Dynamic Screening in International Crises

Noam Reich
NYU Abu Dhabi
noam.reich@nyu.edu

July 16th, 2024

Abstract

In this article, I introduce dynamic screening theory as a theoretical framework for understanding how states behave and assess resolve in crises. Modeling crisis diplomacy as a war of attrition, I ask how long countries will be willing to participate in costly negotiations and invest in sunk costs and audience costs before going to war or conceding. I show that more resolved states display greater impatience with diplomacy, preferring to fight instead. In turn, the least resolved states prefer to concede quickly to avoid having to fight. Finally, moderately resolved states negotiate longer, pay more sunk costs, and accumulate more audience costs. Consequently, moderately resolved states are more likely to obtain concessions, not because belief in their resolve increases, but because they grant their rival more time to concede. The model also features stalemated negotiations, providing new microfoundations for a common crisis outcome.

Keywords: Diplomacy, Dynamic Screening, Costly Signaling, Formal Theory

Supplementary material for this article is available in the appendix in the online edition. Replication files are available in the *JOP* Data Archive on Dataverse (https://dataverse.harvard.edu/dataverse/jop). The empirical analysis has been successfully replicated by the JOP replication analyst.

For decades, costly signaling theory has been the dominant theoretical framework for explaining how states communicate resolve in international crises. This theory maintains that states can demonstrate their willingness to fight by taking costly actions, namely paying sunk costs or accumulating audience costs (Fearon 1994, 1997; Kurizaki 2007; Slantchev 2005). A state that engages in such behavior can convince its rival that it is not bluffing about its intentions to go to war and intimidate it into making a concession. For these reasons, costly signaling is thought to be an important tool of coercive diplomacy that enables resolved states to avoid war.

In this paper, I introduce dynamic screening theory as a rival theoretical framework in which states that invest more in sunk costs and audience costs can appear less resolved. The logic of the argument is straightforward. The more attractive an option war is to a state, the lower its incentive to try to avoid fighting. It follows that the unresolved states most desperate to avoid war are those most willing to invest in sunk costs, audience costs, or any other form of diplomacy that can secure a peaceful concession. In an international crisis where states have private information about their resolve, a state that chooses to spend time and effort on producing costly signals reveals that it believes diplomacy to be more profitable than immediate war. The more a state pursues a non-violent solution to a conflict, the more intimidated by war it must be.

Indeed, states' investments in costly signals are often overshadowed by their efforts to avoid war. For example, during the lengthy prelude to the First Gulf War, US allies remained concerned that the US would waver despite its massive military mobilization and efforts in organizing an international coalition. Though costly signaling theory predicts that these actions should have dispelled any doubt regarding US resolve, US allies remained wary of the lengthy delay and outright panicked when the US organized a summit with Iraq (Baker 1995, 346-353; Freedman and Karsh 1993, 240-243). More recently, the Obama administration failed to convince Israel that it would use force to prevent Iran from acquiring a nuclear weapon. As I argue below, the administration's determined pursuit of peace via open-ended negotiations undermined US assurances and diminished the commitment displayed by its economic sanctions campaign.

However, if resolved states do not engage in costly signaling, how do they behave in crises? Why do states invest in coercive diplomacy if it conveys hesitation? How do states learn about their rivals? To answer these questions, I develop a theory of dynamic screening that builds on the above arguments and offers a framework for explaining how states behave and assess resolve in international crises.

To construct the dynamic screening theory, I model an international crisis between two countries as a war of attrition, an open-ended affair that continues until either one of the two parties involved takes one of two unilateral actions. First, either of the two countries can choose to terminate the crisis at any time by conceding to its rival's demands. Alternatively, they can choose to escalate the crisis by starting a war, modeled as a costly lottery, and in doing so force their rival to go to war as well. I assume that countries have private information about their war payoff and do not know whether their rival prefers to fight or concede or how long they will prolong the crisis before acting. As a result, crises resemble a "war of the nerves" (Fearon 1994) with both countries claiming to be resolved and perpetuating the crisis in the hope that their rival is bluffing and will concede.

Crucially, participation in a crisis is not cost-free. First, I assume that countries must pay sunk costs for every second they choose to delay conceding or escalating the crisis. These include the costs of troop deployments, the economic costs of heightened tensions, and the opportunity costs to leaders for having to manage the crisis. Second, I assume that countries continuously accumulate audience costs that must be paid if a country concedes. How long a country is willing to prolong the crisis depends on its willingness to pay these costs. Dynamic screening refers to how the costs of participating in a crisis compel countries with different levels of resolve to exit the crisis at different times. A state is "screened" whenever its resolve is too low or too high to let it continue negotiating.

The equilibrium to the model is characterized by two dynamic screening processes. First, resolved states, defined here as states who would prefer to go to war than concede, are screened by sunk costs. This occurs because resolved states face a trade-off: though they would rather their rival concede than start a war, delaying war in the hope of a concession requires that they continuously pay sunk costs. Because war is inefficient, resolved states can benefit from granting their rival a short opportunity to concede peacefully. However, there is no guarantee that postponing a war will lead to a concession as countries cannot be sure whether their rival will ultimately concede. Resolved states have limited patience for paying sunk costs and, if enough time passes without a concession, will choose to fight. The higher a state's wartime payoff, the less worthwhile it is to wait for a concession, and the earlier it will to go to war.

The second dynamic screening process occurs when unresolved states, defined here as those states that would rather concede than fight, are screened by the risk of war. One implication of resolved states being screened by sunk costs is that states cannot know whether their rival will abruptly escalate the crisis and start a war until they actually do so. As a result, states that remain in the crisis long enough can have war thrust upon them even if they had been willing to concede peacefully. This presents unresolved states with their own trade-off. On the one hand, unresolved states can benefit from misrepresenting their type – instead of conceding, an unresolved state can perpetuate a crisis and pretend to be resolved, threatening to go to war if their rival does not concede in the hope that their rival is also unresolved and will choose to concede first. However, doing so requires that unresolved states risk having to fight. Since unresolved states with worse payoffs to fighting have more to lose if their rival suddenly decides to start a war, they will opt to concede earlier.[1]

Together, these dynamic screening processes generate three key results. First is a set of novel comparative statics relating a state's resolve to crisis outcomes. In particular, the model predicts that resolved states spend less time and effort negotiating, are more likely to go to

[1] Sunk costs and audience costs also make waiting less worthwhile for unresolved states. However, willingness to delay and pay these costs is independent of a state's war payoff conditional on having chosen to concede.

war, and less likely to obtain concessions. Conversely, moderately resolved types perpetuate crises the longest. In doing so they pay more sunk costs, accumulate more audience costs, and are more likely to obtain concessions. However, moderately resolved states are more likely to obtain concessions because they grant their rival more time to concede, not because their rival's belief in their resolve increases. These results stand in direct contradiction with costly signaling theory, which predicts that more resolved states invest more in sunk costs and audience costs and are more likely to obtain concessions because their rivals will view their actions as an indication of resolve.

Second, under dynamic screening states gradually learn about their rival's resolve from the length of delay. When a state allows a crisis to drag on, it reveals that it prefers to pay sunk costs rather than go to war. Simultaneously, a state that does not concede even after a lengthy crisis reveals that it is at least willing to risk fighting even if it does not want to initiate a war itself. As a result, states will come to view their rival as being moderately resolved over time. This contrasts with costly signaling theory's emphasis on large and discontinuous shifts in beliefs following ostentatious actions undertaken with the purpose of demonstrating resolve.

Third, dynamic screening theory provides a micro-foundation for stalemate outcomes in international crises. I show that there can exist an endogenous date in a crisis after which both states recognize that if their rival has not yet conceded or escalated the dispute, then they will never do so. In this case the crisis continues in perpetuity and neither state receives the good or issue under dispute. Such a stalemate occurs because the remaining types are both (1) unwilling to pay the audience costs required to concede and (2) lack the resolve to start a war. Moreover, a stalemate can occur even when states are penalized for failing to settle the dispute. To the best of my knowledge, this is the first theory to allow for endogenous stalemates as a crisis outcome.[2]

[2] Previous work has shown that audience costs can lock states into fighting, but not perpetual negotiations (Fearon 1994, 1997; Kurizaki 2007; Leventoğlu and Tarar 2009).

The theory in this paper builds on existing war of attrition models that also feature dynamic screening in both international relations and economics. However, my war of attrition model is designed to apply to a diplomatic crisis and so differs from these works in two key respects. First, following Fearon (1994) states have two exit options: concession or war. When states can escalate the dispute, remaining in the war of attrition for a lengthy period of time is perceived as a sign of hesitancy and therefore as irresolution. By contrast, most existing war of attrition models in the international conflict literature are designed to study ongoing interstate wars or civil wars which participants can only end by conceding (Nalebuff and Riley 1985; Slantchev 2003; Langlois and Langlois 2012; Powell 2017). In this case, the costs of remaining in the war of attrition are interpreted as the costs of fighting which screen low-quality types so that remaining in the war of attrition is interpreted as a sign of determination or strength. Second, in my model states that choose to utilize their outside option and go to war impose this outside option on their rival. This differs from wars of attrition in economics, which are typically used to study firm's decision whether to leave a crowded market and switch to a different sector (Fudenberg and Tirole 1986; Takahashi 2015). Since exiting firms do not impose this outside option on competitors, firms with poor outside options do not preemptively concede and the dual screening result I achieve here is not present.

