Milner and Rosendorff 1997.

17. Consistent with much of the formal modeling literature, we assume leaders inherently enjoy holding office. In this context the reward, Ψ, is not a transfer payment and should be thought of as unrelated to other players' payoffs.

18. We assume that leaders are evaluated in every period of the game. However, it is a straightforward generalization of the model to have elections only in certain periods. For example, we could assume that elections occur every fourth period. This could be used to represent elections that occur every four years, while still considering foreign policy as made annually. However, such elaborations add few substantive results and so are ignored here.

19. Although of little substantive relevance, we assume that the next leader is drawn from an infinite pool of candidates. This provides the technical convenience that, once removed, a leader has no chance of returning to office.

As we show later, the magnitude of the cost K determines whether or not leaders are politically accountable.

Next we consider the agent-specific grim trigger (ASGT) strategy and show how, if agents are accountable, it leads to more robust cooperative behavior than the conventional unitary actor approach. The strategy specifies the conditions under which leaders cooperate and defect, and describes the contingencies under which the principals want to remove their agents. We specify the strategy for nation 1. Nation 2 behaves analogously. The ASGT strategy is as follows:

1. The agent in nation 1 cooperates unless this agent or the current agent in nation 2 has ever previously unilaterally defected. If either current agent has unilaterally defected in the past, then the agent in nation 1 defects. (We say that agent 1 unilaterally defects if this agent plays D while agent 2 plays C.)
2. The principals (the citizens) in nation 1 retain their leader provided that the leader has never unilaterally defected against nation 2 (independent of who the leader is in nation 2 at the current time). If the leader in nation 1 has ever unilaterally defected in any previous period, then the principals replace this leader with a new leader.

If both agents are accountable, then, provided that

$$\delta \geq \frac{T - R}{T + \psi - R} \qquad \text{and} \qquad K \leq (R - P)\frac{\delta}{1 - \delta},$$

the ASGT strategy is a subgame perfect Nash equilibrium. Before exploring the substantive implications of this result, we outline aspects of the mathematical proof to this proposition. In the process of doing so we explain the logic behind the equilibrium.

We start by analyzing the principals' decision to remove their agent if the agent has previously unilaterally defected against nation 2. Given the ASGT strategy, once a leader unilaterally defects, no agent in the other nation will ever cooperate with that leader again. As long as the defecting leader remains in power, the outcome will be (D, D) in every period. Thus, if an agent unilaterally defects and the principals retain this agent indefinitely, the nation's payoff is $T + \delta P + \delta^2 P + \delta^3 P + \ldots = T + \delta P/(1 - \delta)$. Given the agent-specific nature of punishments, once an agent is replaced, cooperation restarts. If the principals immediately remove their agent, then their payoff is $T - K + \delta R + \delta^2 R + \delta^3 R + \ldots = T - K + \delta R/(1 - \delta)$.

Provided that the cost of removing a leader is sufficiently small ($K \leq (R - P)\delta/(1 - \delta)$), the principals immediately remove any agent who has unilaterally defected in the past.[20] If the cost of removal is higher than this threshold, then the benefits of restoring cooperation are insufficient to offset the one-time cost of replacing an incumbent. The threshold $[K \leq (R - P)\delta/(1 - \delta)]$ distinguishes between politically accountable and unaccountable agents. When the cost of removing a leader is low, it is clearly optimal for the citizens to remove any agent who has unilaterally defected, since this

20. Technically, the proof only requires consideration of single-period, rather than indefinite, defections from the equilibrium path; however, this generates identical conditions.

restores cooperation. Politically accountable leaders, those for whom the cost of removal is less than $(R - P)\delta/(1 - \delta)$, face severe domestic consequences, or audience costs, from failing to cooperate.

