INDUCING DETERRENCE THROUGH MORAL HAZARD IN ALLIANCE CONTRACTS

BRETT BENSON
ADAM MEIROWITZ
KRISTOPHER W. RAMSAY

Abstract. Do military alliances foster aggressive behavior in allies to the point of undermining the security goal of the alliance? Like others, we find that alliance commitments may cause moral hazard because allies do not fully internalize the costs of actions that can lead to war. But unlike others, we show that the effect of moral hazard can improve security. Moral hazard can be the driving force behind generating deterrence and avoiding costly conflict. Aggressors may refrain from initiating crises if their target enjoys additional resources from its ally and so is more willing to fight back. So rather than incurring costs moral hazard may be the very key to deterring potential aggressors and minimize the risk of conflict. This behavior allows alliance partners to capture a "deterrence surplus," the gains from avoiding conflict.

1. Introduction

Do military alliances cause allies to act so aggressively that their behavior undermines the security goal of the alliance? If so, then leaders may be cautious about joining an alliance, doing so only if they can select safe alliance partners or design the terms of the treaty in a way that captures the deterrence benefits while managing the dangers of over-aggression. In this paper, we develop a theory of security alliances that advances an alternative argument. We show that while alliance commitments may cause alliance partners to behave aggressively, under some conditions the added aggressiveness actually enhances deterrence. From this perspective, the effectiveness of an alliance is related both to the structure and the content of the treaty. Countries may shop for alliance partners and design treaty terms that enable them to capitalize on the deterrence benefits of a prospective ally's increased tendency towards aggression.

This view stands in contrast to the standard explanation that the content of an alliance agreement is often designed to balance the benefit of deterring an enemy against the risk of emboldening an ally (Snyder 1997; Fearon 1997; Yuen 2009). Our approach builds on existing models of alliance formation (Smith 1995; Morrow 1994) but adds intra-alliance contracting over the benefits from successful deterrence. Specifically, avoiding a conflict that they otherwise might expect creates a Coasian surplus equal to the foregone cost of war, which may be divided among alliance partners.

To facilitate an understanding of the intuition, we suggest that military alliances share many similarities with standard insurance contracts. Much as an auto insurance policy stipulates how much a policy holder will receive if she is in an accident, an alliance agreement likewise often describes how much aid an ally will provide to the attacked party if there is a war. For example, the 1656 Treaty of Defensive Alliance between Brandenburg and France enumerates precisely the amount and form of aid each ally would provide to the other if it was attacked: Brandenburg pledged 2400 men and 600 horses to France while France promised 5000 men, 1200 horses, and artillery to Brandenburg.

Furthermore, in an insurance contract, the size of the insurance premium the insured pays usually depends on the amount of risk being indemnified by the insurance provider: the more risk for the insurer, the higher the premium. Our explanation of the content of alliance commitments likewise ties the level of support to the amount of security risk alliance partners face. That is, leaders of threatened countries may look to team up with each other, and the amount of support they promise one another may depend on the amount of threat each faces. However, insurance against risk carries with it the potential problem of moral hazard, which occurs when the guarantee of indemnity distorts the insureds behavior because the insurance policy insulates her from the risks of her actions (Pauly 1968, 1974; Shavell 1979). Just as insured motorists may exercise less caution for their property, states insured by alliance treaties have an incentive to behave more aggressively in negotiating with other states.

Generally, scholars of alliances take the position that moral hazard creates potentially harmful effects. Most notably, Snyder (1984), Snyder (1997), and Christensen and Snyder (1990) claim that alliances “embolden” state leaders to “entrap” unwilling allies in wars that they would prefer to avoid. Yuen (2009) shows that moral hazard increases allied states’ level of aggression in crisis bargaining, and this added aggression may heighten the risk of war or affect the bargaining settlements. As a result of the potential harmful effects of moral hazard, scholars argue that leaders may either avoid alliances, screen alliance partners based on their likelihood of behaving recklessly, or attempt to design treaties carefully so as to balance their dueling goals of deterring external threats while restraining alliance partners (Snyder 1984; Jervis 1994; Snyder 1997; Zagare and Kilgour 2003; Yuen 2009).

A more subtle side effect of moral hazard, however, attracts states to alliances because the tendency of an ally to behave aggressively actually enhances deterrence. The possibility that an allied state will negotiate aggressively may cause third-party adversaries to refrain from initiating a crisis. Likewise, a defensive alliance might make an alliance partner more willing to retaliate if challenged because it benefits from its ally's support in war; this may cause a prospective adversary to be reluctant to initiate a challenge targeting the alliance partner (Smith 1995).

In cases where moral hazard advances the deterrence objective of an alliance, there is little cost to entrapment, because the third-party adversary calibrates its hostility toward the allies based on its expectation about its likelihood of winning a conflict if the target of their challenge does not capitulate. The combination of added resources from an ally and the increased willingness of the target to fight back encourages the third party to refrain from initiating violence. When encouraging an alliance partner to fight back if it is attacked enhances deterrence, then the goal of the contract is to induce a maximal amount of moral hazard so as to deter potential aggressors to such an extant that the risk of conflict is negligible. In this case allies are not called upon to expend costly resources in support of their partners, as no conflict occurs. Thus, *a priori*, it seems equally likely that moral hazard will deter would-be challenges or increase the likelihood of conflict. Therefore, an important challenge for a theory of alliances is to identify the conditions under which moral hazard serves the deterrence purpose of the alliance rather than causing harmful effects that undermine the alliance's objective.

The amount of the deterrence surplus and how it is captured depends on the structure of the alliance and international environment. Alliances may have different structures depending on which alliance partners can be secured from threat by joining. That is, they may include alliance partners that all gain security through the alliance, partners that all remain at risk even after the formation of the alliance, or a combination of partners that include some that become secure and some that do not. If all alliance partners gain security, then all consume the benefits of deterrence without paying any of the costs associated with having to defend an ally. On the other hand, an alliance that includes a partnership between states who may become secure and other states for whom there are no deterrence benefits to partnering can be expensive for the partners who gains security. Nevertheless, if forming an alliance insures the securable countries against the risk of being attacked, they will still pay the insecure partners to join the alliance.[1]

In conceptualizing alliances as insurance contracts, our approach differs from other theories of alliance formation. The literature on this topic is diverse, from the theory of alliances as producers of the public good of security (Olson and Zeckhauser 1966; Sandler 1993; Sandler and Hartley 2001) to the idea that states take advantage of the free rider problem inherent in alliance maintenance to reduce incentives

[1] There is some empirical support for this line of thinking. Poast (2012*b*) finds that vulnerable states often pay alliance partners to "to seal the deal". Moreover, countries that create benefits by including trade provisions in treaties are less likely to be attacked by third parties (Poast (2012*a*)).

among prospective allies to compete over a disputed good (Garfinkel 2004). These theories pay scant attention to the effects of moral hazard on states' behavior, however. In contrast, our approach considers how and when moral hazard impacts the effectiveness of an alliance and dictates its content and structure.

Our approach also differs from previous theories in that we make fewer assumptions about the relationship between the alliance partners. For instance, scholars have shown that one motivation for alliance formation is that alliances can be used as signals to establish commitment to extended deterrence (Morrow 1994; Huth 1991; Smith 1995; Fearon 1997; Smith 1998). Many of these studies also assume defenders intrinsically value the security of their ally. While signaling and commitment problems are important to alliance theory, we abstract away from signaling and commitment problems, because these effects are well-established in the literature. This modification enables us to focus on the effect of moral hazard on deterrence and intra-alliance bargaining when commitments are credible. Additionally, we do not require that prospective allies value one another's security to motivate the formation of their alliance.

Another theory of alliance formation argues that countries may form a partnership to exchange the benefits of security and political influence (Morrow 1991). From this perspective, prospective allies need not value one another's security to have an incentive to form an alliance; the existence of two goods – security and influence – and asymmetric preferences over these goods can create gains from trade through alliance. In our model, countries' concern for their own security is sufficient to motivate an alliance. Additionally, even if joining an alliance does not secure it from threat, a country may still join an alliance if its presence achieves deterrence for an ally. Alliance partnerships may provide deterrence for some partners but not others, because there may be disparities in their capabilities relative to the third-party challenger, differences in the risks of external threat they each face, and dissimilarities in the size of the stakes at issue between each and the challenger. Countries who do not benefit enough from the support of prospective allies to gain security may nevertheless join the alliance because of gains generated by the deterrence surplus generated from protecting the securable allies. States that benefit from this deterrence then share the rents by pledging to support the ally for whom deterrence is incomplete. This is a costly action but worth the price of the deterrence that comes with the alliance.

In many ways, our approach builds on several traditional studies of mutual security alliances (Walt 1990; Christensen and Snyder 1990; Snyder 1984), which argue that alliances are often agreements between self-interested parties who have nothing in common except a demand for security. Some of these studies also observe that alliances may embolden partners to behave aggressively and, because alliances are often formed between states that have very little in common, the implications of emboldenment may be undesirable for at least some alliance partners. We build on these ideas, showing that states may indeed form alliances even if they do not care about one another, and that these alliances often create moral hazard effects. Consequently, countries select alliance partners and bargain over the terms of the treaty precisely to induce optimal levels of moral hazard.

To highlight these implications we focus our analysis on a model in which many of the standard rationales for alliances are absent. We consider a situation where commitments are credible, but the terms of the commitment are chosen strategically.