This paper is the first to demonstrate that resolved states can be screened by sunk costs. Within the literature on crisis diplomacy, Fearon's (1994) seminal article on audience costs is closest to this one. Modeling a crisis as a war of attrition, Fearon sought to demonstrate that audiences could commit states to conflict even when he imposed severe assumptions against fighting. To this end, Fearon assumed that states paid no sunk costs for delay, that no state had a positive payoff from fighting, and that crises end in finite time. Though I relax all three of these assumptions, only the first needs to be relaxed for resolved states to be screened by sunk costs. Intuitively if delay is free, resolved will prefer to wait until all types who wish to concede have done so before fighting.[3] Other scholars studying the two-exit war of attrition, where countries can either concede or escalate, have focused on other dynamics, introducing behavioral types or power fluctuations which do not produce the screening results observed here (Özyurt 2014, 2016; Kim 2018).

In the next section, I present the model setup. I then solve for equilibrium, demonstrating that countries' behavior is governed by two different dynamic screening processes. This is followed by a discussion of the results that compares dynamic screening theory and existing theory. Finally, I present two illustrative case studies of international crises that can be explained by dynamic screening theory and in which costly signaling theory falls short.

The Model

I model an international crisis between two countries as a war of attrition in continuous time. The countries seek to attain an indivisible good of value 1. Both countries will remain locked in the crisis until either one of them exercises one of two exit options: conceding or going to war.[4] If a country chooses to concede, then it surrenders the good to its rival and receives a payoff of 0. If either of the two countries chooses to escalate, then the countries fight and the game ends in a costly lottery (Fearon 1995). Let p_i and $1-p_i$ be the probability that country i and j $(j \neq i)$ respectively win the fight and receive the good. Let c_i $(i = 1, 2)$ denote a country's resolve, the cost that each country pays for fighting regardless of the lottery outcome. Each country's cost for fighting is private information and is selected by a random draw from a common knowledge distribution $c_i \sim C_i$ with continuous and strictly

[3] The assumption that resolved states have negative payoffs for fighting need not prevent screening. However, screening becomes less likely - the lower a state's payoff to fighting the more tolerant of delay it will be and if payoffs for fighting are sufficiently low, then it is possible for all unresolved states to concede before either states becomes impatient with diplomacy. The online appendix demonstrates this point by numerically simulating the model. The assumption that wars of attrition occur end in finite time is unnecessary as sunk costs ensure that resolved states will exit endogenously. Moreover, relaxing this assumption allows for a richer analysis that can accomodate the occurence of stalemates.

[4] To avoid confusion between "war of attrition," the class of model, and "going to war," an exit strategy in the model, I will substitute the term "crisis" for "war of attrition" whenever possible.

positive density over its support $[\underline{c}_i, \bar{c}_i]$. To simplify matters, let each country i's payoff to fighting be denoted with $w_i = p_i - c_i$ and the transformed cumulative distribution F_i have a support over $[\underline{w}_i, \overline{w}_i]$. I assume that $\overline{w}_i \in (0,1)$ and that $\underline{w}_i \in (-1,0)$ such that there always exist both types with a positive and negative expected utility for fighting and no type prefers fighting to obtaining a concession.[5]

Following Fearon (1994), each country accumulates audience costs that must paid if that country concedes. Such costs are designed to capture punishments imposed by domestic audiences on leaders who fail to follow through on a threat that they have made. I assume that these costs accrue at a linear rate a_i so that if country i chooses to concede at time t, they pay $a_i t$ audience costs for doing so. This parameterization of audience costs implicitly assumes that countries pay no audience costs for conceding the conflict immediately at time $t = 0$. This reflects a belief that countries which concede "quietly, without a public contest" incur no penalties from their domestic audiences (Fearon 1994, 585). After this point, the model assumes that domestic audiences punish leaders more severely for conceding after lengthier crises.

In addition, each country must also pay a sunk cost k_i for every moment that they remain in the crisis. These sunk costs represent any and all expenses that might arise in a diplomatic dispute that must be paid regardless of the outcome of the conflict, such as the costs of mobilizing troops. I assume that countries pay no sunk costs at the start of the crisis so that a country who exits at $t = 0$ incurs no sunk costs. Together, these assumptions imply that a country that remains the crisis until time t incurs sunk costs $k_i t$.

Since I am interested in studying stalemate outcomes where both countries remain locked in the crisis forever, I must consider what happens if neither country goes to war or concedes. In such a case, I assume that countries cease to pay sunk costs and incur a one-time penalty $K_i > 0$. Substantively, this term represents the costs that arise when countries fail to settle a diplomatic dispute, such as market actors viewing investments in the country as being

[5] This contrasts with Fearon (1994) who assumed that $\overline{w}_i = 0$.

riskier or a permanent increase in troop deployments. Mathematically, the assumption that countries treat the costs of a stalemate as a one-time penalty, as opposed to paying sunk costs in perpetuity, is useful because it bounds payoffs. This is a necessary condition for stalemates. The online appendix explores an alternative solution in which states discount future payoffs and pay sunk costs in perpetuity and achieves similar results.

A strategy for country i is a function mapping country i's type to its choice of exit time and choice of exit strategy. The former will be denoted with a choice t_i in the set $\mathbb{R}_+ \cup \{\infty\}$ where a choice of $t_i = \infty$ represents a choice not to exit. Exit options will be denoted with $\theta_i \in \{0,1\}$ where $\theta_i = 0$ represents a choice to concede and $\theta_i = 1$ represents a choice to go to war. Formally country i's strategy is defined as $\sigma_i : [\underline{w}_i, \overline{w}_i] \rightarrow \mathbb{R}_+ \cup \{\infty\} \times \{0,1\}$. It will often be useful to work with the inverse image of the strategy function that maps exit times and exit choice to a country's type $\tau_i : \mathbb{R}_+ \cup \{\infty\} \times \{0,1\} \rightarrow [\underline{w}_i, \overline{w}_i]$. Abusing notation, let $\sigma_i(w_i|\theta)$ denote the strategy function mapping type to exit time for types choosing exit option θ. Given these strategies, a simplified version of country i's expected utility function can be written as

$$\begin{aligned} U_i(t_i, \theta_i, \sigma_j|w_i) = &\int_{\{w_j|t_j<t_i,\theta_j=0\}} [f_j(w)(1 - k_i\sigma(w|0))]dw + \int_{\{w_j|t_j<t_i,\theta_j=1\}} [f_j(w)(w_i - k_i\sigma(w|1))]dw \\ &+\mathbb{1}_{\{t_i\neq\infty,\theta_i=1\}} \int_{\{w_j|(t_j=t_i,\theta=1)\vee(t_j>t_i)\}} [f(w)\,(w_i - k_it_i)]\,dw \\ &-\mathbb{1}_{\{t_i\neq\infty,\theta_i=0\}} \int_{\{w_j|t_j>t_i)\}} [f_j(w)(a_it_i + k_it_i)]dw - \mathbb{1}_{\{t_i=\infty\}} \int_{\{w_j|t_j=\infty)\}} [f(w)(K_i + k_i\overline{T})]dw \end{aligned} \tag{1}$$

where the first line represents the payoff if country j exits before country i, the second line represents country i's payoff from going to war at time t_i, the first term on the third line represents country i's payoff from conceding at time t_i, and the last term is the payoff for remaining locked in the crisis forever.[6]

Throughout the paper I solve for Perfect Bayesian Equilibria. This requires that each

[6] A full statement of the expected utility function is provided in the online appendix.

country update its beliefs using Bayes' Rule whenever possible and maximize their expected utility in light of these beliefs. Throughout the crisis, each country will be able to update its beliefs at every instant after learning that its opponent has not yet chosen to exit. Let $g_i(w_j|t)$ denote country i's posterior beliefs that country j has wartime payoff w_j after observing country j remain in the crisis up until time t. Each country must continue to find its choice of exit time and strategy optimal as it updates these posterior beliefs. An equilibrium is therefore a pair (σ_i^*, g_i) for each country.[7]

Characterizing the Equilibrium

In this section I demonstrate that the game has an equilibrium featuring three distinct phases during which countries play different strategies. I characterize behavior in each phase and define the conditions which must be met for the countries to transition between them. These phases must occur in a strict sequence for virtually any possible set of parameters. The game always begins with a peaceful phase during which no country goes to war. If countries have types that are sufficiently resolved, then sunk costs cause the game to transition to one of two different screening phases. First, a screening phase where only one country gradually goes to war and then a second screening phase where both countries gradually go to war. Unresolved types will concede throughout the three phases, though the strategy by which they do so will change from phase to phase. All equilibria must take this sequential form.[8]

The game may end during any one of the three phases. Specifically, the game ends when at least one country no longer has any types remaining who wish to concede. The assumption that states accumulate a strictly increasing quantity of audience costs ensures that such a date must exist. Intuitively, audience costs must eventually grow so large such that there is a date by which no type would ever choose to concede. Following Fearon (1994), I refer to this time as the horizon date and label it $\overline{T}$. Once a country no longer has any

[7] A formal definition of a Perfect Bayesian Equilibrium is provided in the online appendix.

[8] The online appendix contains a formal proof for the argument that the war of attrition must have delay occur with positive probability and that the sequence of phases must be unique.

types remaining who wish to concede, its rival can no longer justify paying the sunk costs required to delay exiting, thereby triggering an end to strategic behavior. This leads us to the following Lemma (all proofs provided in the online appendix):

Lemma 1

In any equilibrium there exists a finite time $\overline{T}$ after which no type exits the crisis.

