A Behavioral Foundation for Audience Costs*

Avidit Acharya† Edoardo Grillo‡

August 2017

Abstract

We give a behavioral foundation for audience costs by augmenting the standard crisis bargaining model with voters who evaluate material outcomes relative to an endogenous reference point. Voters vote to re-elect their leader when their payoff is high, and to replace him when it is low. Backing down after a challenge may be politically costly to a leader because initiating the challenge has the potential to raise voters' expectations about their final payoff, creating the possibility that they suffer a payoff loss from disappointment when their leader backs down. Whether it is costly or beneficial to back down from a threat—and just how costly or beneficial it is—depends on the reference point, which is determined in equilibrium.

Key words: crisis bargaining, audience costs, reference-dependent utility

1 Introduction

Consider the canonical "crisis bargaining" situation in which the leader of a country can challenge a foreign adversary. In this situation, "audience costs" refer to the costs that the leader incurs as a result of his citizens punishing him for backing down from the challenge relative to not making the challenge to begin with. In the seminal works on the topic, Fearon (1994) and Schultz (1999) examine the implications of these costs for crisis bargaining, but they do not provide an explanation for how the costs arise. Subsequent work on the foundations of audience costs has had difficulties in building a

*We thank Jim Fearon, Jack Levy, Kris Ramsay and Ken Schultz for helpful conversations.

†Assistant Professor of Political Science, Stanford University, Encina Hall West, Room 406, Stanford CA 94305-6044 (email: avidit@stanford.edu).

‡Assistant Professor of Economics, Collegio Carlo Alberto, Via Real Collegio, 30, 10024 Moncalieri (Torino), Italy (email: edoardo.grillo@carloalberto.org).

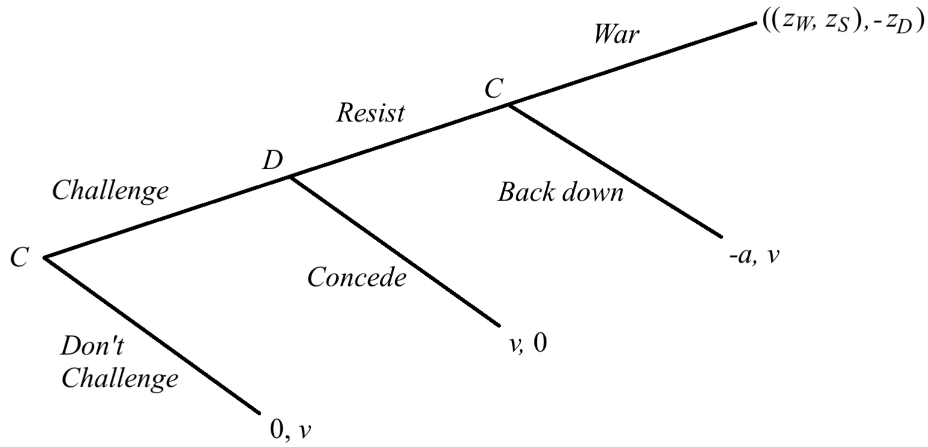

Figure 1

compelling case for these costs under the assumption that citizens are standard expected utility maximizers; we discuss these in Section 4.2 below.

Our objective in this paper is to take a new approach to micro-found audience costs in which we assume that citizens are "behavioral." We augment the standard crisis bargaining model by adding a stage in which voters who care about the outcome of the crisis bargaining game vote to re-elect or replace the incumbent politician. Voters are behavioral in the sense that they have endogenous reference-dependent preferences modeled in the spirit of Kőszegi and Rabin (2006, 2007). Voters decide whether or not to reward the incumbent politician by re-electing him based both on the realized outcome of crisis bargaining and on their expected equilibrium payoff.

The crisis bargaining model that we extend is depicted in Figure 1. In this model, the leaders of two countries, a potential Challenger, C, and a Defender, D, make decisions sequentially. C's leader is one of two types— weak, W, or strong, S—and type is private information, with the prior probability of the strong type denoted $q \in (0, 1)$. C's leader first chooses whether or not to challenge D for a piece of territory that both value at $v > 0$. If C challenges, then the leader of D decides whether to resist or concede the territory. If D resists then C's leader can either escalate to war or back down.

If there is no challenge, then the territory remains with country D, and both types of C's leader get a payoff of 0 while D gets a payoff of v. If the game ends with D conceding, then the territory goes to country C, which results in both types of C's leaders getting a payoff of v and D getting a payoff of 0. If the game ends with war, then the weak leader of C obtains a payoff z_W while the strong type gets z_S and D gets $-z_D$. If the game ends with C's leader backing down, then the payoffs are $-a$ for both types of C's leaders and v for D. The interpretation of this is that the territory remains with country D, and if $a > 0$ then C's leader incurs a cost from backing down in comparison to the payoff from not challenging to begin with. This corresponds to what Fearon (1994) call the "audience cost" (see also Schultz, 1999). Although it is exogenous, Fearon (1994) postulates a foundation in which it results from the possibility that the citizens of C punish their leader for backing down from the initial challenge.