Moreover, we do not focus on the relative risk aversion of actors and various motivations for alliance formation beyond preservation of one's own security. We provide a model of alliance formation where the presence of security threats and the cost of fighting create incentives to form agreements of "mutual-help" in the case of war. We close by considering an extension in which alliance contracts involve transfers that actually change war-fighting and the probability of victory. Although the logic behind the analysis of this extension is consistent with the baseline model, it is easier in this case to generate alliances and their characteristics depend on conditions regarding the technology relating military transfers.

2. Literature

Our approach builds on three strains of research. First is the literature on moral hazard in crises. Second is the literature on diversification of risk in alliance portfolios and the economics literature on insurance. Third is the literature on alliance commitment.

The problem of moral hazard in international relations has a long tradition. Many scholars have observed that committing aid to another state may cause that state to behave more aggressively than it otherwise would (Snyder 1984; Jervis 1994; Snyder 1997; Fearon 1997; Crawford 2001, 2003). This effect occurs not only in alliances and extended deterrence agreements but humanitarian intervention as well. Crawford (2005) for instance, argues that humanitarian intervention may incite unintended rebellions, and Kuperman (2008) provides evidence from Bosnia and Kosovo to show that humanitarian intervention can cause citizen rebellions that trigger retaliation by the state.

While many scholars have highlighted the dangers of moral hazards in commitments and intervention, scholarship has pointed out that intervention can be calibrated to balance the costs of moral hazard with the benefit of increased security (Wagner 2005). In this vein, Snyder (1997) points out that flexibility and ambiguity in alliances often reflect the intention of one or more countries to restrain an alliance partner because of fears of entrapment. Zagare and Kilgour (2003) create a formal model to capture the deterrence-versus-restraint phenomenon in alliances, finding a pooling equilibrium in mixed strategies in which an ally creates some uncertainty about whether it will intervene on the behalf of its alliance partner in a conflict. The authors interpret this equilibrium behavior as a kind of ambiguous alliance designed to restrain overly-aggressive behavior, although they do not model the alliance formation stage. And in her model of third-party intervention with moral hazard, Yuen (2009) shows that alliances not only can strike a balance between deterrence and an ally's over-aggression but, when the ally's costs for fighting are sufficiently high, the alliance can actually induce the ally to make small concessions to the challenger to avoid conflict.

Missing from the literature is a theory that explicitly formalizes the negotiating environment in which prospective allies, anticipating that moral hazard may both enhance deterrence and provoke aggression, bargain over the terms of their alliance. The theory offered by Snyder (1997) comes closest to what we have in mind. In his theory, prospective allies bargain over the terms of aid and the distribution of benefits of the alliance. A country's bargaining power grows as the ratio of its valuation of the alliance to its valuation of its alternative options decreases. Bargaining power is also affected by a country's relative valuation of the alliance compared to the valuations by its prospective allies. Relative valuation is determined by three factors: (1) the level of threat each ally faces from a prospective adversary; (2) how much each ally expects to gain from the other's aid; and (3) the cost each ally pays for sacrificing some autonomy by joining the alliance. Moreover, a country's bargaining power depends on its value of remaining unaligned or allying with another country. A country's bargaining power matters because different countries have different fears and, therefore, desire different alliance structures. If a country's predominant concern is that its alliance partner will entrap it in a war, then it will use its bargaining power to insist on a flexible or ambiguous alliance. On the other hand, if it is primarily worried about its ally abandoning it if conflict occurs, then it will negotiate for a firm and unambiguous alliance.

In our approach, alliance partners negotiate with one another directly about the content of the alliance, and some of Snyder's ideas related to bargaining power are also relevant in our model. However, rather than negotiate over the relative flexibility or transparency of an alliance agreement, the alliance partners in our model bargain specifically over how much assistance they are willing to transfer to one another, as well as the division of the deterrence surplus created by forming a successful alliance. This is an important piece of the puzzle because prospective allies care deeply about the terms of the alliance: the level of their obligation to their allies and of the aid they will receive in return are directly relevant to the likelihood that deterrence will succeed. Moreover, if concerns about moral hazard can be satisfied by striking a deal on just the right level of assistance with just the right alliance partner, then mechanisms designed to restrain alliance partners are not always required even when entrapment fears prevail.

Another advantage of our theory is that it avoids the criticisms leveled by Rauchhaus (Rauchhaus 2005, 2009) and others against scholars who have misunderstood or misused the concept of moral hazard. Rauchhaus's main criticism is directed toward studies that invoke moral hazard as an explanation for the outbreak of conflict without justifying why the adversary would not back down as a result of a more aggressive rebel group or ally. We agree with this point, and our model demonstrates that this is precisely the mechanism by which deterrence is achieved. In response to Wagner (2005)'s concern that many studies blame over-insurance for causing of conflict, we show that, under certain conditions, over-insurance enhances deterrence rather than causing conflict.

The second line of research on alliances that informs our theory emphasizes the ability of alliance portfolios to diversify a country's security risks. Our theoretical analysis starts from the observation that many alliances share similarities with insurance contracts, since an alliance agreement describes how much aid the ally will provide to the attacked party if there is a war. Interstate swaps of these insurance contracts bear some resemblance to exchanges of military securities between countries. This view of alliances is related to the portfolio analysis in Conybeare (1992), in which alliances are viewed as investment portfolios formed by countries for the purpose of diversifying their risks of war. The portfolio model, which does not specify any strategic behavior, predicts that the security risk of an alliance portfolio is decreasing in the number of allies. Our approach goes beyond the portfolio model by allowing allies to bargain over how much security to swap given some exogenous risk of crisis and by adding a conflict subgame in which those securities impact crisis payoffs. We also introduce an additional dimension of risk by incorporating moral hazard; security assurances encourage allies to fight in the crisis subgame because they increase the payoff to war. Therefore, decisions to promise securities to a prospective ally depend on the anticipation of amount of risk created by that ally's behavior.

Finally, we also build on a large body of literature that focuses on commitment. Much of the formal analysis of alliances and deterrence emphasizes the role of a country's commitment to its ally, and the factors that lead to credible (and incredible) commitments. These studies have already established the mechanisms through which allies establish their credibility with one another. Additionally, we know empirically that alliances are usually reliable (Leeds, Long and Mitchell 2000). Since the focus of our theory is on the structure and content of alliance agreements, we need not reproduce proofs of commitment here. Instead, we examine who allies with whom, what promises they make, and how those promises affect conflict given that all relevant players recognize the commitment is credibly.

The advantage of this assumption is two-fold. First, it spares the reader a significant amount of unnecessary analysis given that the mechanisms for credibility are well-established. Second, our approach enables us to address the question of what type of commitment we should expect if alliances are assumed to be credible and there is no risk of abandonment. Understanding the wide variety of details in the structure of commitments made by countries is valuable. Certainly, that said, explanations of the content of alliance members' commitments is understudied and less well-understood. In many ways, the case of credible commitment is the hard case for our objective, because fears about moral hazard prevail when there is no question about the reliability of the alliance. Under these circumstances, over-commitment can embolden allies to try to entrap their partners in wars the partners do not want to fight. Snyder's response to this problem is to inject some uncertainty into the alliance to add a risk of abandonment. In other words, he sees a solution to the problem of moral hazard when commitments are not fully credible. In this paper we seek to determine whether moral hazard might have benefits even when there is no possibility that a country could abandon its ally.

3. Model

Consider a world with three countries, a challenger and two potential targets. With probability π_i, target $i \in \{1, 2\}$ has a crisis with the challenger. We assume that $\pi_i + \pi_i = 1$. Once a crisis starts, the challenger decides whether or not to escalate by threatening target i. If the challenger chooses the status quo, instead of making a threat, the crisis ends peacefully and there is no change in the stakes controlled by either side. We therefore normalize the payoff for the status quo to 0 for the challenger and 1 for the target.

If the challenger makes a threat, on the other hand, the target country can choose to resist the threat and fight to keep the status quo, or capitulate and give in to the challenger's threat. If the target fights the dispute is settled by war. The challenger wins a war against target i with probability p_i and pays a cost k_i. For simplicity we assume that the challenger's costs of fighting are known. We assume that the target countries have private information regarding their costs of war in the crisis. Each target has a cost of war $c_i \in [0, \bar{c}]$. We let $F(c)$ denote the prior on this cost and assume it has a continuous density. If the challenger issues a threat, the target can

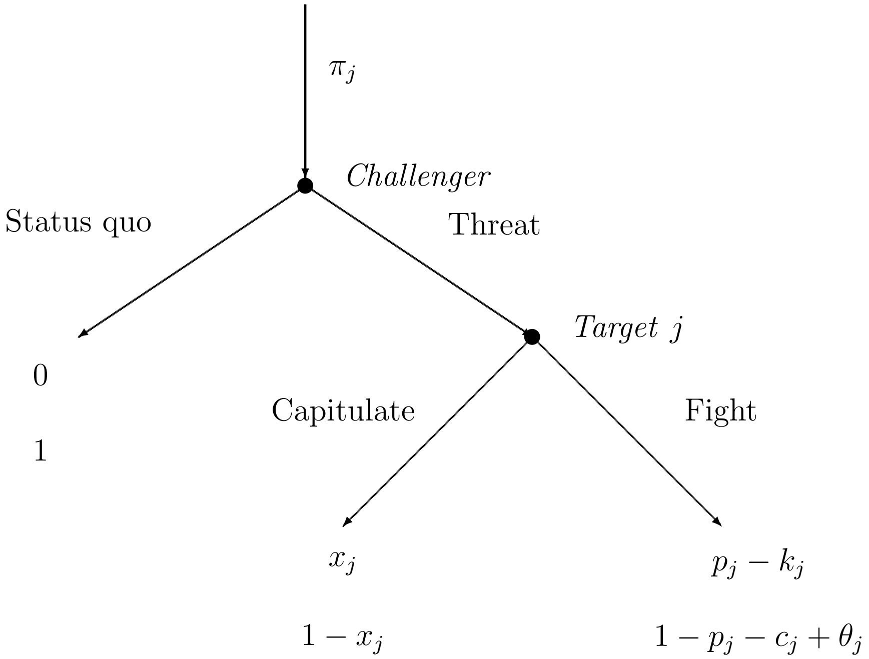

FIGURE 1. Crisis game

avoid war by capitulating, but then the target must forfeit the "stakes" of the crisis, x_j, keeping the fraction $1 - x_j$ for itself. The game is depicted in Figure 1.