In equilibrium, both agents always cooperate. Hence, the principals receive the payoff associated with cooperation in every period, which is worth $R/(1 - \delta)$. Interestingly, if the agent is politically accountable, then the principals might be better off receiving a payoff of $T - K + \delta R/(1 - \delta)$ if their agent defects. Substantively, this means that voters might actually want their leaders to cheat.[21] Yet this has the perverse consequence that leaders who carry out these popular policies are subsequently punished electorally. While the public may favor reneging on agreements, abrogation is political suicide for accountable agents. As we shall show, the threat of this audience cost means that agents honor agreements and cooperate internationally, even though doing otherwise might be more popular.

We now formally examine the incentives of agents. As shown earlier, under the ASGT strategy, if an agent unilaterally defects, then the agent is removed from office. Suppose neither agent has ever defected in the past. According to the ASGT strategy, an agent should cooperate. If the agent does so, then the nation receives R in the first period. In addition, the agent receives a payoff of Ψ associated with holding office. Thus, the agent's total first period payoff is $R + \Psi$. In equilibrium, the agent continues to receive this payoff in every period, which is worth $(R + \Psi) + \delta(R + \Psi) + \delta^2(R + \Psi) + \ldots = (R + \Psi)/(1 - \delta)$. Yet, if the agent defects, then the agent's immediate reward is larger. However, under the ASGT strategy, the principals remove the agent at the next election. Although such a leader benefits from future cooperation (since under ASGT the agent's successor will restore cooperation), given the agent's removal the agent loses office-holding benefits in future periods. Hence, an agent's payoff from defecting is $(T + \Psi) + \delta R + \delta^2 R + \ldots = (T + \Psi) + \delta R/(1 - \delta)$. This represents being in office and getting the temptation payoff in the first period but then being replaced by a new agent who cooperates in every subsequent period.

An agent cooperates if the reward from doing so outweighs the short-term benefits of defection. This condition is satisfied when

$$\delta \geq \frac{T - R}{T + \Psi - R}.$$

Provided that accountable agents care sufficiently about keeping their jobs (large Ψ), the ASGT strategy facilitates cooperation even between relatively impatient nations.[22]

21. This suggests the possibility of collusion. For example, if the citizens promised to compensate a leader for the lose of office if the leader defected on their behalf, then the citizens could both have their cake and eat it. Yet such collusion is impossible. Even ignoring the citizens' commitment problem in compensating their former leaders, such collusion undermines the desire of other nations to cooperate, since they know they will be cheated.

22. We have not shown the optimality of defecting once defection has occurred in the past. We omit a formal discussion of this eventuality because it is straightforward.

Political Accountability and International Cooperation

The mathematical analysis shows that the conditions under which cooperation is possible in the infinitely repeated prisoners' dilemma depends on the extent to which leaders are politically accountable. Next we examine the substantive implications of these results by exploring three topics: the robustness of cooperation, the domestic political consequences of cooperation or its failure, and the pattern of cooperative behavior between states.

When agents are accountable, the ASGT strategy facilitates cooperation between nations through two mechanisms. First, accountable agents who value office holding are reluctant to defect, since they know their jobs depend on their ability to cooperate in future periods. As we shall explore in a moment, this means that cooperation between accountable agents can be sustained under a much wider range of circumstances than when agents are not accountable or when punishments are targeted at principals rather than at agents. Second, since punishments are targeted at individual agents rather than the people they represent, should cooperation break down it can be readily restored by the domestic removal of the defecting agent. Hence, the ASGT's promotion of cooperation is twofold: cooperation is more robust—occurring under a wider range of conditions—and cooperation, once interrupted, is easily reestablished.

Cooperation within the prisoners' dilemma requires both parties to trust each other. Unless both agents are politically accountable, the superiority of the ASGT strategy is lost. When an agent is unaccountable, even if leaders use ASGT punishment strategies, the pattern of behavior, and the conditions under which cooperation can be supported, are identical to those under the conventional grim trigger strategy. The democratic peace literature observes that while democracies are as war prone as other states, they do not fight each other.[23] Our results show a similar dyadic dependence. Unless both agents are accountable, the conditions for cooperation are identical to the extant unitary actor approach. Figure 2 shows how the domestic political institutions of nations 1 and 2 determine the conditions under which cooperation can occur. The ASGT strategy requires that both agents be accountable, since cooperation requires that both agents be trustworthy. We now turn to this finding.