Strategic behavior can end in one of three ways. First, both countries may run out of types willing to concede, and all remaining types can go to war. Second, under certain conditions, only one country may run out of types willing to concede. In response its rival immediately exits, either by going to war or concede at time $\overline{T}$. Due to space constraints, I restrict attention in the main text to the first of these two possibilities. A full characterization of the equilibria that accounts for this second possibility is provided in the online appendix.

Third, it is possible for the countries to remain in the crisis forever. This requires that $K_i < a_i(\overline{T})$ so that there exist types who would prefer to pay the sunk costs required to sustain the crisis forever rather than concede. I will refer to such an outcome as a stalemate because the dispute remains unsettled; neither party is sufficiently resolved to go to war and audience costs prevent both countries from conceding. Stalemates require that both countries have their unresolved types finish conceding by $\overline{T}$. In turn, any resolved types still participating in the crisis to go to war at $\overline{T}$ and the only types remaining past $\overline{T}$ are those who prefer to remain in the crisis forever. In the remainder of this section, I describe countries' behavior during each of the three phases in sequence and then at the horizon date.

Characterizing the Peaceful Phase

The game begins with a peaceful phase during which neither country goes to war. Since war is costly, resolved types prefer that their rival concede peacefully and will start by granting them the opportunity to do so. Knowing that their rival might abruptly choose to declare war at a later time, the least resolved types will choose to concede during the peaceful phase. However, these least resolved types can benefit from delaying their concession if their rival is similarly unresolved and concedes first. As a result, types who concede during the peaceful phase will do so via a mixed strategy, delaying their concession for a random length of time in the hope of outlasting the rival state. Such delay is costly and both countries accumulate sunk costs and audience costs while waiting for their rival to concede. Though resolved types would prefer their rival concede peacefully, their willingness to pay sunk costs while waiting for their rival to concede is limited. When the most resolved type of both countries is no longer willing to incur sunk costs, they will go to war and cause the game to transition to the first screening phase. Figure 1 illustrates these strategies.

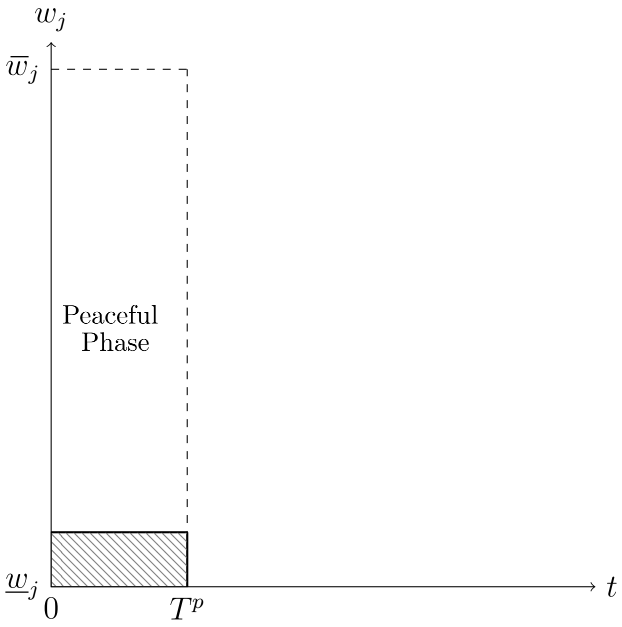

Figure 1: **The Peaceful Phase:** This figure maps the time at which different types exit the crisis, with the y-axis measuring the possible war payoffs and the x-axis measuring the time at which they exit. Crises will always begin with a peaceful phase during which no type goes to war. This is represented by the dashed line at the top of the figure indicating that type $\overline{w}_j$ does not exit. Knowing that war may occur in the future, the least resolved types of each country will concede during the peaceful phase by playing a mixed strategy. This is represented by the patterned box, indicating that the least resolved types of country j concede at a random time in that interval.

Concession Behavior During the Peaceful Phase

I begin by characterizing the strategies for types that concede during the peaceful phase. At the start of the dispute one of the two countries may concede with positive probability.

If that country does not concede, then a crisis begins and the countries start to accrue sunk costs and audience costs. From that point on either one of the two countries may concede at any time. Though the probability of a concession at any *particular* time t is vanishingly small, the probability that either country has conceded *by* a particular time t is strictly increasing and does not cease to increase until the end of the peaceful phase.

Formally, let $Q_i(t)$ be the cumulative distribution function describing the probability that types of country i who have chosen to concede during the peaceful phase do so by time t. Moreover, let T^p denote the date at which the peaceful phase transitions into a screening phase. Lemma 2 establishes some useful properties of $Q_i(t)$.

Lemma 2

Let $T^1 = \min\{T^p, \bar{T}\}$. $Q_i(t)$ must satisfy the following properties in an equilibrium in which both countries finish conceding by the horizon date: (i) $Q_i(t)$ must be continuous and strictly increasing; (ii) $Q_i(t) < 1$ if and only if $t < \min T^1$; and (iii) $Q_i(0)Q_j(0) = 0$.

The following Proposition characterizes $Q_i(t)$ explicitly.

Proposition 1

Let $T^1 = \min\{T^p, \bar{T}\}$. In any equilibrium in which both countries finish conceding by the horizon date, types $w_i \in [\underline{w}_i, \beta_i^p]$ $(i = 1, 2)$ concede on the interval $[0, T^1]$ according to the following strategy

$$\frac{q_j(t)F_j(\beta_j^p)}{1 - F_j(\beta_j^p)Q_j(t)} = \frac{a_i + k_i}{1 + a_i t} \tag{2}$$

Since no type goes to war during the peaceful phase, both countries beliefs are given by

$$g_i(w_j|t) = \begin{cases} \frac{f_j(w_j)[1-Q_j(t_i)]}{1-Q_j(t)F_j(\beta_j^p)} & \text{if } w_j \in [\underline{w}_j^t, \beta_j^p] \\ \frac{f_j(w_j)}{1-Q_j(t)F_j(\beta_j^p)} & \text{if } w_j \in [\beta_j^p, \overline{w}_j] \end{cases} \tag{3}$$

The intuition underlying Proposition 1 is straightforward. Since country i is mixing, it has to be indifferent as to when it concedes during the peaceful phase. This requires that the marginal benefits of delaying concession at any given moment are equal to the marginal costs of doing so. Proposition 1 characterizes these quantities. The left-hand side of equation (2) is country j's hazard rate, representing the probability that country j will concede if country i decides to wait a moment longer. The right-hand side of equation (2) represents the weighted marginal costs to waiting, i.e. the additional sunk and audience costs that would have to paid for conceding later divided by the difference in payoffs to having country j concede as opposed to country i conceding. Together Lemma 1 and equation (2) form a comprehensive strategy for types conceding during the peaceful stage. Finally, over the course of the peaceful phase each country continuously reduces their belief that their rival comes from a subset of the types with the lowest payoffs to fighting - because these types concede with positive probability over the course of the the peaceful, the longer a state holds out the less likely they are to be unresolved.

How Long will the Peaceful Phase Last?

The peaceful phase ends whenever there is a type that is no longer willing to delay going to war. Since types $\overline{w}_i$ $(i = 1, 2)$ have the highest payoffs for fighting, they will be the types of country i and country j that will prefer to go to war earliest. However, types $\overline{w}_i$ and $\overline{w}_j$ may prefer to go to war at different times. The peaceful phase will end whenever the first of these two types decides to go to war or upon arrival at the horizon date, whichever comes first.

Proposition 2 provides a formal characterization for T^p, the length of the peaceful phase that transitions into a first screening phase. Let T_i^p $(i = 1, 2)$ denote the amount of time that type $\overline{w}_i$ is willing to wait before going to war if its rival will play according to equation (2) up until that time. The following proposition provides a formal characterization for these quantities.

Proposition 2

During the peaceful phase, type $\overline{w}_i$ $(i = 1, 2)$ will choose to go to war at time

$$T_i^p = [1 - \overline{w}_i]\left[\frac{1}{k_i} + \frac{1}{a_i}\right] - \frac{1}{a_i} \tag{4}$$

Let $T^p = \min\{T_1^p, T_2^p\}$. If $T^p < \overline{T}$, then at time T^p the game transitions to the first screening phase and type $\overline{w}_i$ goes to war where i is the country for which $T_i^p = T^p$. Otherwise, the game proceeds directly to the horizon date at $\overline{T}$.

The following is the intuition underlying Proposition 2. Country i will seek to wait before going to war until the marginal benefit of doing so is equal to the marginal cost. This will be achieved whenever

$$\frac{F_j(\beta_j^p)q_j(t_i)}{1 - Q_j(t_i)F_j(\beta_j^p)} = \frac{k_i}{1 - \overline{w}_i} \tag{5}$$

The left-hand side of this equation is the probability that country j concedes if country i $(i \neq j)$ at time t and represents the marginal benefit of delaying the choice to go to war at that time. The right-hand side of the equation is the marginal cost, representing the sunk costs that are paid for delaying another moment weighted by the difference in payoffs between having country j concede and having country i go to war. Equation (4) is found by substituting in for the hazard rate (2) into the left-hand side of equation (5).

Equation (4) also reveals three additional facts about the peaceful phase. First, T_i^p is strictly decreasing in $\overline{w}_i$, implying that as the upper bound of a countries' resolve increases, the peaceful phase becomes shorter. Second, it is each country's most resolved type $\overline{w}_i$ that will satisfy the equation at the earliest time. Third, the equation demonstrates that it is the presence of audience costs and the fact that they are strictly increasing that generates the peaceful phase.[9] This is because audience costs make delay more costly for unresolved types thereby requiring that their rival concede at a faster rate to keep the unresolved types indifferent as to when they concede. It is this accelerated rate of concession which makes delaying war worthwhile for resolved types.