We imbed this crisis game into a larger *ex ante* bargaining game in which the two potential targets can make an agreement regarding the amount of aid they will provide to each other if one is engaged in war. Discussion of the bargaining game follows our discussion of the alliances.

In general, the agreement (or alliance) will constitute a transfer from one target (say i) to the other (say j) of an amount $\theta_j \geq 0$ if there is a war. We can think of the *ex ante* alliance agreement as a form of decentralized insurance. Much as family members might help each other financially if one loses a job or experience an accident or illness, these countries are agreeing to help each other through the transfer of resources from one country to the other in the case of war. An alliance agreement then is a pair of transfer, $\boldsymbol{\theta} = (\theta_1, \theta_2) \in R^2_+$. Security alliances are made *ex ante* in the sense that the players do not know their costs of war at the time of agreement, though they have beliefs about the distribution of these costs.

We are interested in settings in which war is possible in the absence of alliances. This requires that with positive probability c_j is low enough so that j prefers to fight (getting payoff $1 - p_j - c_j$) to capitulating (getting $1 - x_j$). This necessitates that $p_j < x_j$ for both targets. We maintain this assumption throughout the paper.

Given an agreement, $\boldsymbol{\theta} \in \Theta^2$, we let $u_i(\boldsymbol{\theta})$ denote the expected payoff to country i from this agreement. Naturally, if the parties do not consent to an agreement, then their payoffs are given by $u_i(0, 0)$. Specifying players' payoffs in both the treaty and null treaty environments facilitates comparisons that enable us to determine whether an agreement is preferred to other agreements and when there is no treaty. In some situations it will be particularly useful to distinguish between alliance agreements that are Pareto Efficient and those that are not.

Definition 1. *An agreement $\boldsymbol{\theta}$ Pareto dominates agreement $\boldsymbol{\theta}'$ if $u_i(\boldsymbol{\theta}) \geq u_i(\boldsymbol{\theta}')$ for $i = 1, 2$ with a strict inequality for at least one of the players. An agreement is Pareto efficient if no agreement Pareto dominates it. Finally, a treaty, $\boldsymbol{\theta}$ is Pareto dominant if for all other agreements, $\boldsymbol{\theta'}$ one of the following is true: $\boldsymbol{\theta}$ Pareto dominates $\boldsymbol{\theta'}$ or $u_i(\boldsymbol{\theta}) = u_i(\boldsymbol{\theta'})$ for $i = 1, 2$.*

The two target countries in our model reach an agreement by bargaining over the levels of support θ_i and θ_j that they will provide to each other in a war. We consider a situation where the bargaining protocol is the alternating offers procedure with a risk of break down (Rubinstein 1982; Binmore, Rubinstein and Wolinsky 1986). In period 1, country 1 makes a proposal that 2 may accept or reject. If 2 accepts, bargaining ends and the crisis subform is played. If 2 rejects the proposal, no agreement is reached and the game continues with a lottery that determines whether bargaining resumes or players proceed to the crisis subform without an agreement. With probability z the crisis subform is reached and the game ends with payoffs, $u_i(0, 0)$. With probability $1 - z$, there is no crisis in this period and the bargaining phase of the game proceeds to period $t + 1$. If a subsequent bargaining period is reached then the roles are reversed: 2 makes an offer that 1 either rejects or accepts. The process continues until either an agreement is reached or bargaining is interrupted by a crisis where one of the countries faces the challenger without an agreement.

A standard representation of this game-form involves nature first randomizing over the costs faced by states 1 and 2 if they enter into a conflict. These costs are unknown by both players and are only revealed to the relevant state if a crisis involving that state occurs. Next the two countries bargain over an alliance agreement, which ends with an agreement or the arrival of a crisis, at which point one of the nations enters into the crisis subform with the challenger. In an equilibrium to the alliance formation game, the bargaining is predicated on continuation values from the behavior in possible crisis subforms that is sequentially rational. During bargaining there is no room for meaningful signaling, as the states do not know yet their costs of conflict. Given the dynamic nature of the game and the information environment, we will apply the equilibrium concept of perfect Bayesian equilibrium and impose the additional condition known as “no signaling what you don’t know,” which in this case requires that posterior assessments of c_j are equivalent to the prior at any history (both on and off the path) in which country j has not yet learned its type. Thus, players treat c_j as drawn from the prior at all histories except for one in which j is in a crisis situation and has therefore observed its cost. We thus use the term **equilibrium** to mean PBE with this additional condition. We do not discount, and thus there is no content to assuming that a crisis occurs immediately following an agreement between allies, but occurs with possible delay if no agreement is reached. What matters is that failure to agree on a treaty leads to the possibility of a crisis without a treaty, and agreement at period t ensures that the contract is present if a crisis occurs.

4. Results

To analyze the incentives that the countries face in the bargaining phase of the game, we begin by analyzing the crisis subforms taking the alliance agreement as fixed. Given a pair of agreements $\boldsymbol{\theta} = (\theta_1, \theta_2)$, the target's decision to go to war is well defined. Specifically, target j will capitulate in equilibrium if

$$1 - x_j \geq 1 - p_j - c_j + \theta_j$$
$$c_j \geq x_j - p_j + \theta_j.$$

Perfect Bayesian equilibrium, or more precisely sequential rationality, implies that if θ_j is greater than $c_j - x_j + p_j$, then j will choose to go to war to maintain the status quo. This observation makes it clear that target countries that anticipate some chance of war have a utility that is increasing in the level of commitments they extract from their ally. For the alliance partner, however, the alliance commitments create different incentives. Because alliances make wars more attractive, the targets will fight back more often; thus, the more the alliance partner promises to its allies, the more frequently it will need to make the transfer. The effect of θ_j on a target country's behavior is analogous to *moral hazard* in insurance markets: the fact that a player is being indemnified in the case of war may make it choose to fight wars that it would otherwise avoid.

As noted by Rauchhaus (2005), alliance contracts also influence the decisions of challengers. Obviously, it could be the case that the presence or absence of an alliance agreement between two targets has no affect on the decision of the potential challenger. On the other hand, an alliance agreement may make threatening sufficiently less attractive for the challenger so that it prefers the status quo over initiating a crisis.

Definition 2. *We will call an alliance agreement* (θ_1, θ_2) deterrent for j *if it is the case that the challenger would make a threat against* j *if* $\theta_j = 0$*, but it does not make a threat against* j *given this alliance agreement. We say an agreement is* deterrent *if it is deterrent for at least one player. We say a country* j *is* large *if there exists a treaty that is deterrent for* j*.*

Importantly, both allies are (weakly) better off, if deterrence is achieved. The target in the crisis is never challenged and the ally never has to follow through on its agreement to transfer resources because war does not happen. How these various security possibilities work out and how the incentives they create for prospective alliance partners shape behavior are at the center of our theory of alliance agreements. We break up the analysis into three natural cases. First, we consider settings in which there exist treaties that are deterrent for both 1 and 2. Then we consider the case where there is no treaty that is deterrent for either player. Finally, we consider the case where there is a treaty that is deterrent for 1 but none that are deterrent for 2.

4.1. **Large targets.** We start by considering the case where the targets can form alliance agreements that change the behavior of the challenger in some circumstances. For our purpose, the term *large* means that the alliance decisions of a target can influence the threat decision of the challenger.[2] In particular, consider the case where

$$p_j < k_j \text{ and } F(x_j - p_j)(p_j - k_j) + (1 - F(x_j - p_j))x_j > 0 \tag{1}$$

hold for both targets. Under these conditions, if the challenger believes that the target will definitely choose war if faced with a threat, the challenger will choose to keep the status quo. On the other hand, the second inequality states that the challenger will find aggression worthwhile if the target does not have an alliance agreement because the odds that the target will choose to fight are sufficiently small. Thus, under the null treaty, the challenger would threaten either target. This is a situation where the challenger is potentially deterrable. More precisely under these conditions there exist treaties that are deterrent for both 1 and 2, so under our nomenclature both 1 and 2 are large.

[2] This classification of targets is similar to how one classifies countries in terms of international trade–i.e., whether their actions can affect world prices. Importantly, equation (1) implies that whether a target is large or small depends on many factors, including the probability of winning a war, the costs of the challenger to fight this target, and the stakes of the conflict. To be clear, it is not simply a major vs. minor power classification.