If both agents are politically accountable, then nations can support cooperation when $\delta \geq (T - R)/(T + \Psi - R)$. Yet if either of the agents is unaccountable (or, alternatively, accountable leaders forgo the benefits of agent-specific punishment and target the nation as a whole), then cooperation requires $\delta \geq (T - R)/(T - P)$. Agent-specific punishments make cooperation possible under a wider range of conditions if $\Psi > R - P$: leaders value office holding more than the difference between a cooperative and a noncooperative outcome. While obviously the superiority of the ASGT strategy rests on this inequality holding, given the primacy of office holding for *Homo politicus*, we believe such conditions generally prevail and that a leader would

23. See Bremer 1992; Bueno de Mesquita et al. 1999 and 2000; Dixon 1994; Lake 1992; Levy 1988; Maoz and Abdolali 1989; Maoz and Russett 1993; Ray 1995; and Rousseau et al. 1996.

Nation 1 \ Nation 2	Accountable	Unaccountable
Accountable	$\delta \geq \frac{T-R}{T+\Psi-R}$	$\delta \geq \frac{T-R}{T-P}$
Unaccountable	$\delta \geq \frac{T-R}{T-P}$	$\delta \geq \frac{T-R}{T-P}$

FIGURE 2. *Conditions under which cooperation can occur as a function of domestic political institutions*

be unlikely to prefer adherence to, for example, a trade agreement instead of remaining in power.

Figure 3 illustrates the discount factor required to support cooperation under both the grim trigger strategy and the ASGT strategy as a function of the value of holding office, Ψ. As a guide to the magnitude of effects, we return to the earlier numerical values, $T = 4$, $R = 3$, $P = 2$, and $S = 1$, in constructing this figure. The unitary actor approach requires a discount factor of $\delta \geq 1/2$ to support cooperation. Yet as the benefits of office holding increase, cooperation can be supported under virtually any circumstances through the ASGT strategy. For example, if the value of office, Ψ, is 10, cooperation is possible when $\delta \geq 1/11 \approx 0.091$.

Figure 3 shows the ability of agent-specific punishments to support cooperation under a greater range of conditions than unitary actor approaches. Yet the ASGT strategy is more robust than grim trigger on another dimension as well. A common problem with grim trigger mechanisms is that they are not renegotiation proof, meaning that once the punishment starts, nations want to avoid the consequences of the punishment.[24] This undermines the credibility of the punishment and, hence, reduces trust. In contrast, the ASGT strategy is renegotiation proof. It explicitly allows principals to curtail punishments—by removing leaders—and this strengthens rather than undermines the credibility of agents.

Principals remove accountable agents who renege on agreements or in other ways break the norms of international cooperation. In the literature, this risk of removal is often referred to as an audience cost.[25] In common with this article, Ashley Leeds uses a prisoners' dilemma model of cooperation to show that as leaders face higher costs from breaking agreements, nations are better able to cooperate.[26] In parallel

24. See Farrell and Maskin 1989; and Abreu, Peace, and Stachetti 1989.

25. See Bueno de Mesquita and Lalman 1992; Eyerman and Hart 1996; Fearon 1994; Guisinger and Smith 1999; Leeds 1999; Mansfield, Milner, and Rosendorff 1998; Martin 1993; Schultz 1998; Smith 1998a,b; and Schultz 1999a,b.

26. Leeds 1999. For similar arguments with respect to alliances, see Gaubatz 1996; and Smith 1995. For similar arguments with respect to trade policy, see Mansfield, Milner, and Rosendorff 1998.