To provide a behavioral foundation for this cost, we set the exogenous audience cost to $a = 0$ and augment the model above with voters who have reference-dependent payoffs in which the reference point is determined endogenously as in Kőszegi and Rabin (2006, 2007). The voters are citizens of C, and are of two types: hawks, whose material payoffs equal the crisis bargaining payoffs of the strong type of leader, and doves, whose material payoffs equal the crisis bargaining payoffs of the weak type of leader. If voters vote to re-elect their politician when their payoff is high, and vote to replace him when their payoff is low, then backing down after a challenge may be costly to the politician, who also values re-election. In particular, if voters are predominantly hawks, then entering a crisis by challenging the territory has the potential to raise the voters' reference points. Backing down could then generate a payoff loss due to "disappointment." On the other hand, citizens cannot be disappointed if the politician does not challenge the territory since this would not raise their expectations in the first place.

Whether or not it is costly to back down, however, and just how costly it is, depends on the value of the reference point and whether voters are predominantly hawks or doves. If sufficiently many voters are sufficiently hawkish about war, then backing down after a challenge generates both a "sunk audience cost" (i.e., a lower payoff after backing down than after not challenging to begin with) as well as a "tying hands audience cost" (i.e., a higher expected payoff difference between backing down and escalating to war in the model with voters than in the model without). Instead, if sufficiently many voters are sufficiently dovish about war, then backing down generates the corresponding "audience benefits." The reason is as follows. After a challenge, the doves form the pessimistic expectation that they are likely to go to war, which they would like to avoid. If the politician backs down, these voters get a payoff gain from the sense of relief that war did not ensue. In the model, therefore, disappointment and relief are two psychological states that occur on the two opposite sides of an endogenous reference point.

After the pioneering work of Kahneman and Tversky (1979), numerous studies uncovered evidence that individual behavior is consistent with the maximization of reference-dependent payoffs. In some relatively recent work, Farber (2008), Fehr et al. (2011) and Pope and Schweitzer (2011) find evidence for reference-dependent payoffs in labor markets, contractual relationships, and non-market settings. Two empirical contributions that are particularly relevant to our application here include Kimball and Patterson (1997) and Waterman et al. (1999), who show that the attitudes of voters towards an elected official are affected by their expectations.

The mounting evidence for reference-dependent preferences has also motivated several other applications in politics prior to ours. Levy (1997) reviews early applications of prospect theory in international relations. Alesina and Passarelli (2015) and Lockwood and Rockey (2016) show that reference-dependent preferences can explain departures from the predictions of standard voting models. Passarelli and Tabellini (2017) use these preferences to explain political unrest. Grillo (2016) shows that reference-dependent voters might punish politicians who make excessively large campaign promises. Martin (2016) shows that, due to reference-dependence, citizens may be more willing to punish corruption when their taxes are higher.

This paper has four more sections. Section 2 presents the model and Section 3 presents the equilibrium analysis. In Section 4, we discuss how some of the recent empirical evidence on audience costs squares with the predictions of our model. We also discuss other approaches that provide foundations for audience costs, and explain how our approach differs from these. Section 5 concludes. Proofs and additional results appear in a supplemental appendix.

2 Model

2.1 The Workhorse Model as a Benchmark

We start by introducing a set of standard assumptions on the workhorse crisis bargaining model described above (and depicted in Figure 1) and reporting its equilibria.

First, we assume that $v > z_S > 0 > z_W$ so that absent any audience cost or benefit, the strong type would choose war at his final decision node, while the weak type would back down; and, the value of the territory for both types of C's leaders is greater than the payoff from war. Second, we assume that $-z_D < 0$ so that D would prefer to concede the territory than to go to war. These reduced form war payoffs can be interpreted as expected payoffs when war is costly and the outcome of war is uncertain.[1] Finally, to avoid having to deal with trivial sources of multiplicity arising from knife-edge cases of indifference, we maintain the assumption throughout the paper that payoffs v, z_W, z_S and $-z_D$ are all generic.

The assumption of generic payoffs implies that the following three cases are exhaustive: (i) $-a > z_S > z_W$, (ii) $z_S > z_W > -a$ and (iii) $z_S > -a > z_W$.

In the first case, both the strong and weak leaders of country C back down at their final decision nodes, so D resists. Furthermore, since $z_S > 0$, this case can arise if and only if $a < 0$. As a result, there is a unique equilibrium in which both types challenge.

In the second case, the audience cost a is so high that both the weak and strong types of C choose war over backing down. Since $-z_D < 0$, there is a unique equilibrium in which D concedes and, consequently, both types of C challenge.

In the third case, the strong type of C chooses war at its final decision node, while the weak type backs down. If $q > v/(v + z_D)$, then there is a unique equilibrium in which D concedes, and both the strong and weak types of C challenge. On the other hand, suppose that $q < v/(v + z_D)$. Then, in equilibrium, D resists with probability $\min\{1, v/(a+v)\}$ and the strong type of C challenges. The weak type, instead, challenges with probability $qz_D/(1 - q)v$ if $a > 0$, with probability 1 if $a < 0$, and with any probability $\sigma_W \geq qz_D/(1 - q)v$ if $a = 0$. In the latter case, there is a continuum of equilibria, including one in which the weak type of C challenges with certainty.[2]

2.2 Augmenting the Model with Behavioral Voters

Our purpose here is to augment the workhorse model in such a way that any cost that the leader of C suffers from backing down arises *endogenously*.