In this situation there exist treaties, (θ_1, θ_2), that induce target countries to fight regardless of their costs. In particular, whenever

$$\theta_j \geq \bar{c}_j - x_j + p_j$$

all types of country j will fight if challenged and thus the challenger will keep the status quo rather than fight regardless of which target country it is paired with in a crisis. This conclusion does not rely on the boundedness of the support of costs. Even if the target country's war costs are so high that it might make a concession if it is challenged, it is still possible for large states to reach a treaty agreement that will deter the challenger. That is, since $p_j - k_j < 0$, a probability of fighting that is less than 1 is still sufficient to deter the challenger and thus, without loss of generality, we can consider the case where c_j is drawn from a distribution with support R_+ and still find a θ_j for which $F(x_j - p_j + \theta_j)(p_j - k_j) + (1 - F(x_j - p_j + \theta_j))x_j < 0$. Let $\underline{\theta}_i$ be the smallest amount of support that target i needs to receive from target j to deter a threat.

In the case of two large targets the equilibrium set has a stark structure.

Lemma 1. *When condition 1 is satisfied for both targets (i.e. both targets are large) every equilibrium is deterrent for 1 and 2. In other words there is no perfect Bayesian equilibrium which places positive probability on ending with a treaty* (θ_1^*, θ_2^*) *with* $0 < \theta_j^* < \underline{\theta}_j$ *for either (or both) target,* j.

If the transfers agreed in the treaty are sufficiently large, on the path of play, the challenger never advances a threat and the agreement is never activated. In this scenario, the targets get their maximal possible payoff, because they do not receive a challenge and do not need to transfer resources to each other or to the challenger. Thus, the agreement generates a deterrence surplus, because the target countries save the expected expense of being challenged and either conceding the stakes or fighting a costly war.

To complete our analysis, we show that agreements that are deterrent for both targets are reached without delay.

Lemma 2. *Suppose that for both targets condition (1) is satisfied (both targets are large). Then in every perfect Bayesian equilibrium it is the case that alliance agreements are reached without delay.*

Using these two lemmas we can establish the following result for alliance agreements between large states.

Proposition 1. *Suppose that condition* (1) *is satisfied (both targets are large). Then there is an equilibrium and in every perfect Bayesian equilibrium alliance agreements are reached without delay and they completely deter challenges.*

Proof. The second part of the result follows from the lemmas. By Lemma 1, all equilibrium agreements completely deter and by Lemma 2 they are achieved without delay. To establish existence, consider the strategy profile in which crisis bargaining follows the form described above, and at any history in which j is the proposer it offers $(\underline{\theta}_1, \underline{\theta}_2)$ and at any history in which j is given the chance to accept or reject an offer, she accepts any offer that is weakly preferred to the continuation payoff from the null alliance in the next period and agreement to the vector $(\underline{\theta}_1, \underline{\theta}_2)$ if the game does not end in the next period. Given the proposal strategies, this acceptance rule is clearly sequentially rational, and, given the acceptance strategies, the proposal strategy implies that at every period in which i is the proposer, her proposal is accepted and a Pareto Efficient treaty is reached. Since this treaty yields the maximal payoff to each player in the crisis bargaining game, this proposal strategy is sequentially rational. □

4.2. **Small targets.** Next we consider the case where the challenger cannot be deterred. This is the case where the challenger finds that fighting is more advantageous than the status quo. That is, we assume that $p_j > k_j$. In this case, we know that for each i, the targets' expected utility as a function of θ_1 and θ_2 is given by

$$\begin{aligned} u_i(\theta_1, \theta_2) = {} & \pi_i(F(x_i - p_i + \theta_i)(1 - p_i - \hat{c}_i(\theta_i) + \theta_i) \\ & + (1 - F(x_i - p_i + \theta_i))(1 - x_i)) + (1 - \pi_i)(1 - (F(x_j - p_j + \theta_j)\theta_j)) \end{aligned}$$

The key difference between this case and that of the two large targets is that the cone of Pareto dominant treaties that could costlessly deter attacks is empty. In the case of two large targets, because deterrence was achievable for both targets, neither had to pay the cost of transferring resources to support their alliance partner in a war. This unique benefit does not occur with two small states; an alliance treaty would require a payment of transfer in any crisis. An essential fact, however, is that the null treaty $(0, 0)$ is Pareto efficient. Every treaty that makes one country better off makes the other country worse off. This conclusion stems from the fact that any treaty other than $(0, 0)$ involves a distortion such that moral hazard induces some types of at least one of the country to resist a threat and fight a war that it should not fight. The fact that the treaty compensates the fighting country represents a redistribution of this inefficiency between the treaty parties and proves that the $(0,0)$ agreement is Pareto efficient. Most importantly for our analysis, the ally that is committed to making the transfer internalizes the inefficiency and will not want to accept such a contract.

A feature of a perfect Bayesian equilibrium to the alternating-offers bargaining games is that a weak participation or individual rationality constraint must be satisfied in equilibrium. If i is the proposer in period t then j will not accept an offer that does not give it a payoff in the crisis game that is at least as high as the payoff from the null treaty. But the proposer would also never propose a treaty that gave it a lower payoff. Thus we see that every treaty, $\boldsymbol{\theta}$ that is reached in an equilibrium must satisfy the constrain $u_i(\boldsymbol{\theta}) \geq u_i(0,0)$ for $i = 1, 2$. Gaining θ_i comes at a cost and for at least one player this cost is not justified as the contract induces wars that should not be fought (with positive probability).

Proposition 2. *Assume that $p_j > k_j$. There is a perfect Bayesian equilibrium and in every such equilibrium the alliance agreement is* $(0,0)$.

It is instructive to contrast the situation where there are two large states with the case of two small states. With the latter, the only effect of an alliance is that it increases the probability of war through moral hazard, which is detrimental for both parties. In contrast, we saw that with two large states, an alliance agreement deters the challenger from making a threat, which stops war altogether. This generated a deterrence surplus, which are the generation of new "rents" accrued by the target when the challenger changes its decision from "threaten" to "accept the status quo." From these results, it is clear that if both states are small, having an alliance does not deter the challenger, which means there is no increase in the total welfare of the targets, and hence, no value of having an alliance at all.

4.3. **One large and one small target.**

Consider the situation where 1 is large enough so that if $\theta_1 \geq \underline{\theta}_1$ initiation against 1 will be deterred, but 2 is sufficiently small such that no treaty will deter a challenger from threatening it. To keep the meaning clear, we will use the label L for country 1, the large country, and s for country 2, the small country.

We begin the analysis of this mixed case by showing that in every perfect Bayesian equilibrium the large country ends up with a deterrent agreement.

Lemma 3. *Suppose that condition (1) holds for country L but not for country s. Then in every perfect Bayesian equilibrium the alliance agreement has*

$$\theta_L^* \geq \underline{\theta}_L.$$

From this lemma we see that the gains from an alliance for the large country are always captured. The question that remains is how these gains might be distributed. To answer this question it is sufficient to fix $\theta_L^* \geq \underline{\theta}_L$ and define

$$\bar{\theta}_s = \{\theta_s | u_L(\theta_L^*, \theta_s) = u_L(0, 0)\}$$

as the maximal transfer to s that L is willing to make in exchange for entering the crisis game with an alliance that deters initiation against it. Recall that because the challenger is deterred from threatening L, it is "free" for s to join the alliance – it will not need to make its promised transfer – whereas supporting s involves actual costs to L. Thus the two target countries bargain over how much L will pay in support of s so that L obtains deterrence. This reduces conceptually to bargaining between L and s over a pie of size $\bar{\theta}_s > 0$. For the remainder of this section we focus on the problem of how L and s divide this pie. Though an alliance is an agreement of the form $(\underline{\theta}_L, \theta_s)$., it is natural to think of them bargaining over the set

$$Y = \{(y_L, y_s) \in R^2 | y_1 + y_2 = \bar{\theta}_s \text{ and } y_i \geq 0 \text{ for } i = L, s\}.$$

This agreement results in the alliance $(\underline{\theta}_L, y_s)$. The game structure of alternating offers with the risk of breakdown has been well studied and our analysis consists mostly of showing that our particular application satisfies a known set of sufficient conditions for existence and uniqueness of equilibria. The results will hold when z, the probability that negotiations exogenously end after an alliance proposal is rejected, is sufficiently large. Following the notation of Osborne and Rubinstein (1990), we denote an alliance bargaining game with the risk of breakdown at a probability z as $\Gamma(z)$.

To give our first result regarding the characteristics of equilibrium alliance agreements we need some additional notation. We use notation that is standard in the bargaining literature in order to make the connection with results from that literature clear. Let $v_i(\theta_s, 1)$ be the amount of support given to the small country by the large country in period 0 that would make country i indifferent between this particular agreement at the beginning of the game and (possibly) getting the settlement that gives the small country θ_s in period 1. Recall that under both $v_i(\theta_s, 1)$ and θ_s, country L gets a deterrent transfer. For each player we have

$$(v_i(\theta_s, 1), 0) \sim_i (\theta_s, 1).$$

where we use the notation $\sim$ to denote indifference. In other words this amount $v_i(\theta_s, 1)$ satisfies the equivalence

$$u_i(\underline{\theta}_L, v_i(\theta_s, 1)) = z u_i(0, 0) + (1 - z) u(\underline{\theta}_L, \theta_s).$$

Lemma 4. *(Existence and Uniqueness) If*

$$\max_i \{ \sup_{\theta_s} | \frac{u_i'(\underline{\theta}_L, \theta_s)}{u_i'(\underline{\theta}_L, v_i(\theta_s, 1))} | \} < \frac{1}{1 - z} \tag{2}$$

then there is an essentially unique perfect Bayesian equilibrium to the alliance agreement game.