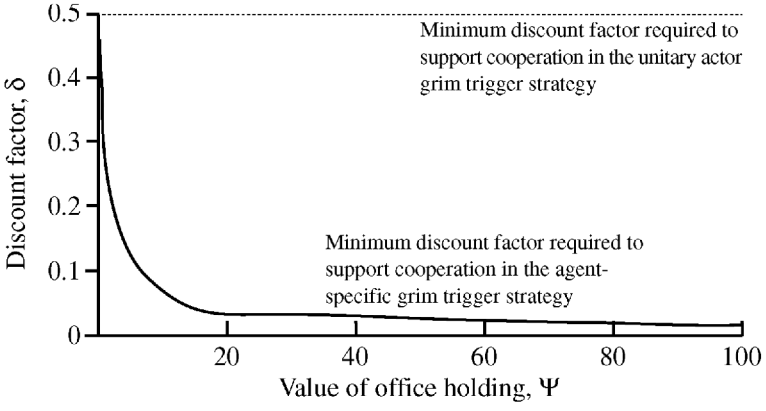

FIGURE 3. *The relationship between the value of office holding and the minimum discount factor required to support cooperation under the grim trigger and the agent-specific grim trigger strategies*

with other studies, she asserts that democratic leaders, on the basis of their heightened domestic vulnerability, allow democratic states to cooperate to a greater extent than other pairs of regime types. Her large-sample empirical tests support this hypothesis.

While leaders in democracies are clearly more sensitive to domestic pressures than leaders in autocracies, the majority of extant audience-cost explanations fail to show why citizens should want to punish leaders. After all, in the case of the prisoners' dilemma, if a leader wants to defect, say, against a grim trigger, then the citizens also want to defect. (This is strictly true only if leaders and citizens have identical preferences over international outcomes; however, it is assumed to be the case in nearly all existing work.) So, the relevant question is, should citizens punish leaders for breaking agreements that they themselves wanted broken? Agent-specific punishments provide this link. As discussed earlier, although breaking an agreement might be popular with the principals, they will still remove a leader who does so in order to restore cooperation. Hence, in contrast to much of the extant literature, our theory of audience costs is endogenous, not only showing how audience costs affect behavior, but also explaining the origin of these costs.

Agent-specific punishments suggest that leaders in democracies are more trustworthy than less accountable leaders and hence can cooperate at higher levels. While this prediction is consistent with empirical findings, this does not distinguish it from extant explanations for cooperation.[27] However, the theory differentiates itself from unitary actor models in several ways. For example, one implication of the ASGT theory is that accountable leaders who defect are likely to be removed. Unfortunately, directly testing this prediction is problematic.[28] Such instances, being off the

27. Such as Leeds 1999.

28. Through the use of a simple game-theoretic model and Monte Carlo simulation, Schultz shows the difficulty in estimating the magnitude of audience costs when leaders seek to avoid them. Schultz 1999b.

equilibrium path, are likely to be rare. Simply put, when the magnitude of audience costs is large, leaders avoid them.[29] Yet patterns of cooperation distinguish agent-specific punishments from existing theories of cooperation.

Once disrupted, the restoration of cooperation depends on domestic political arrangements. In contrast to their more accountable counterparts, autocrats do not jeopardize their domestic political future when they abrogate international agreements or violate international norms. As such they are more likely to undertake such actions; and once they have done so, the agent-specific punishments model predicts that relations between nations will remain acrimonious as long as these leaders remain in power. Yet when such leaders are replaced, it signals the restoration of cooperation and the normalization of relations. In contrast to extant theories, the agent-specific punishments theory predicts that changes in leadership reduce the dependence of future relations on past behavior.

Conclusions

Trust is the key to international cooperation. Until nations are confident that they will not be exploited, they will not collaborate for mutual gains. In extant approaches the threat of future punishment keeps nations honest. If the value of future collaborations exceeds the rewards from short-term exploitation, cooperation is possible. However, such unitary actor approaches fail to account for the growing evidence that the extent of cooperation depends on domestic political institutions.