To this end, suppose that there are no exogenous audience costs, $a = 0$, and that after the leaders of C and D make their decisions, a continuum of citizens of C cast votes

[1] For example, suppose that the probability that country C wins the war is p. Then, if the cost of war incurred by the leader of country D is c_D, the expected payoff for that leader is $-z_D := (1 - p)v - c_D$. Similarly, if the cost of war incurred by the weak (strong) leader of country C is c_W (c_S), then the expected payoff under war for the weak (strong) leader of country C is $z_W := pv - c_W$ ($z_S := pv - c_S$). Our assumptions on z_W, z_S and z_D can then be translated to assumptions on p, c_W, c_S and c_D.

[2] In the cases where a type is indifferent between challenging and not challenging because $a = 0$, challenging with certainty is weakly dominant. So weak dominance as a criterion for equilibrium selection would select the equilibrium in which that type challenges.

to re-elect or replace their leader. The politician is re-elected if and only if a majority of voters vote to re-elect him. If the politician is not re-elected, he obtains only a payoff equal to the payoff that he gets from the crisis bargaining game. If he is re-elected, then he gets an additional payoff that we normalize to 1.

We consider the voters to be mechanical actors (whose behavior we specify below) so they are not players in the game.[3] Thus, an equilibrium of the game will specify only the behavior and beliefs of the leaders of the two countries. Let σ_θ denote the equilibrium probability with which the type $\theta \in \{W, S\}$ leader of country C challenges, σ_θ^w the equilibrium probability with which this type chooses war, and σ_D the equilibrium probability with which D resists. The equilibrium strategy profile is therefore $\sigma = \langle(\sigma_\theta, \sigma_\theta^w)_{\theta=W,S}, \sigma_D)\rangle$. Let $\tilde{q}$ denote the equilibrium posterior probability with which the leader of C is considered to be the strong type after choosing to challenge the territory. D's belief about C's type matters only at the information set at which D's chooses to resist or concede, so we may write an equilibrium to be simply the pair $\rho = (\sigma, \tilde{q})$.

There are two types of voters: those whose material payoffs are given by the payoffs of the strong type of leader of country C in the crisis bargaining game, and those whose material payoffs are given by the payoffs of the weak type in the same game. Fraction λ of voters are of the former type, while $1 - \lambda$ are of the latter type. We will refer to these two types of voters as hawks and doves respectively, and use the labels S and W for voters as well. Voters also have a psychological component of payoffs, which is reference-dependent. The material and psychological components of voter payoffs are additively separable for each type θ. We write the sum of these two components as

$$u_\theta = \pi_\theta + \eta(\pi_\theta - \mathbb{E}^\rho[\pi_\theta|\mathcal{I}]), \qquad \theta \in \{W, S\} \tag{1}$$

where π_θ is the material payoff of the type θ leader in the crisis bargaining game, $\mathbb{E}^\rho[\cdot|\mathcal{I}]$ denotes the expectation operator evaluated at an information set $\mathcal{I}$, and given an equilibrium of the game ρ; and $\eta \geq 0$ is the weight on the psychological component of payoffs. We assume that the voters' reference points are determined at the information sets that arise immediately after the initial choice of C's leader to challenge or not challenge, which we label $\mathcal{I}_{ch}$ and $\mathcal{I}_d$ respectively. This assumption reflects the salience of the initial decision of C's leader in forming voters' expectations. At the information set

[3] There are a continuum of them so any voting rule could have been selected in a model in which the voters are considered to be players. Further, we will assume below that voting is probabilistic so the assumption that voters are mechanical is standard.

$\mathcal{I}_d$ each voter knows that her payoff will be 0, so $\mathbb{E}^\rho[\pi_\theta|\mathcal{I}_d] = 0$ for both $\theta \in \{W, S\}$ and all equilibria ρ. Finally, we assume that voters share D's belief about C's type at $\mathcal{I}_{ch}$: the posterior probability with which they think that C is strong is also $\tilde{q}$.

Voting is probabilistic. Each voter receives a stochastic preference shock ε that is drawn uniformly from the interval $\left[-\frac{1}{2\alpha}, \frac{1}{2\alpha}\right]$ and independently across voters; then, each voter votes to reelect the incumbent politician if and only if his payoff (the deterministic part plus the shock) exceeds a stochastic threshold u that is drawn uniformly from the interval $\left[-\frac{1}{2\beta}, \frac{1}{2\beta}\right]$. Here, β measures the overall responsiveness of the electorate to the outcome of the crisis. We make the standard assumption that α and β are sufficiently small so that the probability that the politician is re-elected is

$$\frac{1}{2} + \beta\big[\lambda u_S + (1-\lambda)u_W\big] \tag{2}$$

This quantity is also the additional expected payoff that the leader gets due to the fact that he may be re-elected. Since the term in squared brackets of this expression is simply the population-weighted (utilitarian) average of voters' payoffs, the leader of country C maximizes a payoff equal to the payoff that he receives in the crisis bargaining game, which depends on his type, plus β times the utilitarian average of voters' payoffs, which includes both the material and psychological parts.