Although the sufficient condition for the lemma may be quite difficult to verify, the result is useful; our next proposition will use the result to provide a sufficient condition with an immediate substantive interpretation z, the risk of conflict in the immediate period, close to 1.. First, we highlight an important implication of the lemma, which illustrates that the alliance agreement can be described in a tractable way.

Corollary 1. *Assume the condition of Lemma 4 and let $\langle\langle x, t\rangle\rangle$ denote the lottery over receiving a transfer x in the t^{th} period and bargaining breakdown before t. Then the unique equilibrium solution solves*

$$\langle\langle y^*(z), 0\rangle\rangle \sim_L \langle\langle x^*(z), 1\rangle\rangle \text{ and } \langle\langle x^*(z), 0\rangle\rangle \sim_s \langle\langle y^*(z), 1\rangle\rangle. \tag{3}$$

And country s accepts any proposal such that $x_L \geq x_L^*(z)$ and country L accepts any proposal such that $y \leq y_s^*(z)$.

From the corollary, we may use the implicit function theorem to obtain comparative statics on the equilibrium transfers to the small state. To do so, we write the two indifference conditions as the following system of equations:

$$u_L(\underline{\theta}_L, y^*) - [z u_L(0,0) + (1-z) u_L(\underline{\theta}_L, x^*)] = 0 \tag{4}$$

$$u_s(\underline{\theta}_L, x^*) - [z u_s(0,0) + (1-z) u_s(\underline{\theta}_L, y^*)] = 0 \tag{5}$$

where the individual terms are given by

$$u_L(\underline{\theta}_L, y^*) = \pi_L + (1-\pi_L)(1 - F(x_s - p_s + y^*)y^*)$$

$$u_L(0,0) = \pi_L(F(x_L - p_L)[x_L - p_L - \hat{c}_L(0)] - x_L) + 1$$

$$u_L(\underline{\theta}_L, x^*) = \pi_L + (1-\pi_L)(1 - F(x_s - p_s + x^*)x^*)$$

$$u_s(\underline{\theta}_L, x^*) = (1-\pi_L)[F(x_s - p_s + x^*)(x_s - p_s - \hat{c}_s(x^*) + x^*) - x_s] + 1$$

$$u_s(0,0) = (1-\pi_L)[F(x_s - p_s)(x_s - p_s - \hat{c}_s(0)) - x_s] + 1$$

$$u_s(\underline{\theta}_L, y^*) = (1-\pi_L)[F(x_s - p_s + y^*)(x_s - p_s - \hat{c}_s(y^*) + y^*) - x_s] + 1$$

In the discussion that follows, we will refer to the system (4) and (5) as $H_1 = 0$ and $H_2 = 0$, where the dependence on parameters and endogenous values x^* and y^* are clear. For what follows, we will assume that the costs are drawn from the uniform distribution on support $[0, 1]$. Thus, the cut-points are given by $\hat{c}_s(y^*) = \frac{x_s - p_s + y^*}{2}$ and $\hat{c}_L(x^*) = \frac{x_L - p_L + x^*}{2}$ and $F(w) = w$ if w is in $[0, 1]$.

Before we find the comparative statics for solutions to this system, it is instructive to think about the equilibrium conditions. We see from the corollary that an equilibrium is characterized by a pair of indifference conditions. As second proposer, s needs to be indifferent between accepting the offer that is made by the first proposer, or rejecting it, and L, needs to be indifferent between accepting and rejecting the offer that s would make if it got the chance in the next round of bargaining. This indifference, of course, hinges on the offers that each player would make in the next round if they rejected their partner's offer in the current round. The amount of support that L pledges to s on the equilibrium path is the minimal amount that s is willing to accept given anticipation of equilibrium play following a rejection of L's proposal. Accordingly, the comparative statics on the amount of support promised to s, denoted x^*, depend on an analysis of the indifference condition for player s. This condition is equation (5). If there is an exogenous increase in π_L, which is the probability that L faces a crisis with the challenger, we observe that this unambiguously increases all three terms in equation (5). If s is less likely to be involved in a crisis, then s is better off when it has an alliance agreement with L because deterrence occurs. On the other hand, when L is less likely to be in a crisis, it is better off if there is no agreement between the two states because it will be more likely to make transfers to support s in a war. But since the left hand side of (5) is the difference of the current and continuation utilities, it is ambiguous whether x^* needs to increase or decrease to offset this exogenous shock.

In contrast, the indifference condition for L, equation (4), is less ambiguous. For z close enough to 1, the primary effects occur in the first and second terms of the left hand side of (4). Here an increase in π_L is good for L if an agreement is reached, since it is deterrent for L and bad if no agreement is reached. Thus, the left hand side of (4) increases so to maintain the equality with 0, y^* needs to decrease. This latter effect – that y^* or the amount of support offered by s to L decreases – is not picked up on the equilibrium path if L is the first proposer, because in equilibrium s accepts L's offer when L proposes. It would, however, be picked up in a different game in which s moved first in the bargaining game.

We now apply the implicit function theorem to analyze the comparative statics. Importantly, we must capture both direct and indirect effects of the parameters of interest. For example, a change in the parameter p_s will have a direct effect on the equilibrium values x^* and y^*, but, since x^* and y^* are also related in equilibrium, the effect of p_s on x^* will have an indirect effect on y^* that is weighted by the effect that a change in x^* has on y^*. To capture all of these direct and indirect effects we treat the left-hand side of the system (4) and (5) as a 2-by-1 vector $H(y^*, x^*, \gamma)$ and the right-hand side as a vector of 2 zeros. Here we let γ be a vector capturing the exogenous parameters $(x_s, x_L, p_s, p_L, \pi_L)$. The implicit function theorem tells us that as long as a technical condition known as the transversality condition is satisfied, then the derivative of $y^*(\theta)$ with respect to a coordinate of θ is given by total differentiation.

We focus on the comparative statics on x^* since this is the offer that is made and accepted on the equilibrium path. Using the implicit function theorem we obtain the following qualitative comparative statics for equilibrium values of $x^*(x_L, x_s, p_s, p_L, \pi_L)$. We focus on the effects of changes in p_s, p_L, and π_L.

Proposition 3. *Assume z is sufficiently close to 1. Then there is an essentially unique equilibrium and the offer that is made in period 1 and accepted, x^*, is increasing in p_s and p_L. For any fixed values of p_s and p_L, x^* is monotone in π_L, but* *it is either always decreasing in π_L or increasing (depending on parameters) in π_L for low values of p_s and then decreasing in π_L for higher values of p_s.*

5. Alliances that change the probability of victory

Up to this point, we have treated alliance agreements as a cost-sharing measure. In many ways, this is a useful theoretical choice, because it appeals to our intuition about how alliances affect alliance parters. This is a natural approach if we think of alliances as obligating countries to share the burden of war when there is a conflict and they are called upon to do so. Our analysis shows that even when an alliance between two potential targets does not change the payoffs to the challenger directly, the cost sharing and the resulting changes in the targets' incentives and actions can change the challenger's incentives to make a threat. In other words, by distorting the behavior of target countries, alliances can distort the behavior of attackers even when the alliance does not directly alter the payoffs of the attacker.

We might, however, think that alliances serve more of a capabilities aggregation function by enhancing alliance partners' strength relative to the challenger. In this case, agreements directly change the incentives of both the target and the challenger. A modeling strategy for alliances that have this alternative effect involves assuming that an alliance agreement $\boldsymbol{\rho} = (\rho_1, \rho_2)$ results in an increase in the probability that i wins a war by the quantity ρ_i. In this model then the set of possible alliances would involve $\rho \in [0, r_1] \times [0, r_2]$ with $r_i \leq 1 - p_i$. Let $c_i(r_i)$ denote the strictly increasing cost function capturing the cost to i's alliance partner, of increasing the probability that i wins by ρ_i. Like before, we may assume the costs are only borne if target i ends up at war with the challenger.

Much of the intuition from the private values case, in which an alliance only affects the targets' costs of war, extends here. The concept of a "large" country in the private values model translates here to the case where it is possible to increase the probability that i wins enough to deter initiation by the outsider. Thus the case of two large countries involves the assumption that for each i there exists a ρ_i such that $\rho_i \geq p_i - k_i$. As a result, there are alliances that fully deter war and end up costing the alliance partners nothing. Given alliances of this form exist, and if the second part of condition 1 holds, then every equilibrium must involve an alliance of this form.

The natural extension of our concept of small countries involves the assumption that $r_i < p_i - k_i$. In this case, it is not possible for j to make i strong enough so that an outsider believes i will fight rather than acquiesce. In other words, i prefers 0 over the lottery between x and war. Under this condition, it is not possible to deter the initiator but, in contrast to the case of private values, a treaty that changes the probability that i wins need not be inefficient. In particular, consider when ρ_i satisfies the condition $p_i - \rho_i > k_i$, so that the initiator is not deterred from attacking i. The promise ρ_i has two effects on i's payoffs. It expands the set of costs-types – the set of targets with different war costs – that will reject the challenger's threat (and thus fight) and it increases the payoff to i if it fights. Conditional on fighting, the payoff to i increases by ρ_i times the stakes of war (which are 1) and conditional on j transferring $c_i(\rho_i)$ to i the value gained by i is exactly ρ_i. Accordingly, an alliance between two small countries can be efficient if and only if $c_i(\rho_i) \leq \rho_i$.