We reconceptualize the problem of international cooperation as a game between leaders fighting, not just for international gains but also for their domestic political survival. The key intuition is that interactions between states differ when foreign policy is targeted against a specific leader rather than against the nation as a whole. When leaders are the targets of agent-specific punishments, they know their citizens have an increased incentive to oust them. This incentive exists because under the ASGT strategy punishment for defection continues only as long as the incumbent agent remains in power. Accountable agents, wishing to stay in office, do not exploit others even when from a unitary actor perspective their nation could not commit to cooperation. Hence, domestic political institutions influence the extent to which international cooperation is possible. They also influence the repercussions of a breakdown in cooperation. Leaders in democracies are more accountable than their counterparts in autocracies, and hence their survival in office is highly contingent on maintaining good relations. Democrats are more likely than autocrats to lose their jobs for exploiting another nation. Yet precisely because audience costs are higher for democrats are they likely to avoid such contingencies in the first place.

29. The fate of John Major's Conservative government is perhaps one instance. On 16 September 1992 Major withdrew Britain from the European Exchange Rate Mechanism and adopted a generally recalcitrant position on European integration. Although the economy was extremely buoyant and the Conservative's stance on Europe was generally more popular than that of the opposition Labour party, support for the Conservatives plunged approximately 15 percent and from there remained consistently behind the Labour party. See Butler and Kavanagh 1997; Waller 1995; and Worcester 1997.

While to a large extent the existing literature shows that cooperation is possible, our model is more concerned with the contingent circumstances and the domestic consequences of the success or failure of cooperation. The political accountability of agents influences the pattern of behavior we expect to see. Where leaders are accountable, cooperation thrives and instances of punishment are uncommon and short in duration. Where replacing agents is difficult, cooperation is more fragile. Fewer conditions support cooperation, and once stalled, it is difficult to restart. This can lead to prolonged hostilities and periods of bitter relations between states. Yet, if only from an actuarial perspective, unaccountable leaders never survive indefinitely, and, as the case of King Hussein's death reveals, this offers the prospect of rehabilitating relations. Leader-specific punishments never do worse, in terms of promoting cooperation, than unitary actor equivalents, and they offer the prospects of doing better.

References

Abreu, D., D. Peace, and E. Stachetti. 1989. Renegotiation and Symmetry in Repeated Games. Mimeo, Department of Economics, Harvard University, Cambridge, Mass.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, Robert, and Robert O. Keohane. 1986. Achieving Cooperation Under Anarchy: Strategies and Institutions. In *Cooperation Under Anarchy*, edited by Kenneth A. Oye, 226–54. Princeton, N.J.: Princeton University Press.

Bendor, Jonathan. 1993. Uncertainty and the Evolution of Cooperation. *Journal of Conflict Resolution* 37 (4):709–34.

Bremer, Stuart. 1992. Dangerous Dyads: Conditions Affecting the Likelihood of Interstate War, 1816–1965. *Journal of Conflict Resolution* 36 (2):309–41.

Bueno de Mesquita, Bruce, and David Lalman. 1992. *War and Reason: Domestic and International Imperatives*. New Haven, Conn.: Yale University Press.

Bueno de Mesquita, Bruce, and Randolph M. Siverson. 1995. War and the Survival of Political Leaders: A Comparative Study of Regime Types and Political Accountability. *American Political Science Review* 93 (4):841–55.

Bueno de Mesquita, Bruce, James D. Morrow, Randolph M. Siverson, and Alastair Smith. 1999. An Institutional Explanation of the Democratic Peace. *American Political Science Review* 89 (December): 791–808.

———. 2000. Political Survival and International Conflict. In *War in the Changing World*, edited by Zeev Maoz and Azar Gat. Ann Arbor: University of Michigan Press.