[9] To see this multiply both sides of equation (4) by a_i and then set $a_i = 0$.

Characterizing the First Screening Phase

Once type $\overline{w}_i$ or $\overline{w}_j$ chooses to go to war, the peaceful phase ends and the game transitions into the first screening phase.[10] Without loss of generality, let country j be the country who's most resolved type preferred to go to war first ($T_j^p < T_i^p$). During this phase, sunk costs screen resolved types of country j - at any given time, country j's most resolved type still participating in crisis the crisis will go to war because it is no longer willing to pay sunk costs. In turn, the threat of war screens country i's unresolved types - at any given time, country i's least resolved type still participating in the crisis will concede because it wants to avoid the risk that its rival will abruptly declare war. However, resolved types of country i are still willing to give their rival an opportunity to concede peacefully and do not go to war in this phase. Without the threat of war from their opponent, unresolved types of country j continue to concede by playing a mixed strategy. When the most resolved type of country i is willing to incur sunk costs no longer, they will escalate and cause the game to transition to the second screening phase. Figure 2 illustrates these equilibrium strategies.

The Switch to Pure Strategies

I begin by characterizing the properties which must be satisfied by the various exit strategies. Once again, though the probability that country j goes to war, or that country i or j concedes, *at* any particular moment is vanishingly small, the probability that they do so *by* a particular moment is strictly and continuously increasing right up until the very end of the first screening phase. Consequently, resolved types of country j can choose precisely how much sunk costs they are willing to pay in exchange for a probability of a concession by delaying war. Resolved types with higher costs of fighting have more to lose from going to war and will be willing to remain in the crisis longer. Similarly, unresolved types of country i can choose exactly how much risk they are willing to take by remaining in the war of

[10] If the most resolved types of both countries both want to transition from the peaceful phase at at the same time ($T_i^p = T_j^p, i \neq j$), then the game transitions directly to the second screening phase. However, this is highly unlikely given the multidimensional and continuous parameter space.

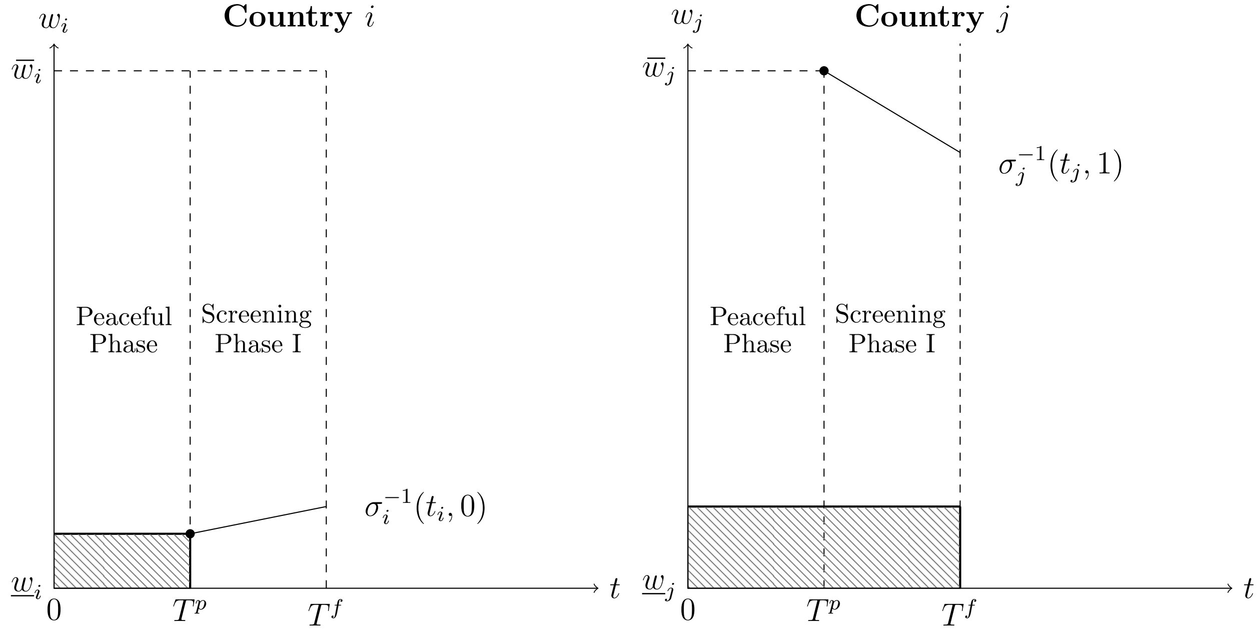

Figure 2: **The First Screening Phase:** The peaceful phase is followed by the first screening phase during which only one country has types that go to war. Beginning at time T^p, sunk costs begin to screen resolved types of country j. This is represented by the upper curve in the right sub-figure depicting the time at which resolved types of country j choose to go to war. In turn, the threat of war screens unresolved types of country i who switch from playing a mixed strategy to a pure strategy. This is represented by the lower curve in the left sub-figure depicting the time at which unresolved types of country i concede. Resolved types of country i remain peaceful during the first screening phase as depicted by the extension of the dashed line at the top of the left sub-figure into the first screening phase. Absent the threat of war, unresolved types of country j continue to concede via mixed strategy as depicted by a similar extension of the patterned box at the bottom of the right sub-figure.

attrition by choosing to delay until a particular time. Unresolved types of country i with higher payoffs to fighting will be willing to incur a little more risk and therefore delay longer.

Formally, let $S_j(t)$ denote the mixed strategy adopted by conceding types of country j. Let T^f, defined more precisely below, denote the date at which the first screening phase transitions into the second screening phase. Lemma 3 lays out the key characteristics that govern the dual screening process.

Lemma 3

Let $T^2 = \min\{T^f, \overline{T}\}$. If there exists a $T^p < \overline{T}$, then (i) both $S_j(t)$ and $\sigma_i(\cdot|0)$ must be continuous and strictly increasing on $[T^p, T^2]$(ii); $S_j(t) < 1$ if and only if $t < \min T^2$; (iii)

$S_j(T^p) = 0$; (iv) $\sigma_j(\cdot|1)$ must be continuous and strictly decreasing on $[T^p, T^2]$.

Proposition 3 characterizes the strategies of unresolved types of countries i and j and resolved types of country j that choose to exit during the first screening phase. Let β_i^f and β_j^f denote the lowest cost-of-fighting type of each country to concede during the first screening phase. Let $\underline{w}_i^t$ denote the least resolved type of country i yet to concede at time t. Proposition 3 provides a formal characterization of exit strategies for types exiting during the first screening phase.

Proposition 3

Let $T^2 = \min\{T^f, \overline{T}\}$. If there exists a $T^p < \overline{T}$, then during $[T^p, T^2]$, country i concedes by playing $\tau_i(\cdot, 0)$ as given by

$$\frac{f_i(\tau_i(t,0))\tau_i'(t,0)}{1 - F_i(\tau_i(t,0))} = \frac{a_j + k_j}{1 + a_j t} \tag{6}$$

Types $w_j \in [\beta_j^p, \beta_j^f]$ and resolved types of country j exit by playing

$$\begin{aligned} &\frac{[F_j(\beta_j^f) - F_j(\beta_j^p)]s_j(t)}{F_j(\tau_j(t,1)) - [F_j(\beta_j^f) - F_j(\beta_j^p)]S_j(t) - F_j(\beta_j^p)} = \frac{a_i + k_i}{1 + a_i t} \\ +&\frac{f_j(\tau_j(t,1))\tau_j'(t,1)}{F_j(\tau_j(t,1)) - [F_j(\beta_j^f) - F_j(\beta_j^p)]S_j(t) - F_j(\beta_j^p)} \times \frac{\underline{w}_i^t + a_i t}{1 + a_i t} \end{aligned} \tag{7}$$

$$\sigma_j(w_j|1) = [1 - w_j]\left[\frac{1}{k_j} + \frac{1}{a_j}\right] - \frac{1}{a_j} \tag{8}$$

Each country's posterior beliefs posterior beliefs during this period are given by

$$g_i(w_j|t) = \begin{cases} \frac{f_j(w_j)[1 - S_j(t)]}{F_j(\tau_j(t_i,1)) - [F_j(\beta_j^f) - F_j(\beta_j^p)]S_j(t) - F_j(\beta_j^p)} & \textit{if } w_j \in [\beta_j^p, \beta_j^f] \\ \frac{f_j(w_j)}{F_j(\tau_j(t_i,1)) - [F_j(\beta_j^f) - F_j(\beta_j^p)]S_j(t) - F_j(\beta_j^p)} & \textit{if } w_j \in [\beta_j^f, \overline{w}_j^t] \\ 0 & \textit{otherwise} \end{cases} \tag{9}$$

$$g_j(w_i|t) = \begin{cases} \frac{f_i(w_i)}{1 - F_i(\tau_i(t,0))} & \textit{if } w_i \in [\underline{w}_i^t, \overline{w}_i] \\ 0 & \textit{otherwise} \end{cases} \tag{10}$$

The following is the intuition underlying the result. First, recall that without the threat of war, unresolved types of country j must be indifferent as to when they concede. From the discussion in the previous section, we know that this requires that unresolved types of country i concede at the rate given in equation (2). However, Lemma 3 requires that country i play a pure strategy with unresolved types of country i that have higher costs of fighting conceding earlier. Therefore, (2) is rewritten as (6). Surprisingly, this implies that the threat of war does not change the rate at which country i concedes, only that its unresolved types now do so via a pure strategy. Second, because the rate at which country i concedes does not change when the game transitions to the first screening phase, the trade-offs affecting how long a resolved type of country j should wait before going to war are identical to those faced by type $\overline{w}_i$ in the previous section. Therefore, equation (8) is analogous to (4).