Thus, in the case of two small countries, a non-trivial treaty is possible in equilibrium. In particular, contrasted with the private values model, 1 and 2 now bargain over treaties in which a treaty ρ results in the gain to i of $\pi_i F_i(x_i - p_i + \rho_i)\rho_i$ and it results in a cost to j of $\pi_j F_j(x_j - p_j + \rho_j)c_j(\rho_j)$. In order for i to be willing to accept (or propose) a treaty of this form in equilibrium, the treaty must provide weakly positive gains to i. Moreover, as this condition must be satisfied for both players, any treaty that is accepted with positive probability in an equilibrium must simultaneously satisfy the inequalities

$$\pi_1(F_1(x_1 - p_1 + \rho_1)\rho_1 \geq \pi_2(F_2(x_2 - p_2 + \rho_2)c_2(\rho_2)$$

$$\pi_2(F_2(x_2 - p_2 + \rho_2)\rho_2 \geq \pi_1(F_1(x_1 - p_1 + \rho_1)c_1(\rho_1).$$

An immediate observation is that in the case where $c_1(\rho) = c_2(\rho) = \rho$, either both equations are satisfied with equality or neither is satisfied. In this case, small countries may form an alliance, but the payoffs are the same as in the trivial treaty, $\rho = (0, 0)$. We can then conclude that if a technology of war takes a transfer ρ from small country i and turns it into an increase in the probability of victory for j that is exactly equal to the cost of the transfer, then there is no incentive for i and j to form an alliance, even if j's chances of winning a war are improved. The case where $c_i(\rho_i) > \rho_i$ for $i = 1, 2$, will also not support non-trivial treaties, as these treaties involve inefficiencies.

We are left with the case where the technology of war makes the probability of winning increase at a faster rate than the cost of transfers to the ally. Here, small countries will have an incentive to increase their collective payoffs by supporting each other during fighting. For example, consider a $c_i(\rho_i) = \alpha\rho_i$ for $i = 1, 2$ with $\alpha < 1$, and $x_1 = x_2 = x; p_1 = p_2 = p; \pi_1 = \pi_2$, with uniform distributions on the players costs. Then the above system of inequalities becomes

$$\frac{x - p + \rho_1}{x - p + \rho_2} \geq \frac{\alpha\rho_2}{\rho_1},$$
$$\frac{\rho_2}{\alpha\rho_1} \geq \frac{x - p + \rho_1}{x - p + \rho_2}.$$

Now a set of alliances that are Pareto superior to the trivial treaty exists. For example, alliances with $\rho_1 = \rho_2$ are in the interior of the set of treaties satisfying these constraints. Under our bargaining protocol, some agreement in which $\rho_i > 0$ for at least one player (and both in protocols that are not too extreme) would emerge. Even without characterizing the equilibrium to such a model here, we can make two observations concerning the conditions that lead to productive alliances in international relations. First, alliances are attractive when it is possible for a transfer between target country 1 and target country 2 to alter the behavior of a potential challenger. This is what makes alliances between a large country and a small country viable in the private values model, and it is also what makes an alliance between two small countries supportable, under certain conditions, in the common values model. However, it is also important to observe that the formation of an alliance in the common values model relies on a technological benefit of the treaty; without such a technology, small countries would not find alliances beneficial.

6. Conclusion

By viewing security alliances as a form of decentralized insurance and focusing on alliance commitments in which aid may cause moral hazard, we make four key findings. First, when two large countries are threatened, they form alliances that look like what international relations scholars might call threat balancing. Each side commits to aid the other to a degree that the challenger is deterred from escalating a dispute. Second, the ability of large states to manipulate the incentives of the challenger by making allies more aggressive is a key element to explaining how security commitments arise. In contrast, an alliance of two small country targets fails to deter the challenger, with the result that moral hazard increases the incidence of war and the alliance serves to redistribute the war's costs from the party at war to its ally. But since the cost of war is completely internalized by the target who is not at war, these social welfare decreasing agreements cannot arise in equilibrium.

Third, when analyzing the asymmetric alliance case, we see that alliance agreements always generate private benefits through deterrence for the large country. The formation of an alliance in this environment then turns on the bargaining between the now-safe large country and the still-threatened small country regarding how much of this benefit will be returned in the form of transfers to the small country. Finally, we see that when the risk of a crisis is severe, that is, when z is sufficiently large, there is unique equilibrium and we can make strong prediction regarding the bargaining outcome between targets. This may explain, in part, why many actual alliance agreements in history have been forged on the eve of a crisis and why they describe in detail the conditions of activation as well as the amount and form of aid.

This paper has sought to explain why countries form alliances and what terms they choose when they do so. Our analysis employed a basic agreement structure in which allies exchange security guarantees, which represents a foundational class of alliances agreements in international relations. Although empirically alliance agreements take many forms, the basic environment examined here addresses fundamental questions concerning the role of moral hazard in producing deterrence and thus in alliance formation. As suggested by the empirical evidence, a more complete understanding of the role of alliances in international politics will require further investigation into why and what kind of agreements are written in this basic environment.

7. Appendix

This appendix contains proofs of lemmas and propositions not given in the main text.

Lemma 1.

Proof. Suppose not. That is, suppose there were some equilibrium where at period t the two targets reached an alliance agreement where, for some j, $\theta_j^* < \underline{\theta}_j$. There are two cases.

Case (1): Suppose there is an equilibrium which puts positive probability on a treaty satisfying $\theta_1^* < \underline{\theta}_1$ and $\theta_2^* < \underline{\theta}_2$. Then at some period t some i proposes an agreement (θ_1^*, θ_2^*) and this proposal is accepted (with positive probability). As a result

$$u_j(\theta_1^*, \theta_2^*) \geq z u_j(0,0) + (1-z)W_j(t+1)$$

where $W_j(t+1)$ is j's continuation value for the game that starts after she is the veto player in period t.

Now suppose at time t country i proposes $(\underline{\theta}_i, \theta_j^*)$. First, $u_j(\underline{\theta}_1, \theta_2^*) > u_j(\theta_1^*, \theta_2^*)$ because on the path j never has to pay $\underline{\theta}_i$, but pays $\theta_1^* > 0$ with positive probability, while at the same time j's payoff to their own crisis does not change. Thus we can conclude that $(\underline{\theta}_i, \theta_j^*)$ is accepted by j at time t.

All that remains is to show that at t, i is strictly better-off proposing $(\underline{\theta}_i, \theta_j^*)$. For country i the expected utility of (θ_i^*, θ_j^*) is

$$\begin{aligned}(6)\quad \pi_i[F(x_i - p_i + \theta_i^*)(1 - p_i - \hat{c}_i(\theta_i^*) + \theta_i^*) + (1 - F(x_i - p_i + \theta_i^*))(1 - x_i)]\\ + (1-\pi)[1 - F(x_j - p_j + \theta_j^*)\theta_j^*],\end{aligned}$$

where $\hat{c}_i(\theta_i) = \mathbb{E}[c_i | c_i < x_i - p_i + \theta_i]$ denotes the expected cost of player i conditional on the cost being sufficiently low that i fights.

The expected utility of i for $(\underline{\theta}_i, \theta_j^*)$ is

$$\pi_i[1] + (1 - \pi_i)[1 - F(x_j - p_j + \theta_j^*)\theta_j^*]. \tag{7}$$

By assumption $[F(x_i - p_i + \theta_i^*)(1 - p_i - \hat{c}_i(\theta_i^*) + \theta_i^*) + (1 - F(x_i - p_i + \theta_i^*))(1 - x_i)] < 1$, and this proposal is a profitable deviation, a contradiction.

Case (2): Now suppose there is an equilibrium which puts positive probability on a treaty satisfying $\theta_i \geq \underline{\theta}_i$ and $\theta_j < \underline{\theta}_j$. There are two sub-cases.

Sub-case (i): Suppose this agreement is reached at a time t when j is the proposer. By an argument parallel to the one in Case (1), j has a profitable deviation, a contradiction.

Sub-case (ii): Now suppose that this agreement is reached at a time t when i is the proposer and $\theta_j^* > 0$. If i increases the proposed support to country j to some $\theta_j \geq \underline{\theta}_j$, then j will never be attacked and i's expected payout to j is $0 < (1 - \pi_i)F(x_j - p_j + \theta_j^*)\theta_j^*$. This is a profitable deviation for i, a contradiction. If $\theta_j^* = 0$, but j is a proposer in some future period, j will reject this offer to get the lottery over zero and being the proposer at some future t'. This contradicts that the agreement is reached at period t.

Together these cases prove the lemma. □

Lemma 2.

Proof. Suppose not. That is, suppose there is a PBE that reaches an agreement, (θ_1^d, θ_2^d), at period $s > 0$. From Lemma 1 we know that this perfect Bayesian equilibrium agreement will be deterrent for both targets. Let j be the veto player in the first period. From period 0 the veto player i has expected utility given by

$$zu_i(0,0) + (1-z)([zu_i(0,0) + (1-z)[zu_i(0,0) + (1-z)[\ldots + (1-z)^s u_i(\theta_1^d, \theta_2^d)]]]\ldots$$

$$= zu_i(0,0) + (1-z)u_i(0,0) + (1-z)^2 u_i(0,0) + \ldots + (1-z)^s u_i(\theta_1^d, \theta_2^d)$$

$$= zu_i(0,0)\sum_{t=0}^{s-1}(1-z)^t + (1-z)^s u_i(\theta_1^d, \theta_2^d).$$

To show that a deterrent agreement would be accepted in period $t = 0$, suppose it were rejected by i. Then

$$u_i(\theta_1^d, \theta_2^d) \leq zu_i(0,0)\sum_{t=0}^{s-1}(1-z)^t + (1-z)^s u_i(\theta_1^d, \theta_2^d),$$

$$u_i(\theta_1^d, \theta_2^d) \leq \frac{zu_i(0,0)\sum_{t=0}^{s-1}(1-z)^t}{1-(1-z)^s},$$

$$= zu_i(0,0)\Big[\frac{1-(1-z)^s}{z}\frac{1}{1-(1-z)^s}\Big],$$

$$= zu_i(0,0)\frac{1}{z} = u_i(0,0).$$

But by assumption $u_i(\theta_1^d, \theta_2^d) > u_i(0,0)$, a contradiction.