Remark 1 Other assumptions could give rise to the result that C's leader maximizes the payoff from the crisis plus (2). One is that the leader of C has weighted utilitarian preferences and places weight β on the average voter payoff. Another is that C's citizens have non-electoral means of rewarding and punishing their leader such that the leader internalizes the average citizen payoff.[4] However, under these alternative assumptions, note that β would no longer be a measure of the responsiveness of the electorate to the outcome of the crisis. Instead, it would measure the extent to which the politician's payoffs were other-regarding, or the extent to which the political process generated incentives for politicians to consider the average voter's interests.

Remark 2 The assumption that voters update their reference point at the two information sets that arise after C's initial choice is natural in our application. It captures the idea that citizens form their expectations based on what they learn after observing their own leader's initial policy choice, but not on the details of inter-state crisis bargaining, which, in practice, are typically opaque. That said, there does not yet exist a theory

[4] Under this assumption, our results are also applicable to cases such as dictatorships where leaders are not directly chosen by voters (see, e.g., Weeks, 2008).

about how to select the information sets at which the endogenous reference points are updated in sequential move games. Given this, in Appendix B we discuss the equilibrium consequences of choosing other sets of information sets in which the reference point is updated. There, we show that audience costs arise even under alternative assumptions about the updating of the reference point.

Remark 3 Voting in our model is retrospective. However, our results extend to the case where voting is prospective and, following the game depicted in Figure 1, the incumbent leader of country C runs for reelection against a challenger with a randomly drawn type. We examine this version in Appendix C, and show that endogenous audience costs arise if voters, besides exhibiting reference dependence, are also "loss averse;" that is, they suffer more from negative deviations from their reference point more than they benefit from equal-size positive deviations.[5]

The intuition for why loss aversion is necessary in this setting is as follows. If voters are retrospective, the payoff threshold u is random and does not depend on the reference point; if they are prospective, a change in the reference point impacts the evaluation of the incumbent and the challenger symmetrically. Therefore, for audience costs to emerge, the challenger has to be more appealing to voters when their reference point has shifted up. This is exactly what happens under loss aversion.

2.3 Endogenous Payoffs and Equilibrium Definition

Substituting (1) into (2) and simplifying, the politician's probability of re-election is

$$\frac{1}{2} + \beta\left[\pi^{\lambda} + \eta\left(\pi^{\lambda} - \mathcal{R}^{\rho}[\mathcal{I}]\right)\right] \tag{3}$$

where

$$\pi^{\lambda} := \lambda\pi_S + (1-\lambda)\pi_W \tag{4}$$

is the population-weighted average of material payoffs given any outcome of the crisis bargaining game, and

$$\mathcal{R}^{\rho}[\mathcal{I}] := \lambda\mathbb{E}^{\rho}[\pi_S|\mathcal{I}] + (1-\lambda)\mathbb{E}^{\rho}[\pi_W|\mathcal{I}] \tag{5}$$

[5] Loss aversion has been found to be a relevant behavioral phenomenon whenever individuals exhibit reference dependence. See, e.g., Kahneman and Tversky (1991), Kahneman et al. (1991), Camerer (2004) and the references therein.

is the the population weighted average value of the endogenous reference point evaluated at an equilibrium ρ and information set $\mathcal{I}$.

As mentioned above, for both types of voters, $\theta \in \{W, S\}$, the endogenous reference point in the case where the C's leader chooses to not challenge the territory is $\mathbb{E}^{\rho}[\pi_\theta|\mathcal{I}_d] = 0$. This implies that $\mathcal{R}^{\rho}[\mathcal{I}_d] = 0$ independently of the equilibrium ρ. Therefore, if either type of leader chooses not to challenge the territory, he is re-elected with probability $\frac{1}{2}$ and obtains a payoff of $\frac{1}{2}$ from not challenging.

At the information set $\mathcal{I}_{ch}$, voters observe that C's leader decided to challenge. Thus, the endogenous reference point of a type θ voter after a challenge is a weighted average of the payoffs that arise at the terminal nodes following the initial challenge, with weights given by the probabilities with which these nodes are reached according to voters' equilibrium beliefs. Formally:

$$\mathbb{E}^{\rho}[\pi_\theta|\mathcal{I}_{ch}] = (1-\sigma_D)\, v + \sigma_D\big[\tilde{q}(\sigma_S^w z_\theta + (1-\sigma_S^w)0) + (1-\tilde{q})(\sigma_W^w z_\theta + (1-\sigma_W^w)0)\big]. \quad (6)$$

This implies that the population weighted average value of the endogenous reference point after country C's leader challenges the territory is

$$\mathcal{R}^{\rho}(\mathcal{I}_{ch}) = (1-\sigma_D)\, v + \sigma_D\, [\tilde{q}\sigma_S^w + (1-\tilde{q})\sigma_W^w]\, z^\lambda \quad (7)$$

where $z^\lambda = \lambda z_S + (1-\lambda) z_W$. Therefore, if the game ends with country D conceding, the expected payoff to both types of country C's leaders is

$$v + \frac{1}{2} + \beta\Big[v + \eta\big(v - \mathcal{R}^{\rho}(\mathcal{I}_{ch})\big)\Big] \quad (8)$$

If the game ends with C's leader backing down, the expected payoff to both types of C's leaders is

$$0 + \frac{1}{2} + \beta\Big[0 + \eta\big(0 - \mathcal{R}^{\rho}(\mathcal{I}_{ch})\big)\Big] \quad (9)$$

and if the game ends with war, the expected payoff to each type θ of C's leaders from choosing war is

$$z_\theta + \frac{1}{2} + \beta\Big[z^\lambda + \eta\big(z^\lambda - \mathcal{R}^{\rho}(\mathcal{I}_{ch})\big)\Big] \quad (10)$$

The payoffs from the various outcomes of the game to the leader of D are simply D's payoffs in the crisis bargaining game, depicted in Figure 1.