Finally, because unresolved types of country j that concede during the first screening phase are playing a mixed strategy and are indifferent as to when they concede, their strategy only needs to ensure that unresolved types of country i find $\tau_i(t_i, 0)$ as given in equation (6) optimal. Unresolved types of country i will choose to delay their concession until the marginal costs of doing so are equal to the marginal benefits. The proof of the proposition shows that Country i's expected utility function is maximized by waiting until (7) is satisfied. Equation (7) is similar to equation (2) with the addition of a new term, an additional marginal cost to waiting that accounts for the possibility that delay will lead to war. Perhaps contrary to intuition, this new term is positive, implying that country j's rate of concession increases when compared to the peaceful phase. This increase is necessary to compensate country i for the increased risk of war now involved in delay.

These strategies determine how countries posterior beliefs change over the course of the first screening phase. First, country i continuously lowers its belief for country j's highest possible level of resolve. Because the most resolved types of country j are screened by sunk costs, i can eliminate the possibility of j being a type that should have already gone to war. Second, country j can similarly continuously increase its belief regarding the lowest possible level of country i's resolve - the fact that country i has not exited demonstrates that it has not yet been screened by the risk of war. Finally, as in the peaceful phase, country i continuously reduces its belief that country j is from the subset of the least resolved types as the crisis continues and country j does not concede.

How Long Will the First Screening Phase Last?

Eventually, resolved types of country i will grow tired of paying sunk costs while waiting for country j to concede. As before, type $\overline{w}_i$ will be the type who wants go to war at the earliest date because it has the lowest cost of fighting. The time at which type $\overline{w}_i$ goes to war is T^f and it marks the time at which the game transitions to the second screening phase.

The following Proposition characterizes the length of first screening phase.

Proposition 4

Type $\overline{w}_i$ will go to war whenever the following condition is satisfied

$$\overline{w}_i = 1 - \frac{k_i[1 + a_i t]}{k_i + a_i + [\underline{w}_i^t + a_i t]\frac{f_j(\tau_j(t_i,1))\tau_j'(t_i,1)}{F_j(\tau_j(t_i,1)) - (F_j(\beta_j^f) - F_j(\beta_j^p))S_j(t) - F_j(\beta_j^p)}} \tag{11}$$

at which point the game transitions to the second phase.

The intuition underlying Proposition 4 is similar to that underlying Proposition 2. Type $\overline{w}_i$ will delay going to war until the marginal costs of doing so are equal to the marginal benefits. This expected utility function will be maximized at time T^f when equation (11) holds. When equation (4) is rearranged to isolate $\overline{w}_i$, it produces an identical expression to that in equation (11) without the right-most term in the denominator. This new term reflects the fact that resolved types are willing to wait longer before going to war if there is some chance that their rival is going to initiate a war anyway. Therefore, type $\overline{w}_i$ delays their exit time relative to when they would have exited if it had been the case that $T_i^p < T_j^p$ and country i would have initiated the transition to the first screening phase. The date at which equation (11) holds is the date at which the game transitions from the first to second screening phases, providing us with a formal definition of T^f.

Characterizing the Second Screening Phase

Once the most resolved type of country i decides to go to war, the game transitions to the second screening phase.[11] In this phase, sunk costs screen both countries resolved types, causing the most resolved type of each country still participating in the crisis at any particular moment to go to war to avoid continuing to pay sunk costs. This generates a risk of war that screens both countries' unresolved types, causing the least resolved type of each country still participating in the crisis at any particular moment to concede. Both of these screening processes continue until the horizon date, when all types who intended to concede have done so. Figure 3 illustrates these results.

As in previous phases, screening involves strategies wherein countries gradually exit the crisis. Specifically, countries play strategies such that the timing of a war or concession is effectively zero at any particular moment, but that the probability of a war or a concession is strictly and continuously increasing right up until the horizon date. The following lemma establishes this result and dynamic screening's monotonicty properties:

Lemma 4

If there exists a $T^f < \overline{T}$, then during $[T^f, \overline{T}]$ $\sigma_i(\cdot|0)$ is continuous and strictly increasing and $\sigma_i(\cdot|1)$ is continuous and strictly decreasing.

The following Proposition characterizes the strategies for those types exiting during the second screening phase.

Proposition 5

If there exists a $T^f < \overline{T}$, then types of country i who exit during $[T^f, \overline{T}]$ play according to

[11] Or alternatively, if $T_i^p = T_j^p$ the game proceeds directly to this phase instead of the first screening phase.

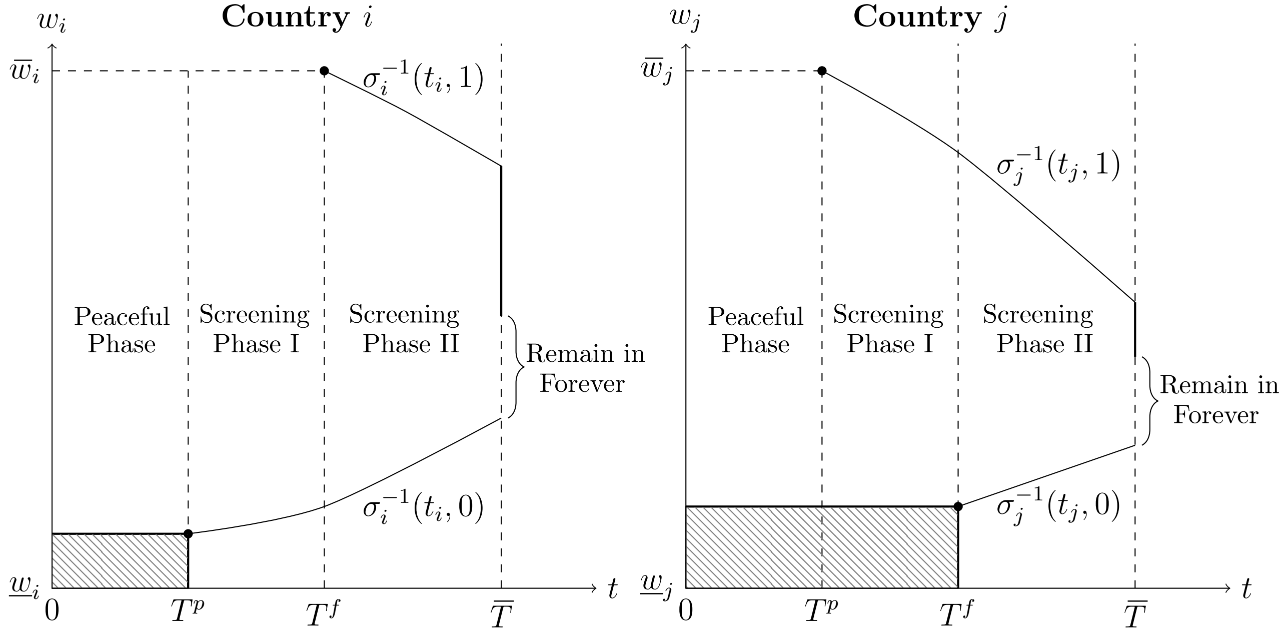

Figure 3: **The Second Screening Phase and the Horizon Date:** At time T^f sunk costs begin screening resolved types of country i causing them to gradually go to war. This is depicted by the upper curve in the left sub-figure. In turn, the risk of war that this generates begins to screen unresolved types of country j as depicted by the lower curve in the right-sub figure. Resolved types of country j and unresolved types of country i continue to be screened as they were in the first screening phase. Both countries continue being screened until the horizon date $\overline{T}$ when the last unresolved type of both countries concedes. At this point remaining types either go to war immediately or remain in crisis forever. Though the figure depicts the horizon date at the end of the second screening phase, it is possible for it to arrive earlier such that countries never arrive at either the first or second screening phases.

the strategy $\tau_i(\cdot,\theta)$ $(i=1,2)$ as defined by

$$\frac{f_j(\tau_j(t,0))\tau_j'(t,0)}{F_j(\tau_j(t,1))-F_j(\tau_j(t,0))} = \frac{k_i}{1-\overline{w}_i^t} \tag{12}$$

$$\frac{f_j(\tau_j(t,1))\tau_j'(t,1)}{F_j(\tau_j(t,1))-F_j(\tau_j(t,0))} = \frac{k_i[\overline{w}_i^t+a_it]-a_i[1-\overline{w}_i^t]}{[1-\overline{w}_i^t][\underline{w}_i^t+a_it]} \tag{13}$$

and country i's $(i=1,2)$ posterior beliefs during $[T^f,\overline{T}]$ are given by

$$g_i(w_j|t) = \begin{cases} \frac{f_j(w_j)}{F_j(\tau_j(t,1))-F_j(\tau_j(t,0))} & \textit{if } w_j \in [\underline{w}_j^t, \overline{w}_j^t] \\ 0 & \textit{otherwise} \end{cases} \tag{14}$$

The intuition for this result is similar to that of Proposition 3. Resolved and unresolved types of both countries will remain in the war of attrition until the marginal costs of doing exceed the marginal costs. For resolved types, this means delaying until equation (12) is satisfied which is analogous in interpretation to equation (5). For unresolved types, this means delaying concession until equation (13) is satisfied.