Butler, David, and Dennis Kavanagh. 1997. *The British General Election of 1997*. New York: St. Martin's Press.

Dixon, William J. 1994. Democracy and the Peaceful Settlement of International Conflict. *American Political Science Review* 88 (1):14–32.

Eyerman, Joe, and Robert A. Hart, Jr. 1996. An Empirical Test of the Audience Cost Proposition: Democracy Speaks Louder than Words. *Journal of Conflict Resolution* 40 (4):597–616.

Farrell, Joseph, and Eric Maskin. 1989. Renegotiation in Repeated Games. *Games and Economic Behavior* 1 (4):327–60.

Fearon, James D. 1994. Domestic Political Audiences and the Escalation of International Disputes. *American Political Science Review* 88 (3):577–92.

———. 1998. Bargaining, Enforcement, and International Cooperation. *International Organization* 52 (2):269–305.

Fudenberg, Drew, and Eric Maskin. 1986. The Folk Theorem in Repeated Games with Discounting and with Incomplete Information. *Econometrica* 54 (3):533–54.

Gaubatz, Kurt Taylor. 1996. Democratic States and Commitment in International Relations. *International Organization* 50 (1):109–39.
Gourevitch, Peter Alexis. 1996. Squaring the Circle: The Domestic Sources of International Cooperation. *International Organization* 50 (2):349–73.
Gowa, Joanne. 1994. *Allies, Adversaries, and International Trade*. Princeton, N.J.: Princeton University Press.
Guisinger, Alexandra, and Alastair Smith. 1999. Honest Threat: The Interaction of Reputation and Political Institutions in International Crises. Unpublished paper, Department of Political Science, Yale University.
Lake, David A. 1992. Powerful Pacifists: Democratic States and War. *American Political Science Review* 86 (1):24–37.
Laver, Michael, and Norman Schofield. 1990. *Multiparty Government: The Politics of Coalition in Europe*. Oxford: Oxford University Press.
Leeds, Brett Ashley. 1999. Domestic Political Institutions, Credible Commitments, and International Cooperation. *American Journal of Political Science* 43 (4):979–1002.
Levy, Jack S. 1988. Domestic Politics and War. *Journal of Interdisciplinary History* 18 (4):653–73.
Lijphart, Arend. 1990. The Political Consequences of Electoral Laws, 1945–85. *American Political Science Review* 84 (2):481–96.
Mansfield, Edward D. 1998. The Proliferation of Preferential Trading Arrangements. *Journal of Conflict Resolution* 42 (5):523–43.
Mansfield, Edward D., Helen Milner, and B. Peter Rosendorff. 1998. Why Democracies Cooperate More: Electoral Control and International Trade Agreements. Paper delivered at the 94th Annual Meeting of the American Political Science Association, Boston.
Maoz, Zeev, and Nasrin Abdolali. 1989. Regime Types and International Conflict, 1816–1976. *Journal of Conflict Resolution* 33 (1):3–36.
Maoz, Zeev, and Bruce Russett. 1993. Normative and Structural Causes of Democratic Peace, 1946–1986. *American Political Science Review* 87 (3):624–38.
Martin, Lisa L. 1993. Credibility, Costs, and Institutions: Cooperation on Economic Sanctions. *World Politics* 45 (3):406–32.
McGillivray, Fiona. 1998. How Voters Shape the Institutional Framework of International Negotiations. In *Strategic Politicians, Institutions, and Foreign Policy*, edited by Randolph M. Siverson, 79–96. Ann Arbor: University of Michigan Press.
Milner, Helen. 1992. International Theories of Cooperation Among Nations: Strengths and Weaknesses. *World Politics* 44 (3):466–96.
———. 1997. *Interests, Institutions, and Information: Domestic Politics and International Relations*. Princeton, N.J.: Princeton University Press.
Milner, Helen V., and B. Peter Rosendorff. 1997. Democratic Politics and International Trade Negotiations. *Journal of Conflict Resolution* 41 (1):117–46.
Oneal, John R., and Bruce Russett. 1997. The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950–1985. *International Studies Quarterly* 41 (June):267–93.
Pahre, Robert. 1994. Multilateral Cooperation in an Iterated Prisoner's Dilemma. *Journal of Conflict Resolution* 38 (2):326–52
Pape, Robert A. 1997. Why Economic Sanctions Do Not Work. *International Security* 22 (2):90–136.
Powell, G. Bingham, Jr., and Guy D. Whitten. 1993. A Cross-National Analysis of Economic Voting: Taking Account of the Political Context. *American Journal of Political Science* 37 (2):391–414.
Ray, James Lee. 1995. *Democracy and International Conflict: An Evolution of the Democratic Peace Proposition*. Columbia: University of South Carolina Press
Remmer, Karen L. 1998. Does Democracy Promote Interstate Cooperation? Lessons from the Mercosur Region. *International Studies Quarterly* 42 (1):25–52.
Rheinhardt, Eric. 1996. Posturing Parliaments: Ratification, Uncertainty, and International Bargaining. Ph.D. diss., Columbia University.
Roubini, Nouriel, and Jeffrey D. Sachs. 1989. Political and Economic Determinants of Budget Deficits in the Industrial Democracies. *European Economic Review* 33 (May):903–33.