We can thus conclude that if a target i would accept the deterrent equilibrium agreement in period $s > 0$, then it would accept it in period 0. We are left to show that the proposer in period $t = 0$ is better off making the proposal today. But this is clear from an argument parallel to the one that shows the veto player is willing to accept a deterrent proposal today if it is willing to accept it in the future.

This contradiction proves the lemma. □

7.1. **Proposition 2.**

Proof. We prove the second part first. An immediate consequence of the pair of inequalities that precede the proposition is that if a treaty, $\boldsymbol{\theta}$ is passed in any equilibrium then

$$u_1(\boldsymbol{\theta}) + u_1(\boldsymbol{\theta}) \geq u_1(0,0) + u_2(0,0).$$

But since the null treaty is Pareto efficient, this implies that no treaty which is not payoff equivalent to the null treaty can be accepted in any equilibrium. To establish existence consider the profile in which each target proposes the null treaty at every history in which she is proposer; each proposer accepts any treaty that yields a weakly higher continuation payoff than continuation with the null treaty and bargaining in the crisis game is as described above. Given this profile and the inequalities, there is no treaty that will be accepted which gives the proposer a payoff higher than the null treaty. Given the proposal strategies, the acceptance rule described is sequentially rational. □

7.2. **Lemma 3.**

Proof. Suppose not. That is, suppose that for country L condition (1) holds, but not for country s, and there is a perfect Bayesian equilibrium agreement (θ_L^*, θ_s^*) and $\theta_L^* < \underline{\theta}_L$. We show that a profitable deviation exists. There are two cases.

Case (1): First suppose that at some time $t \geq 0$ the two targets reach an agreement of (θ_L^*, θ_s^*) and $\theta_L^* < \underline{\theta}_L$.

Sub-case (i): Suppose L is the proposer in period t. Because s accepts (θ_L^*, θ_s^*) it must be the case that

$$u_s(\theta_L^*, \theta_s^*) \geq zu_s(0,0) + (1-z)W_s(t+1).$$

where $W_s(t+1)$ is the continuation payoff to s in this conjectured equilibrium. Evaluating s's expected utilities at both this profile and $(\underline{\theta}_L, \theta_s^*)$ we see that

$$E[u_s(\underline{\theta}_L, \theta_s^*)] - E[u_s(\theta_L^*, \theta_s^*)]$$

$$= \pi_L - \pi_L(1 - F(x_L - p_l + \theta_L^*)\theta_L^*)$$

$$= \pi_L F(x_L - p_L + \theta_L^*)\theta_L^* > 0$$

and s will also accept $(\underline{\theta}_L, \theta_s^*)$.

All that remains is to show that the large country is better-off proposing $(\underline{\theta}_L, \theta_s^*)$. As the large country's utility is increasing in θ_L this is a profitable deviation contradicting that (θ_L^*, θ_s^*) is a perfect Bayesian equilibrium when the agreement is accepted in period t and L is the proposer.

Sub-case (ii): Suppose that s is the proposer at period t. By a similar argument as above, if the large country accepts (θ_L^*, θ_s^*), then it will accept $(\underline{\theta}_L, \theta_s^*)$.

We are left to show that when s is the proposer at time t, s's expected utility is greater if it proposes $(\underline{\theta}_L, \theta_s^*)$. As above, we see that the difference in s's expected utility from these two alliances is

$$F(x_L - p_L + \theta_L^*)\theta_L^* > 0$$

and s is better-off with the agreement $(\underline{\theta}_L, \theta^*_s)$ because s never has to pay any transfer in this case. This contradiction shows there is no perfect Bayesian equilibrium with an agreement (θ^*_L, θ^*_s) where s is the proposer of the accepted offer at time t.

Case (2): The second case considers the situation where no agreement is ever reached and the target countries's payoffs are $u_i(0,0)$. This means that there is some even period where L makes a proposal of $(0,0)$ or some other $(\hat{\theta}_L, \hat{\theta}_s)$ which is rejected by s. In this period s's expected utility of rejecting is $u_s(0,0)$.

Suppose L proposes $(\underline{\theta}_L, \epsilon)$. For s, this proposal raises its expected utility and would be accepted. This is also a strictly profitable deviation for L, contradicting that there is an equilibrium with $(0,0)$ or perpetual disagreement.

This completes the proof of the lemma.

□

Lemma 4.

Proof. From Osborne and Rubinstein (1990) Theorem 3.4, we have the useful result that if an alternating offers bargaining game satisfies the conditions that

A1 Disagreement is the worst outcome,

A2 pie is desirable,

A3 time is valuable,

A4 the preference order is continuous,

A5 the preferences are stationary,

A6 and there is increasing loss to delay,

then the bargaining game has an (essentially) unique subgame perfect equilibrium. It is easy to see that conditions A1–A5 are satisfied in the alliance game $\Gamma(z)$ for any z in the interval. The increasing difference condition (condition A6) holds if the difference

$$\theta_s - v_i(\theta_s, 1) \tag{8}$$

is an increasing function of θ_s –the share of the surplus from an alliance to country s. Recall that v_i is the relation defined above. For condition A6 to be true for the small state it is sufficient to show that

$$\frac{\partial v_s(\theta_s, 1)}{\partial \theta_s} < 1.$$

Let the $v_s(\theta_s, 1)$ be denoted by $\hat{\theta}_s^s$. Then we have $\hat{\theta}_s^s$ implicitly defined by ,

$$u_s(\underline{\theta}_L, \hat{\theta}_s^s) = zu_s(0,0) + (1-z)us(\underline{\theta}_L, \theta_s)$$
$$u_s(\underline{\theta}_L, \hat{\theta}_s^s) - zu_s(0,0) - (1-z)u_s(\underline{\theta}_L, \theta_s) = 0$$

By the implicit function theorem we have

$$\frac{\partial \hat{\theta}_s^s}{\partial \theta_s} = (1-z)\frac{u_s'(\underline{\theta}_L, \theta_s)}{u_s'(\underline{\theta}_L, \hat{\theta}_s^s)} \text{ for all } \theta_s. \tag{9}$$

Combining this equation and the previous inequality we obtain the sufficient condition,

$$|\frac{u_s'(\underline{\theta}_L, \theta_s)}{u_s'(\underline{\theta}_L, \hat{\theta}_s^s)}| < \frac{1}{(1-z)} \text{ for all } \theta_s. \tag{10}$$

A parallel argument for the large country yields the other sufficient condition

$$|\frac{u_L'(\underline{\theta}_L, \theta_s)}{u_L'(\underline{\theta}_L, \hat{\theta}_s^L)}| < \frac{1}{(1-z)} \text{ for all } \theta_s. \tag{11}$$

If z satisfies the condition of the proposition, then both of theres inequalities hold, and A6 of Osborne and Rubinstein's Theorem 3.4 is satisfied. As a result the alliance game has a unique perfect Bayesian equilibrium. □

Proposition 3.

Proof. Existence and uniqueness for sufficiently large z follows from the previous lemma.

We now obtain the comparative statics. The implicit function theorem states that for a given parameter $\eta \in p_s, p_L, \pi_L$ we get

$$\begin{bmatrix} \frac{\partial y^*}{\partial \eta} \\ \frac{\partial x^*}{\partial \eta} \end{bmatrix} = - \begin{bmatrix} \frac{\partial U_L}{\partial y^*}\left(\widehat{\theta}_L, y^*\right) & -\left(1-z\right)\frac{\partial U_L}{\partial x^*}\left(\widehat{\theta}_L, x^*\right) \\ -\left(1-z\right)\frac{\partial U_S}{\partial y^*}\left(\widehat{\theta}_L, y^*\right) & \frac{\partial U_S}{\partial x^*}\left(\widehat{\theta}_L, x^*\right) \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial H_1}{\partial \eta} \\ \frac{\partial H_2}{\partial \eta} \end{bmatrix},$$

if the Jacobian,

$$\begin{bmatrix} \frac{\partial U_L}{\partial y^*}\left(\widehat{\theta}_L, y^*\right) & -\left(1-z\right)\frac{\partial U_L}{\partial x^*}\left(\widehat{\theta}_L, x^*\right) \\ -\left(1-z\right)\frac{\partial U_S}{\partial y^*}\left(\widehat{\theta}_L, y^*\right) & \frac{\partial U_S}{\partial x^*}\left(\widehat{\theta}_L, x^*\right) \end{bmatrix}$$

has full rank (and thus its inverse exists). Differentiating the particular expressions above yields the following Jacobian

$$\begin{bmatrix} \left(1-\pi_L\right)\left(p_S - x_S - 2y^*\right) & -\left(1-z\right)\left(1-\pi_L\right)\left(p_S - x_S - 2x^*\right) \\ -\left(1-z\right)\left(\pi_L - 1\right)\left(p_S - x_S - y^*\right) & \left(\pi_L - 1\right)\left(p_S - x_S - x^*\right) \end{bmatrix}.$$