The strategies stated in the proposition also determine how countries will update their posterior beliefs during the second screening phase. Now that unresolved and resolved types of each country are being screened, both countries will continuously increase their belief regarding their rival's lowest possible level of resolve and continuously decrease their belief regarding their rival's highest possible resolve. As before, this is because a country that remains in the crisis demonstrates that it is sufficiently resolved to have incurred the risk of war so far, but also lacks the resolve to have yet initiated a war.

Characterizing the Horizon Date

The game ends once all unresolved types have conceded. At this point, any type of either country that intends to go to war can no longer benefit from delay. As a result, a mass of types, possibly all those who remain, will go to war at the horizon date. However, it is also possible that some types still participating in the crisis may opt for a stalemate outcome and sustain the crisis in perpetuity. This latter outcome is illustrated in Figure 3.

Let β_i $(i = 1, 2)$ denote the lowest cost-of-fighting type to concede of country i. The following proposition summarizes countries' behaviour throughout the crisis and at the horizon date.

Proposition 6

Strategic behaviour ends at $\overline{T}$, which can arrive during any phase. In an equilibrium where both countries finish conceding by the horizon date, countries' choice of exit strategy and their behavior at the horizon date is determined by the following:

(i) All types exit: There exists an equilibrium where types $w_i \in [\underline{w}_i, \beta_i]$ concede and types

$w_i \in (\beta_i, \overline{w}_i]$ go to war where $\beta_i = -a_i\overline{T}$. Any type still participating in the crisis at $\overline{T}$ goes to war at that time.

(ii) Some types remain in forever: If $K_i < a_i\overline{T}$ for both $i = 1, 2$, then there exists an equilibrium where types $w_i \in [\underline{w}_i, \beta_i]$ concede for β_i as given by

$$\frac{F_j(\overline{w}_j^{\overline{T}}) - F_j(-\overline{K}_j)}{F_j(\overline{w}_j^{\overline{T}}) - F_j(\beta_j)}\beta_i - \frac{F_j(-\overline{K}_j) - F_j(\beta_j)}{F_j(\overline{w}_j^{\overline{T}}) - F_j(\beta_j)}K_i = -a_i\overline{T} \tag{15}$$

Types $w_i \in (\beta_i, -K_i]$ remain in the crisis forever and types $w_i \in (-K_i, \overline{w}_i]$ go to war. Any type from the latter set still participating in the crisis at $\overline{T}$ go to war at that time.

The explanation for the result is as follows. If all types remaining at $\overline{T}$ choose to go to war, then type β_i must be indifferent between going to war or conceding and paying the audience costs required to concede at the horizon date. It follows that any type with a higher cost of fighting than type β_i must have already conceded by the horizon date and that any type with a lower cost of fighting than β_i will go to war at $\overline{T}$ if it has not already done so. Alternatively, the game can end in a stalemate whenever sunk costs are sufficiently low $K_i < a_i\overline{T}$ for both $i = 1, 2$. In this case type $w_i = -K_i$ is indifferent between fighting and sustaining the crisis forever. This implies that any type with a higher payoff to fighting than $-K_i$ strictly prefers to fight at $\overline{T}$ if they have not already exited the war of attrition. Type β_i is then defined by equation (15) as the type that is indifferent between paying the audience costs accumulated by the horizon date and opting for a stalemate while risking war with all the types that choose to fight at the horizon date. It follows that any type less resolved than β_i must have already conceded by the horizon date and that any type in $(\beta_i, -K_i]$ will choose to remain in the crisis forever.

Note that Proposition 6 implies that the costs of a stalemate being sufficiently low ($K_i < a_i\overline{T}$ for both $i = 1, 2$) are a necessary but insufficient condition for a stalemate outcome. Stalemates require that both countries choose not to fight at $\overline{T}$. If both countries' strategies require them to fight at the horizon date, then no single country has the ability to prevent a war by unilateral deviation from $\sigma(w_i) = \{\overline{T}, 1\}$ to $\sigma(w_i) = \{\infty, \theta\}$. [12]

Discussion

The war of attrition model offers a parsimonious and widely applicable framework with which to study the dynamics of diplomacy. True to the anarchic nature of the international system, the model imposes little structure on the countries' interactions. The crisis has no exogenously imposed end date. Nor can states commit to taking any future action. Instead, states are free to go to war or concede at any time. Moreover, the model is capable of incorporating standard aspects from the costly signaling literature, including sunk costs and audience costs, into a continuous dynamic setting.

By contrast, costly signaling models are defined by a sender who can take a costly action to attempt to convey information to a receiver. For analytical convenience, signaling models have a discrete order of moves imposed on the players and typically only afford the sender one opportunity to take a costly action. This is a weakness of costly signaling models as it assumes that countries are able to instantaneously take large costly actions.[13] Canonical costly signaling models have demonstrated that countries can communicate that they value an issue highly by taking costly actions (e.g. Fearon 1997; Slantchev 2005).

However, whether a high quality challenger will be able to communicate its willingness to fight depends on the source of uncertainty and type of signal considered (Arena 2013; Caroll and Pond 2021; Reich 2023). For example, the online appendix demonstrates that even if one of the two countries were offered the opportunity to engage in sunk cost signaling prior to the war of attrition beginning, then resolved types would fail to distinguish themselves and a war of attrition with positive probability for delay must still occur. This is because

[12] This is why countries went to war at the horizon date in Fearon's (1994) model even though no type had a positive expected utility for fighting ($\overline{w}_i = 0$). However, without sunk costs ($K_i = k_i = 0$), Fearon's model could also support an outcome where both countries remain in the war of attrition forever. The potential for a stalemate is not recognized or discussed in the article.

[13] Alternatively, countries could produce signals over a longer period during which any of the receiver's interim actions are unimpactful so that the signal may as well have been instantaneously produced.

resolved types' preference to invest less in sunk costs is not driven by the game form, but by their weaker incentive to spend to avoid war. However, this is not to say that countries cannot communicate with costly signaling under the right assumptions. Therefore, it is worth highlighting four general results from the model and how they compare to existing theory.[14]

The first general result is that more resolved states prefer to invest less in diplomacy and go to war earlier. Specifically, the model shows that more resolved states spend less time negotiating, pay less sunk costs, accumulate less audience costs, and are more likely to go to war. These findings represent a a major departure from costly signaling models whose results suggest that resolved states should invest more in sunk costs and audience costs and will be more likely to achieve peaceful outcomes (Fearon 1994, Fearon 1997; Slantchev 2005; Reich 2022).

The second general result is that unresolved states with higher costs of fighting will choose to concede earlier to avoid the risk of war. Specifically, the model shows that unresolved types with higher costs of fighting spend less time negotiationg, pay less sunk costs, and accumulate less audience costs, are less likely to obtain concessions themselves, and are less likely to go to war.[15] At first glance, this result would seem to support findings in the costly signaling literature, wherein less resolved or weaker types invest less in costly signaling. However, the mechanism that drives this behavior differs across the two theories. In costly signaling models unresolved states don't signal because they are deterred by the price of sunk costs and audience costs, even when these lower the risk of war. By contrast, in dynamic screening theory unresolved states concede early to avoid the risk of war, not because they are deterred by paying sunk costs or audience costs.[16]

[14] For brevity's sake, the introduction treats the first two as a single result.

[15] Note that this relationship is not completely monotonic because Lemmas 2 and 3 require that countries play a mixed strategy absent the threat of war. Conditional on conceding during the peaceful phase or the first screening phase, there need not be a relationship between a country's cost of fighting and its exit time.

[16] Dynamic screening theory more closely resembles the literature on brinkmanship, which argues that states

Taken together, these two general results imply that moderately resolved states spend the most time negotiating. This comes across clearly in Figure 3. Accordingly, moderately resolved states pay more sunk costs, accumulate more audience costs, and are more likely to obtain concessions. This increased probability of obtaining a concession occurs because they grant their rival more time to concede not because states come to believe their rival is more resolved as they invest more in sunk costs and audience costs.

Indeed, the third general result concerns the gradual convergence in each countries posterior beliefs towards their being a moderately resolved type. Once sunk costs begin to screen resolved types, countries conclude that their rival must lack sufficient resolve to have yet started a war and respond by decreasing their belief in the highest level of their rival's possible resolve. Similarly, each country concludes that their rival is at least sufficiently resolved to have not yet conceded, and responds with a continuous reduction in its belief that its rival is unresolved.[17]

Additionally, this slow and steady rate of learning also implies that countries use the length of delay – how long their rival has chosen to prolong the crisis - as their primary metric for assessing resolve. Formally, the propositions demonstrate that the length of delay is a sufficient statistic for a state's posterior distribution of their rival's possible resolve. This means that, conditional on knowing a rival's strategy, the length of delay contains all the information contained within the model necessary for a state to form beliefs about its rival - once a state knows how long its rival has negotiated, it can infer the amount of sunk costs

can demonstrate resolve by generating an *exogenous* risk of war (Schelling 1960). Formal models of brinkmanship showed that when remaining in a crisis required states to withstand the probabilistic risk of war, states with higher costs of fighting concede earlier (Powell 1988). This is similar to the method by which unresolved states are screened in my model. However, brinkmanship views remaining in a crisis and incurring risk as the primary method by which states demonstrate resolve. By contrast, in my model the risk of war is generated endogenously by resolved states who are tired of diplomacy and prefer to fight.