Since the payoffs to the two types of leaders of country C are endogenous to the equilibrium strategy and beliefs of D, we say that $\rho = (\sigma, \tilde{q})$ is an equilibrium of the model if (i) $\tilde{q}$ is consistent with Bayesian updating given σ, and (ii) no type of either player has a profitable deviation from the strategy profile σ given beliefs $\tilde{q}$ when the payoffs to all of the outcomes of the game are computed at the equilibrium ρ. In this sense, an equilibrium of a model in which players have reference-dependent preferences with endogenous reference points has the fixed point characteristic that is typical of a rational expectations equilibrium: the reference points are derived from equilibrium behavior and equilibrium behavior is consistent with the endogenous references points.

2.4 Endogenous Audience Costs

It is now already apparent that the politician may suffer an endogenous audience cost from backing down after making a threat. The cost for the leader of C from backing down after a challenge is the payoff difference from backing down after a challenge and not challenging at the start of the game, which is

$$a_s^\rho = \beta\eta\mathcal{R}^\rho(\mathcal{I}_{ch}) \tag{11}$$

If this quantity is positive, it represents an endogenous audience cost that is sunk the moment C's leader decides to challenge. For this reason, we refer to a_s^ρ as a *sunk audience cost* if it is positive, or *benefit* if it is negative.

Similarly, the payoff difference between going to war and backing down for a leader of type θ is $z_\theta + \beta(1+\eta)z^\lambda$. This exceeds the same payoff difference in the workhorse model without voters by the quantity

$$a_t = \beta(1+\eta)z^\lambda \tag{12}$$

Therefore, if there are sufficiently many hawks in the population so that z^λ is positive then the leader of C has an extra incentive to go to war over backing down. In particular, electoral incentives can commit even the weak type of politician to war.[6] On the other hand, if there are sufficiently many doves in the population then z^λ will be negative, so electoral incentives can commit even the strong type politician to back down rather than choose war. For this reason, we refer to a_t as a *tying-hands audience cost* or benefit.[7]

[6] Even though $z_W < 0$, it is possible that $z_W + \beta(1+\eta)z^\lambda > 0$ so that the weak leader's payoff in (10) is greater than his payoff in (9). This means that in the augmented model with electoral incentives, even a weak type may choose war over backing down. See the analysis of case (ii) in Section 3.

[7] The distinction between the sunk and tying hands functions of the audience cost was first pointed out by Fearon (1997).

Both the sunk and tying hands audience costs are endogenous quantities, but the sunk audience cost is also an equilibrium quantity since it depends on $\mathcal{R}^\rho(\mathcal{I}_{ch})$, which is an equilibrium quantity. In fact, the sign of a_s^ρ is completely determined by the sign of $\mathcal{R}^\rho(\mathcal{I}_{ch})$, so whether there is an audience cost or benefit is also determined in equilibrium. In addition, the assumption that voters have reference-dependent payoffs is necessary for generating a sunk audience cost since $a_s^\rho = 0$ if $\eta = 0$. But this assumption is not necessary for generating a tying hands cost since $\eta = 0$ does not imply that $a_t = 0$.[8]

3 Results

3.1 Equilibrium Characterization

Our main result, Proposition 1 below, characterizes the equilibrium set in three cases that mirror the three cases analyzed in the benchmark model: (i) $-a_t > z_S > z_W$, (ii) $z_S > z_W > -a_t$, and (iii) $z_S > -a_t > z_W$. Since the sunk audience cost a_s^ρ is an equilibrium quantity, we also report its equilibrium value.

Proposition 1. *(i) If $-a_t > z_S > z_W$ then there is a double continuum of equilibria in which both the strong and weak types of C back down, D resists, and both types of C are indifferent between not challenging and challenging, so each may challenge with any probability. In all of these equilibria, $a_s^\rho = 0$.*

(ii) If $z_S > z_W > -a_t$ then there is a unique equilibrium in which both types of C choose war, D concedes, and both types of C challenge, so $a_s^\rho = \beta\eta v$.

(iii) If $z_S > -a_t > z_W$ then in any equilibrium, the strong type of C chooses war and challenges, while the weak type backs down. In addition, if $q > v/(v + z_D)$, then there is a unique equilibrium in which D concedes and the weak type of C also challenges, so again $a_s^\rho = \beta\eta v$. If $q < v/(v + z_D)$ then we have three subcases:

[8] Since the tying hands audience cost is not an equilibrium quantity, our model says that a leader cannot strategically generate commitment to one of the two actions at the final decision nodes of the game. Instead, the magnitude and sign of the tying-hands audience cost is determined directly by the model's fundamental parameters (the responsiveness of voters to the outcome of the crisis, β; the weight on the psychological part of their payoffs, η; and the population average of material war payoffs, z^λ). In Appendix C, we show that if voters are prospective instead of retrospective, the tying hands audience cost becomes an equilibrium quantity as well.