Taking the inverse of the Jacobian, we have

$$\frac{1}{D}\left[\begin{array}{cc} \left(p_s-x_s-x^*\right) & -\left(1-z\right)\left(p_s-x_s-2x^*\right) \\ \left(1-z\right)\left(p_s-x_s-y^*\right) & -\left(p_s-x_s-2y^*\right) \end{array}\right]$$

where

$$D=\left(1-\pi_L\right)\left(\left(p_s-x_s-2y^*\right)\left(p_s-x_s-x^*\right)-\left(p_s-x_s-2x^*\right)\left(p_s-x_s-y^*\right)\left(1-z\right)^2\right)$$

The implicit derivatives for x^* come from the second entry of the vector defined above (corresponding to equation (11)). Now, to see that x^* is increasing in p_S we first observe that $\frac{\partial H_1}{\partial p_S}=\left(1-\pi_L\right)\left(x^*\left(z-1\right)+y^*\right)$ and $\frac{\partial H_2}{\partial p_S}=\left(\pi_L-1\right)\left(x^*+y^*\left(z-1\right)\right)$. Thus we obtain,

$$\frac{\partial x^*}{\partial p_S}=-\left(\frac{\left(y^*+x^*\left(z-1\right)\right)\left(z-1\right)\left(x_s-p_s+y^*\right)-\left(x^*+y^*\left(z-1\right)\right)\left(x_s-p_s+2y^*\right)}{\left(x_s-p_s+x^*\right)\left(x_s-p_s+2y^*\right)-\left(z-1\right)^2\left(x_s-p_s+y^*\right)\left(x_s-p_s+2x^*\right)}\right).$$

Our proof of uniqueness of perfect Bayesian equilibrium involved taking z sufficiently close to 1. We thus continue to work with large enough values of z. As $z\longrightarrow 1$, $\frac{\partial x^*}{\partial p_S}$ converges to $\left(\frac{(x^*)(x_s-p_s+2y^*)}{(x_s-p_s+x^*)(x_s-p_s+2y^*)}\right)$. Since the cut-point $x_S-p_S+y^*>0$, it must be the case that $\left(\frac{(x^*)(x_s-p_s+2y^*)}{(x_s-p_s+x^*)(x_s-p_s+2y^*)}\right)>0$, which implies $\frac{\partial x^*}{\partial p_S}>0$.

To see that x^* is increasing in p_L, we first observe that $\frac{\partial H_1}{\partial p_L}=-\pi_L z(p_L-x_L)$ and $\frac{\partial H_2}{\partial p_L}=0$. Using the implicit function theorem we obtain

$$\frac{\partial x^*}{\partial p_L}=\left(\frac{\left(1-z\right)\left(-\pi_L z\left(p_L-x_L\right)\right)\left(x_s-p_s+y^*\right)}{\left(1-\pi_L\right)\left(\left(x_s-p_s+x^*\right)\left(x_s-p_s+2y\right)+\left(1-z\right)^2\left(x_s-p_s+y^*\right)\left(x_s-p_s+2x^*\right)\right)}\right).$$

Since the denominator is positive and $-\pi_L z\left(p_L-x_L\right)>0$, we therefore obtain $\frac{\partial x^*}{\partial p_L}>0$.

Finally, for the comparative static on π_L we again take $z \to 1$ and obtain

$$\frac{\partial x^*}{\partial \pi_L} = \frac{-x^*\left(x_s - p_s + \frac{1}{2}x^*\right)}{\pi_L\left(x_s - p_s + x^*\right)}.$$

The denominator is positive as it is the value of the cut-point. From our analysis of the comparative static on p_S we know that as z approaches 1, $\frac{\partial x^*}{\partial p_S}$ approaches $\frac{x^*}{x_s - p_s + x^*}$. So for z large enough x^* increases in p_S with slope greater than 1 (as $p_S > x_s$ has been assumed throughout the paper). Thus $\frac{\partial x^*}{\partial \pi_L}$ is either negative always or it is positive for some low values of p_s and then negative for larger values of p_s.

$\square$

References

Binmore, Ken, Ariel Rubinstein and Asher Wolinsky. 1986. "The Nash Bargaining Solution in Economic Modelling." *Rand Journal of Econometrics* 17(2):176–188.

Christensen, Thomas J. and Jack Snyder. 1990. "Chain Gangs and Passed Bucks: Predicting Alliance Patterns in Multipolarity." *International Organization* 44(2):pp. 137–168.

Conybeare, John A. C. 1992. "A Portfolio Diversification Model of Alliances: The Triple Alliance and Triple Entente, 1879-1914." *The Journal of Conflict Resolution* 36(1):pp. 53–85.

Crawford, Timothy. 2001. "Pivotal Deterrence and the Kosovo War: Why the Holbrooke Agreement Failed." *Political Science Quarterly* 116(4):499–523.

Crawford, Timothy. 2005. "Moral Hazard, Intervention and Internal War: A Conceptual Analysis." *Ethnopolitics* 4(2):175–193.

Crawford, Timothy Wallace. 2003. *Pivotal Deterrence: Third-Party Statecraft and the Pursuit of Peace.* Ithaca, NY: Cornell University Press.

Fearon, James D. 1997. "Signaling foreign policy interests: tying hands versus sinking costs." *The Journal of Conflict Resolution* 41(1):68–90.

Garfinkel, Michelle R. 2004. "Stable Alliance Formation in Distributional Conflict." *European Journal of Political Economy* 20:829–852.

Huth, Paul K. 1991. *Extended Deterrence and the Prevention of War.* New Haven, CT: Yale University Press.

Jervis, Robert. 1994. What do we want to deter and how do we deter it? In *Turning Point: The Gulf War and U.S. Military Strategy*, ed. L. Benjamin Edington and Michael J. Mazarr. Boulder, CO: Westview pp. 122–124.

Kuperman, Alan J. 2008. "The Moral Hazard of Humanitarian Intervention: Lessons from the Balkans." *International Studies Quarterly* 52:49–80.

Leeds, Brett Ashley, Andrew G. Long and Sara McLaughlin Mitchell. 2000. "Reevaluating Alliance Reliability: Specific Threats, Specific Promises." *The Journal of Conflict Resolution* 44(5):pp. 686–699.

Morrow, James D. 1991. "Alliances and Asymmetry: An Alternative to the Capability Aggregation Model of Alliances." *American Journal of Political Science* 35(4):pp. 904–933.

URL: *http://www.jstor.org/stable/2111499*

Morrow, James D. 1994. "Alliances, Credibility, and Peacetime Costs." *The Journal of Conflict Resolution* 38(2):pp. 270–297.

Olson, Mancur and Richard Zeckhauser. 1966. "A Theory of Alliance Formation." *Review of Economics and Statistics* 266-279.

Osborne, Martin J. and Ariel Rubinstein. 1990. *Bargaining and Markets.* Emerald Group Publisher.

Pauly, Mark V. 1968. "The Economics of Moral Hazard: Comment." *The American Economic Review* 58(3):531–537.

Pauly, Mark V. 1974. "Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection." *The Quarterly Journal of Economics* 88(1):44–62.

Poast, Paul. 2012*a*. "Can Issue Linkage Improve Treaty Credibility? Buffer States Alliances as a 'Hard Case'." *working paper* .

Poast, Paul. 2012*b*. "Does Issue Linkage Work? Evidence from European Alliance Negotiations, 1860 to 1945." *International Organization* forthcoming.

Rauchhaus, Robert. 2005. "Conflict Management and the Misapplication of Moral Hazard Theory." *Ethnopolitics* 4(2):215–224.

Rauchhaus, Robert. 2009. "Principle-Agent Problems in Humanitarian Intervention: Moral Hazards, Adverse Selection, and the Commitment Dilemma." *International Studies Quarterly* 53(4):871–884.

Rubinstein, Ariel. 1982. "Perfect Equilibrium in a Bargaining Model." *Econometrica* 50(1):97–109.

Sandler, Todd. 1993. "The Economic Theory of Alliances: A Survey." *Journal of Conflict Resolution* 37:446–483.

Sandler, Todd and Keith Hartley. 2001. "Economics of Alliances: The Lessons for Collective Action." *Journal of Economic Literature* 39:869–896.

Shavell, Steven. 1979. "On Moral Hazard and Insurance." *The Quarterly Journal of Economics* 93(4):541–562.

Smith, Alastair. 1995. "Alliance Formation and War." *International Studies Quarterly* 39(4):pp. 405–425.

Smith, Alastair. 1998. "Extended Deterrence and Alliance Formation." *International Interactions* 24(4):315–343.

Snyder, Glenn H. 1984. "The Security Dilemma in Alliance Politics." *World Politics* 36(4):pp. 461–495.

Snyder, Glenn H. 1997. *Alliance politics.* Ithaca, NY: Cornell Univ. Press.

Wagner, Harrison. 2005. "The Hazards of Thinking about Moral Hazard." *Ethnopolitics* 4(2):237–246.

Walt, Stephen M. 1990. *The origins of alliances.* Ithaca: Cornell University Press.

Yuen, Amy. 2009. "Target Concessions in the Shadow of Intervention." *Journal of Conflict Resolution* 53(5):745–773.

Zagare, Frank C. and D. Marc Kilgour. 2003. "Alignment Patterns, Crisis Bargaining, and Extended Deterrence: A Game-Theoretic Analysis." *International Studies Quarterly* 47(4):pp. 587–615.

URL: *http://www.jstor.org/stable/3693637*