[17] The sole exception is at time $t = 0$ when one country can have a mass of types concede. This implies that if a country meets the challenge of a crisis head on and does not concede immediately, there can be a discontinuous change in beliefs.

it has paid, the amount of audience costs it has accumulated, and the risk of war it has incurred such that these quantities provide no additional information.[18] The gradual nature of learning under dynamic screening contrasts sharply with costly signaling models in which large discontinuous shifts in beliefs are thought to occur following countries undertaking dramatic actions, e.g. following a sudden deployment of troops.

A fourth general result is the model's ability to incorporate stalemates as an endogenous outcome of crises. After the horizon date, countries recongize that they will be locked in crisis in perpetuity as it becomes common knowledge that neither country is sufficiently resolved to start a war and that the crisis has gone on long enough for audience costs to have grown too large for either state to concede. When a stalemate occurs neither state receives the good under dispute and both are assumed to continue to pay some penalty for failing to settle the dispute. Proposition 6 demonstrates that so long as this penalty is not too large, then a stalemate is possible.

An examination of the data on militarized interstate disputes reveals that such stalemates are incredibly common (Palmer et al 2015). Fully 1,508 out of 2,185, or 69 percent, of MIDs that are settled peacefully are coded as ending in a stalemate outcome, i.e. as not having "any decisive changes in the pre-dispute status quo and [are] identified when the outcome does not favor either side in the dispute" (MIDs Dispute Coding Manual). Moreover, 1,442 of these disputes are coded as not having any negotiated outcome, so that "none of the pre-conditions that fueled the conflict are resolved nor is there any agreement between the parties that the dispute should be terminated." This suggests that stalemates as defined by the model are the modal crisis outcome.

Dynamic Screening in Practice

Both dynamic screening and costly signaling theories describe how states manage uncertainty in international crises. Because both theories incorporate uncertainty, sunk costs, and

[18] Note that states may use additional sources of information to form beliefs that are not included in the model.

audience costs as essential components of crisis behavior they allow for direct comparisons of their predictions. This section examines the performance of the two theories in two high-profile cases: the First Gulf War and the Iranian Nuclear Crisis. I demonstrate that in both cases the US's decision to delay war and pursue diplomacy detracted from its large investments in sunk costs or audience costs and caused it to be perceived as unresolved. Together these cases illustrate the shortcomings of costly signaling theory and the explanatory power of dynamic screening theory.

Negotiating the JCPOA

When President Obama acceded to the presidency he decided to reach out to Iran and attempt to negotiate a peaceful resolution to the ongoing crisis over its nuclear program. In the fall of 2009, the P5+1 and Iran held a number of summits until negotiations collapsed when Iran rejected the P5+1's offer for a fuel swap.[19] Iran's rejection of this "confidence-building measure" coupled with the secrecy and nature of its nuclear program convinced the P5+1 that Iran was negotiating in bad faith and in June 2010 the UN Security Council (UNSC) sanctioned Iran. In the 15 months that followed, the US and Europe organized and imposed additional sanctions while Iran continued to develop its nuclear program without further negotiation. Starting in April 2012, negotiations resumed and the P5+1 and Iran held several meetings that produced little progress. Ultimately, a surprise change in Iranian leadership in 2013 jump-started diplomacy. Five months after the election of Iranian President Rouhani, the P5+1 and Iran reached the "Joint Action Plan" agreement and managed to avoid war.

Underlying these negotiations was the US threat to use force to prevent Iran from acquiring nuclear weapons should negotiations fail. Per costly signalling theory, a rational observer of the US should have believed the US threat to be credible. President Obama invested a

[19] By this point Iran had collected a sufficient quantity of low-enriched uranium to produce one nuclear weapon. This deal would have required Iran to ship the lion's share of its stockpile out of the country in exchange for fuel pads for the Tehran Research Reactor which produced medical isotopes. This would have, in principle, provided more time for negotiations and reduced the threat of war (Parsi 2012, 114-116).

significant amount of sunk costs in the crisis. His administration spent a great deal of time and political capital on coordinating the passage of sanctions in the UNSC and advocating for sanctions more broadly. According to Deputy National Security Advisor Ben Rhodes, sanctions against Iran were a top priority for every meeting Obama held with a foreign leader in 2011 (Parsi 2017, 120). Furthermore, Obama would have likely faced a great deal of audience costs if he would have conceded to unrestricted or unmonitored Iranian nuclear enrichment. In both public and private Obama promised that he would use military means to prevent Iran from acquiring a nuclear weapon, stating that "as President of the United States, I don't bluff" (Goldberg 2012; Parsi 2012, 77). Congress also repeatedly exerted pressure on the White House by threatening or passing sanctions bills at the height of sensitive negotiations (Parsi 2012, 73, 108-111, 132-133, 157-161; Parsi 2017 148).

However, US allies did not perceive it to be resolved. Specifically, Israel pressured the US to consider military options to terminate Iran's nuclear program and threatened to attack Iran on its own. Though Israel preferred that the US be the one to attack Iran and the US sought to avoid an Israeli strike, US attempts at reassurance failed to convince Israel that the US would attack should negotiations fail.[20] Though Israel did not ultimately strike Iran, this was not because of its belief in US resolve.[21]

Dynamic screening theory can help explain Israel's lack of confidence in US resolve. Though the US managed to orchestrate a tough multilateral sanctions regime it refused

[20] Israel preferred that the US be the one to attack Iran because of its superior capacity to damage Iranian's nuclear program and set back Iran's ability to acquire the bomb for longer (Parsi 2012, 28-30; Parsi 2017, 151; Barak 2018, 433). The US believed that an Israeli strike would undermine diplomacy, weaken the sanctions regime, and could compel the US to go to war anyway (Parsi 2017, 152-154).

[21] In November 2010, Israeli Prime Minister Netanyahu, the Minister of Defense Ehud Barak, and the Minister of Foreign Affairs Avigdor Liberman supported a strike and held a meeting with the heads of Israel's security organizations to order immediate preparations for one (Barak 2018, 426-427; TOI Staff 2012). However, the plan was opposed by the heads of Israel's security organizations who insisted on a vote in the security cabinet, a group of ministers authorized to approve acts of war, where their opposition would prevent Netanyahu from securing a majority for a strike (Netanyahu 2011, 477-488; TOI Staff 2015).

to commit to a timeline for military action, turning down repeated Israeli requests for a deadline for diplomacy.[22] According to Gary Samore, the White House Coordinator for Arms Control and Weapons of Mass Destruction, the US recognized that if it were ever to admit that negotiations had failed, then they would be forced either to attack Iran or to concede to an unrestricted and unmonitored nuclear program (Parsi 2017, 117). As a result, US policy was to sustain diplomacy and sanctions, even when these offered no sensible path towards a resolution of the dispute.[23] According to dynamic screening theory, this delay conveyed hesitation thereby undermining Obama's reassurances that he would be willing to resort to military means to reign in the Iranian nuclear program.

Gulf War

Within days of Iraq's invasion of Kuwait (August 2nd, 1990), the United States mounted a tough response, beginning the process of deploying tens, and eventually hundreds, of thousands of troops to Saudi Arabia to deter further Iraqi aggression. Simultaneously, the US spearheaded an international coalition that passed numerous UNSC Resolutions condemning the invasion, imposing severe economic sanctions, and authorizing the use of force to enforce these sanctions. In an address to the nation, President Bush made clear that the goal was the "immediate, unconditional, and complete withdrawal" of Iraq from Kuwait (Freedman and Karsh 1993, 93). When it became evident that sanctions would not compel Iraq to leave Kuwait, the US successfully advocated for a UNSC Resolution that authorized the use of force if Iraq would not leave of its own volition.

Per costly signalling theory, these actions should have communicated strong American resolve on this issue. Bush put the US's reputation on the line and accrued audience costs by repeatedly stating that Iraq would have to leave Kuwait without any preconditions and that he would use force to compel it to do so if necessary.[24] Moreover, the Bush administration

[22] See for example Parsi (2012, 50-51, 74-78, 165-169), Parsi (2017, 154-156)

[23] For example, in the summer of 2012, the US scheduled additional diplomatic meetings solely to keep negotiations alive and deny Israel the political cover for a strike (Parsi 2017, 148).

[24] In addition to the UNSC resolutions and many private assurances given to world leaders, Bush made many

incurred a great deal of sunk costs in addressing the crisis. Beyond the time and effort invested in organizing an international coalition, the US deployed a massive force to the Gulf. According to Chairman of the Joint Chiefs of Staff Colin Powell, the US Deployment would be large enough to allow the US to “win decisively” and ensure that it would never be “operating in the margins" (Freedman and Karsh 1993, 207-208). Costly signalling theory would therefore predict that Iraq and US allies should have adjusted their beliefs and taken these actions as evidence that the US was willing to use force if its demands were not met.