(a) *If* $\eta = 0$*, then there is a continuum of equilibria in which D resists, and the weak type of* C *challenges with any probability* $\sigma_W \geq qz_D/(1-q)v$*. In all of these equilibria,* $a_s^\rho = 0$*, so there is no sunk audience cost or benefit.*

(b) *If* $\eta > 0$ *and* $z^\lambda < 0$*, there is a unique equilibrium in which D resists and the weak type of* C *challenges, so there is a sunk audience benefit,* $a_s^\rho = \beta\eta q z^\lambda < 0$*.*

(c) *If* $\eta > 0$ *and* $z^\lambda > 0$ *there is a unique equilibrium in which D resists with probability*

$$\sigma_D = \frac{(1+\beta)(v+z_D)}{(1+\beta)(v+z_D)+\beta\eta z^\lambda}$$

and the weak type of C *challenges with probability* $\sigma_W = qz_D/(1-q)v$*. In this case, the sunk audience cost is*

$$a_s^\rho = (1-\sigma_D)\left[1+\beta(1+\eta)\right]v = \frac{\beta\eta z^\lambda}{(1+\beta)(v+z_D)+\beta\eta z^\lambda}\left[1+\beta(1+\eta)\right]v$$

The results for the first two cases are straightforward. In case (i), both the strong and weak types back down, so D resists. As a result, each type of C's leader is indifferent between challenging and not challenging, giving rise to multiple equilibria. However, since all equilibria yield the same payoffs to voters, there is neither a sunk audience cost, nor benefit. Since $z_S > 0 > z_W$, this case arises only when there are sufficiently many doves in the population, so that $z^\lambda < 0$. In this case, a predominantly dovish electorate forces even a strong leader not to choose war.

In case (ii), both types of C's leaders choose war at their final decision nodes. Therefore, in the unique equilibrium, D concedes and both types challenge. This case arises only if there are sufficiently many hawks in the population, so that z^λ is sufficiently high. Unlike the first case, in this case the electorate works as a commitment device that enables C to credibly threaten to escalate the crisis to war, forcing D to concede. Moreover, since the electorate is predominantly hawkish, the leader faces a sunk audience cost by backing down.

The proof of Proposition 1 in Appendix A, therefore focuses on case (iii) in which the proportion of hawks in the population, z^λ, is neither high nor low. In this case, the strong type of C challenges with certainty, while the weak type does so with a weakly lower probability. This means that the weak leader bluffs with positive probability by pretending to be strong in the hopes of getting a concession from D. Note that this case is compatible with a sunk audience cost, a sunk audience benefit, or neither.

3.2 Comparison with the Benchmark Model

According to Proposition 1, equilibrium payoffs in the augmented model are unique and the equilibria are analogous to the equilibria of the benchmark model.

In the augmented model, the tying hands audience cost, a_t, defines the threshold that separates the three cases where the weak and strong types both back down, both choose war, or make different choices at their final decision nodes, exactly as the exogenous audience cost a does in the benchmark model.

Comparing the equilibrium predictions of the augmented model with those of the benchmark model, we see that in case (i) the equilibrium set of the augmented model is larger than it is in the benchmark model since C's leader can now challenge with probability lower than 1. In case (ii) the two models deliver exactly the same predictions.

In case (iii), there is no sunk audience cost or benefit when $\eta = 0$. This establishes the necessity of reference-dependent payoffs to produce a sunk audience cost in our setting. In this case, for each equilibrium of the augmented model, there is a behaviorally identical equilibrium of the $a = 0$ case of the benchmark model; and vice versa. When $\eta > 0$, the sign of z^λ determines whether there is a sunk audience cost or benefit. When there is a sunk audience benefit, equilibrium behavior in the augmented model is identical to equilibrium behavior in the benchmark model for the case of $a < 0$. When there is a sunk audience cost, equilibrium choices in the augmented model are also similar to equilibrium choices in the benchmark case for $a > 0$. The weak type of C mixes with the same probability in both models, but the probability with which D mixes is different, even after accounting for the equilibrium value of the sunk audience cost. This is because the indifference condition that pins down D's mixing probability in the augmented model (equation (18) in the Appendix) is qualitatively different from the analogous indifference condition in the benchmark model. One key difference is that the augmented model includes re-election payoffs that differ across terminal nodes. Another is that the psychological part of voter payoffs that enters in C's payoff when D concedes also contains the sunk audience cost as a component, whereas in the benchmark model the exogenous audience cost a does not enter C's payoff when D concedes.

3.3 Comparative Statics

Since the tying hands and sunk audience costs are endogenous quantities in the augmented model, we can use Proposition 1 to study their comparative statics.