Rousseau, David L., Christopher Gelpi, Daniel Reiter, and Paul K. Huth. 1996. Assessing the Dyadic Nature of the Democratic Peace, 1918–88. *American Political Science Review* 90 (3):512–33.

Russett, Bruce M. 1993. *Grasping the Democratic Peace: Principles for a Post–Cold War World*. Princeton, N.J.: Princeton University Press.

Russett, Bruce M., John R. Oneal, and David R. Davis. 1998. The Third Leg of the Kantian Tripod for Peace. *International Organization* 52 (3):441–67.

Schultz, Kenneth A. 1998. Domestic Opposition and Signaling in International Crises. *American Political Science Review* 92 (4):829–44.

———. 1999a. Do Democratic Institutions Constrain or Inform? Contrasting Two Institutional Perspectives on Democracy and War. *International Organization* 53 (2):233–66.

———. 1999b. Looking for Audience Costs: A Research Note. Unpublished manuscript, Department of Political Science, Princeton University.

Schultz, Kenneth A., and Barry R. Weingast. 1998. Limited Governments, Powerful States. In *Strategic Politicians, Institutions, and Foreign Policy*, edited by Randolph M. Siverson, 15–49. Ann Arbor: University of Michigan Press.

Signorino, Curtis S. 1996. Simulating International Cooperation Under Uncertainty: The Effects of Symmetric and Asymmetric Noise. *Journal of Conflict Resolution* 40 (1):152–205.

Smith, Alastair. 1995. Alliance Formation and War. *International Studies Quarterly* 39 (4):405–25.

———. 1998a. International Crises and Domestic Politics. *American Political Science Review* 92 (3): 623–38.

———. 1998b. The Effect of Foreign Policy Statements on Foreign Nations and Domestic Electorates. In *Strategic Politicians, Institutions, and Foreign Policy*, edited by Randolph M. Siverson, 221–54. Ann Arbor: University of Michigan Press.

———. 1999. Personalizing Crises. Unpublished manuscript, Department of Political Science, Yale University.

Verdier, Daniel. 1998. Democratic Convergence and Free Trade. *International Studies Quarterly* 42 (1): 1–24.

Waller, Robert. 1995. Taxing Polls. *New Statesman and Society* 8 (354):viii–ix.

Worcester, Robert. 1997. Follow the Polls, Go for EMU. *New Statesman* 126 (4359):21.

Wu, Jianzhong, and Robert Axelrod. 1995. How to Cope with Noise in the Iterated Prisoner's Dilemma. *Journal of Conflict Resolution* 39 (1):183–89.