However key actors in the conflict continued to doubt US resolve. On November 30th, the day after the UNSC authorized the use of force to expel Iraq from Kuwait, Bush announced that, though he was “not hopeful,” he would reach out to Iraq and attempt to go the “extra mile for peace” (Freedman and Karsh 1993, 235). This policy was undertaken to shore up US domestic support for the war by showing that every effort had been made to attain peace. However, upon learning of Bush’s diplomatic initiative, US allies, influential pundits, and even members of the administration began to suspect that the US “didn’t really want to use force” and was “desperately searching for an escape route” (Baker 1995, 346-353). Saudi Ambassador Bandar told National Security Advisor Scowcroft that sending Secretary of State Baker to meet with Iraqi officials would “suggest [to Saddam] you’re chicken” (Freedman and Karsh 1993, 241). Though Bush made clear that Iraq’s unconditional withdrawal from Kuwait was not on the table, onlookers grew concerned that the US would seek a compromise or allow negotiations to extend beyond the deadlines imposed by the UNSC.

Dynamic screening theory can explain why why US allies were so concerned by Bush’s attempt to negotiate with Iraq, interpreting diplomacy and delay as a sign of irresolution. Moreover, it can also explain why the allies disregarded the US’s sunk cost investment in the conflict and the audience costs it accumulated. For example, as early as October, Prime Minister Thatcher was already pressuring Bush to go to war as soon as US forces arrived in sufficient number so as to avoid looking unresolved (Bush and Scowcroft 1999, 385). When

public statements (Bush and Scowcroft 1999, 340-341, 345, 350, 368, 370-371, 388).

informed about the US decision to double the number of troops in the theatre, Thatcher was not impressed by the US commitment as costly signalling theory would predict. Instead she grew concerned about the delay required for the troops to arrive and the potential that the Americans would “wobble" during that time (Freedman and Karsh 1993, 209, 228). Finally, it should be noted that US allies remained wary of US resolve, correctly assessing Bush’s reluctance to go to war. For example, while Saudi Arabia thought it important that Saddam not be allowed to withdraw from Kuwait unpunished and with his army intact, they recognized that Bush would have gladly allowed him to do so (Baker 1995, 352). [25]

Conclusion

In this article, I developed a theory of dynamic screening in international crises and argued that more resolved states should invest less in diplomacy. To do so, I modeled a crisis as a war of attrition in which states decide how long to pursue diplomatic options while incurring sunk costs, audience costs, and potentially risking war. The model is characterized by two different dynamic screening processes. First, more resolved states will go to war earlier, preferring their assuredly high payoff to fighting over paying sunk costs to prolong a crisis in the hope that their rival concedes. Second, the threat of war posed by this potential breakdown in diplomacy causes the least resolved states to concede earlier to avoid the risk of having war thrust upon them. These dynamic screening processes make the length of delay an important source of information about state’s resolve.

This article is but a first step in the study of dynamic screening in international crises. First, this paper focused on the effects of audience costs and sunk costs as these are the two most prevalent signaling costs in the literature. Other potential processes might be capable of reshaping the screening dynamics. For example, it is possible that democratic institutions might constrain leaders and prohibit them from going to war before a threshold

[25] According to Baker, “war was the last thing" Bush wanted. Instead “all [Bush] really wanted was to get Iraq out of Kuwait" and would have refrained from going to war if Iraq withdrew (Baker 1995, 349). Scowcroft shared the belief that an Iraqi withdrawal could have been a disaster (Bush and Scowcroft 1999, 437-438).

of sufficient negotiations has been reached. I leave it to future research to explore such dynamics. Second, there is room to incorporate bargaining into the war of attrition, as in Langlois and Langlois (2012), so as to study its impact on dynamic screening. Recent work in mechanism design and conflict has studied the properties of different bargaining protocols, demonstrating, for example, that ultimatums deliver proposers their best distribution of outcomes (Fey and Ramsay 2011, Fey and Kenkel 2021). In light of these dynamic screening results, the properties of the war of attrition as a bargaining protocol merit more attention. Finally, dynamic screening is the first theory capable of explaining the variation in the length of crises, generating novel comparative statics and predictions for which states are likely to go to war or concede and when. As such it deserves future empirical study.

Acknowledgements

I would like to thank Muhammet Bas, Ekrem Baser, Brendan Cooley, Haonan Dong, Matias Iaryczower, Kathy Ingram, Michael Kistner, Adam Meirowitz, Kris Ramsay, Jacob Shapiro, Jessica Sun, and three anonymous reviewers for their helpful comments. Additional thanks to the attendees of Peace Science 2019, SPSA 2020, APSA 2020, APSA 2021, Online Peace Science Colloqium 2020, and the summer 2020 Virtual Formal Theory Workshop.

References

Arena, Philip (2013). *Costly Signaling, Resolve, and Martial Effectiveness.*

Baker, James Addison (1995). *The Politics of Diplomacy : Revolution, War, and Peace, 1989-1992.* Ed. by Thomas M. DeFrank. New York: G.P. Putnam's Sons.

Barak, Ehud (2018). *My Country, My Life : Fighting for Israel, Searching for Peace.* New York: St. Martin's Press.

Bush, George and Brent Scowcroft (1999). *A World Transformed.* New York: Vintage Books.

Carroll, Robert and Amy Pond (July 2021). "Costly signaling in autocracy". *International Interactions* 47.4, pp. 612–632.

Fearon, James D. (1994). "Domestic Political Audiences and the Escalation of International Disputes". *American Political Science Review* 88.3, pp. 577–592.
— (1995). "Rationalist explanations for war". *International Organization* 49.3, pp. 379–414.

Fearon, James D. (1997). “Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs”. *Journal of Conflict Resolution* 41.1, pp. 68–90.

Fey, Mark and Brenton Kenkel (2021). “Is an Ultimatum the Last Word on Crisis Bargaining?” *The Journal of Politics* 83.1, pp. 87–102.

Fey, Mark and Kristopher W. Ramsay (2011). “Uncertainty and Incentives in Crisis Bargaining: Game-Free Analysis of International Conflict”. *American Journal of Political Science* 55.1, pp. 149–169.

Freedman, Lawrence and Efraim Karsh (1993). *The Gulf Conflict, 1990-1991 : Diplomacy and War in the New World Order*. Princeton, N.J.: Princeton University Press.

Fudenberg, Drew and Jean Tirole (1986). “A Theory of Exit in Duopoly”. *Econometrica* 54.4, pp. 943–960.

Goldberg, Jeffrey (Mar. 2012). “Obama to Iran and Israel: ’As President of the United States, I Don’t Bluff’”. *The Atlantic.*

Kim, Jin Yeub (2018). “Counterthreat of Attack to Deter Aggression”. *Economics Letters* 167, pp. 112–114.

Kurizaki, Shuhei (2007). “Efficient Secrecy: Public versus Private Threats in Crisis Diplomacy”. *American Political Science Review* 101.3, pp. 543–558.

Langlois, Jean-Pierre P. and Catherine C. Langlois (2012). “Does the Principle of Convergence Really Hold? War, Uncertainty and the Failure of Bargaining”. *British Journal of Political Science* 42.3, pp. 511–536.

Nalebuff, Barry and John Riley (1985). “Asymmetric equilibria in the war of attrition”. *Journal of Theoretical Biology* 113.3, pp. 517–527.

Netanyahu, Benjamin (2022). *Bibi: My Story*. Threshold Editions.

Özyurt, Selçuk (2014). “Audience Costs and Reputation in Crisis Bargaining”. *Games and Economic Behavior* 88, pp. 250–259.

— (2016). “Building Reputation in a War of Attrition Game: Hawkish or Dovish Stance?” *The B.E. Journal of Theoretical Economics* 16.2, pp. 797–816.

Palmer, Glenn et al. (2015). “The Mid4 Dataset, 2002–2010: Procedures, Coding Rules and Description”. *Conflict Management and Peace Science* 32.2, pp. 222–242.

Parsi, Trita (2012). *A Single Roll of the Dice : Obama’s Diplomacy with Iran*. New Haven: Yale University Press.

— (2017). *Losing an Enemy : Obama, Iran, and the Triumph of Diplomacy*. New Haven: Yale University Press.

Powell, Robert (1988). “Nuclear Brinkmanship with Two-Sided Incomplete Information”. *American Political Science Review* 82.1, pp. 155–178.

Powell, Robert (2017). “Taking Sides in Wars of Attrition”. *The American Political Science Review* 111.2, pp. 219–236.

Reich, Noam (2022). “Signaling Strength with Handicaps”. *Journal of Conflict Resolution* 66.7-8, pp. 1481–1513.

— (Mar. 2023). “When Can States Signal with Sunk Costs?”

Schelling, Thomas C. (1960). *The Strategy of Conflict.* Cambridge: Harvard University Press.

Slantchev, Branislav L. (2003). “The Principle of Convergence in Wartime Negotiations”. *American Political Science Review* 97.4, pp. 621–632.

— (2005). “Military Coercion in Interstate Crises”. *American Political Science Review* 99.4, pp. 533–547.

Takahashi, Yuya (2015). “Estimating a War of Attrition: The Case of the US Movie Theater Industry”. *American Economic Review* 105.7, pp. 2204–2241.

Tarar, Ahmer and Bahar Leventoğlu (2009). “Public Commitment in Crisis Bargaining”. *International Studies Quarterly* 53.3, pp. 817–839.

TOI, Staff (Nov. 2012). “Security Chiefs Refused Order from PM in 2010 to Prepare Military to Strike Iran within Hours, TV Report Says”. *Times of Israel.*

— (Aug. 2015). “Barak: Netanyahu Wanted to Strike Iran in 2010 and 2011, but Colleagues Blocked Him”. *Times of Israel.*

Biographical Statement: Noam Reich is a postdoctoral associate in the division of social sciences at New York University Abu Dhabi, Abu Dhabi, United Arab Emirates.