We start by reporting the comparative statics of the tying hands audience cost, a_t. The sign of a_t is determined by the sign of z^λ, so there is an audience cost when citizens are predominantly hawks, and an audience benefit when they are predominantly doves. In addition, the magnitude of this cost or benefit is increasing in how predominant the hawks or doves are. The magnitude is also increasing in β, which means that when voting behavior is more responsive to the outcome of the crisis, or when the politician weights the average voter payoff more, there is a larger audience cost or benefit. Lastly, the magnitude is increasing in η, which means that when the psychological part of voter payoffs becomes more important, there is a greater audience cost or benefit. Therefore, the tying hands audience cost works as a commitment device that commits the politician to war when sufficiently many citizens are hawkish about war, when the outcome of the crisis matters more in their voting decisions, and when expectations matter more in determining their payoffs.

The sunk audience cost a_s^ϱ is an equilibrium quantity that can vary with the equilibrium updated belief $\tilde{q}$ that C's leader is strong, the equilibrium probability σ_D with which D resists, and the equilibrium choices of C's types at their final decision nodes. So we must take this into account when studying the comparative statics of a_s^ϱ. These comparative statics are by and large similar to those of the tying hands audience cost, with only a few notable differences. Again, the sign of a_s^ϱ is determined by the sign of z^λ. As well, the magnitude of this audience cost is again increasing with the magnitude of z^λ. When $z^\lambda > 0$, the sunk audience cost a_s^ϱ is increasing in both β and η. However, when $z^\lambda < 0$, it is piecewise constant in these parameters, with a jump to 0 when $-a_t$ crosses z_S. This jump is negative if $q > v/(v + z_D)$ and positive if $q < v/(v + z_D)$. Thus, if voters are predominantly doves and the prior probability of C's leader being strong is high, the sunk audience cost is weakly decreasing in β and η, reflecting the idea that a dovish electorate is lenient toward a leader who backs down. On the other hand, if the electorate is largely dovish, but C's leader is ex ante likely to be weak, then there is an audience benefit that decreases with β and η: as these parameters increase, the leader is more tempted to try to reap the audience benefit by challenging and then backing down. Since the reference point is endogenously determined, voters will anticipate this opportunistic behavior and the audience benefit will decrease. Finally, if there is a sunk audience cost, then it is increasing in v. Indeed, if voters are predominantly hawks, then as the value of the disputed territory goes up, the expected payoff of the voters after a concession by country D goes up as well. As a result, the reference point of the voters after a challenge, and hence the sunk audience cost, increases.[9]

4 Discussion

4.1 Empirical Evidence

Although the predictions of our model are mostly novel, some of our assumptions and predictions find support in experimental investigations of the audience cost.

The first experimental study of the audience cost was done by Tomz (2007), who estimates the sunk audience cost from survey data.[10] Tomz (2007) estimates a positive audience cost, and finds that the audience cost is higher among more politically active respondents. Though political engagement may not be the obvious way to measure voter responsiveness, this finding provides some evidence that is consistent with our prediction that the audience cost is increasing in the responsiveness of the electorate, β, to the crisis outcome.[11] That said, Tomz (2007) also presents some evidence that the public punishes leaders for bluffing because they think that bluffing hurts the leader's (and country's) reputation. While this differs from the disappointment-based mechanism in our paper, this evidence is based on agents self-reporting their disapproval of bluffing, and these subjects may not have had a consistent way of expressing their disappointment for the leader not following through on a challenge.

Building on Tomz's (2007) approach, Trager and Vavreck (2011) estimate a leader's public approval at every outcome of the crisis bargaining game, enabling them to estimate both the sunk audience cost and the tying hands audience cost as defined in this paper. They find positive values for both of these costs. They also find that presidential approval is highest when the adversary concedes, and can be lowest when the leader backs down —even lower than approval after the war outcome. They also show that among respondents who oppose a hawkish foreign policy rhetoric, there is an audience

[9]This comparative static result with respect to v would continue to hold even if we substituted the payoffs following a war with the lottery payoffs described in footnote 1.

[10]Tomz (2007) argues that despite concerns about external validity, the experimental approach sidesteps several of the challenges in estimating the audience cost in observational studies, such as partial observability and strategic selection (Schultz, 2001).

[11]This result was replicated in the UK by Davies and Johns (2013), who found that the audience cost was highest among the most politically engaged British respondents. However, one result of theirs that goes against the grain of our predictions concerning the relationship between responsiveness and the audience cost is that political knowledge, which may also be correlated with responsiveness, did not substantially moderate the audience cost.

benefit rather than cost. Taking their presidential approval measure to be a proxy for the election payoff given in (2), these findings support our model. In particular, they support the prediction that the sign of z^λ determines the sign of the audience cost. If there are sufficiently many doves in the electorate, then there can be an audience benefit, though in their data Trager and Vavreck (2011) find that hawks outnumber doves.

Also building on Tomz (2007), Davies and Johns (2013) estimate the audience cost in the U.K. with variation in crisis type. They find that among voters, the disapproval for bluffing by the British prime minister was lower in a nuclear crisis scenario than in an ally defense crisis scenario, which was in turn lower than in a hostage crisis scenario. This suggests that the audience cost may potentially vary with the importance or scale of the issue, measured in our model by v. However, whether their findings show that it increases or decreases with scale remains unclear.

These studies provide some suggestive evidence for our theory, but more can be done to directly test the assumptions and predictions of our model.

4.2 Other Approaches

One influential theory of the audience cost is that leaders suffer a cost from the damage to their reputation that bluffing causes.[12] A simple and natural extension to the benchmark model that captures this story says that voters prefer to re-elect the strong type and replace the weak type. Suppose that the strong type challenges, and chooses war over backing down. If the weak type separates from the strong type at his final decision node, then he is not re-elected. However, he would not be re-elected even if he separated at the initial decision node, as this decision would also reveal his type to the voters. Therefore, this simple reputation-based extension does not produce an endogenous audience cost.

Smith (1998) circumvents this problem by assuming that there are a continuum of types in an ally defense scenario. When the politician is inferred to be stronger, he is re-elected with higher probability. The set of types is partitioned into those that announce that they will support the ally against the adversary, and those that announce that they will not. In equilibrium, those that announce that they will support the ally follow through. If a deviation takes place, however, then Smith (1998) has the voters think that the type is the weakest possible type and re-elect him with the lowest probability. Thus, he generates audience costs with the help of off-path beliefs. However, for every

[12]As mentioned above, this is the theory that Tomz (2007) claims to find the strongest empirical support for based on open-ended survey responses.

profile of parameters (i.e. payoffs and initial beliefs) his game also has equilibria in which audience costs do not arise. Moreover, these equilibria cannot be ruled out using standard refinements.[13]

Guisinger and Smith (2002) also develop a theory of audience costs based on reputation, but depart further from the standard crisis bargaining scenario. In their model, two countries play a repeated demand bargaining game with adverse selection. In the one shot game, communicating a credible threat is not possible; but since the game is repeated, credible communication can be supported by an equilibrium strategy profile that reverts to babbling if the lying side is caught. Since payoffs are lower in the babbling equilibrium, voters would like to replace the lying politician after he is caught and start afresh with a new leader. Again, audience costs are supported by the selection of one of many possible equilibria of the game; and, in fact, equilibria that are renegotiation-proof in the sense of Farrell and Maskin (1989) do not support audience costs.

Other papers that provide foundations for audience costs include Ashworth and Ramsay (2010) and Slantchev (2006). Slantchev (2006) studies a game between a voter, a politician, an opposition party, and media, abstracting away from the foreign adversary. He shows that an audience cost for implementing bad policies arises when voters are not perfectly informed about the quality of the policy implemented by the politician and a non-strategic media source can convey information about the policy. His justification for audience costs relies on assumptions about what kind of evidence can and cannot be provided to citizens, as well as the existence of exogenous and unbiased news providers. Ashworth and Ramsay (2010) take a mechanism design approach and show that an optimizing voter would design incentives to punish a politician for bluffing. However, they do not investigate the classical case in which there is adverse selection and the voter does not possess commitment power; e.g., the situation (corresponding to our extension in Appendix C) in which voting is prospective and the incumbent politician possesses private information about his type.

Our paper differs from the prior literature in at least three ways. First, we directly extend the canonical crisis bargaining model, endogenizing the audience costs in such a way as that equilibrium behavior in the extended model is directly analogous to equi-

[13] Smith's (1998) game is neither a standard signaling game, nor a standard cheap talk game, though it has some features of both. This means that standard equilibrium refinements for signaling games must be adapted to his specific game. Furthermore, in his model, if the weakest possible type is better off by threatening the intervention and then not following through despite the bluff being called, types above it may also want to do the same. As a result, refinements like the ones proposed by Banks and Sobel (1987) do not uniquely select equilibria that support audience costs.

librium behavior in the benchmark model with exogenous audience costs. Second, we do this with behavioral voters. Third, and most importantly, our model provides a psychological theory for audience costs, based on disappointment and relief, and can then justify audience benefits as well.

5 Conclusion

When rational politicians make threats, voters with reference-dependent preferences may raise their expectations about how successful the politician will be in extracting concessions from the adversary. If the politician eventually backs down from the threat, voters who formed high expectations are "disappointed." If voting behavior is based on this disappointment such that the politician's re-election probability is decreasing in it, then the politician suffers an audience cost from making the challenge and subsequently backing down.

This is the logic upon which we have developed our model of audience costs. We developed the theory by adding to the standard crisis bargaining model a voting stage in which voters have reference-dependent payoffs. The model endogenizes both a sunk audience cost and a tying hands audience cost, and it produces new comparative statics predictions about the sign and magnitudes of these costs. If voters are predominantly hawkish about war, then both audience costs are positive, and are a consequence of voters being "disappointed" that their expectations were not met. But if they are predominantly dovish about war, then both audience costs are negative, turning them into audience benefits. In this case, dovish voters who were worried about the possibility that the crisis would end in war get a payoff benefit from the "relief" that they experience from seeing their leader back down. Leaders can then face audience costs or benefits depending on how hawkish or dovish their voters are.

The magnitudes of these audience costs or benefits depend on the model's parameters. The audience cost or benefit is increasing in the responsiveness of the electorate to the outcome of the crisis, as well as in the salience of the psychological component of payoffs. The magnitude of the sunk audience cost is also increasing in the value of the territory, or the importance of the issue to voters. These comparative statics predictions of the model can be tested empirically.

References

Alesina, A. and F. Passarelli (2015): "Loss Aversion in Politics," *NBER Working Paper # 21077.*

Ashworth, S. and K. W. Ramsay (2010): "Should Audiences Cost? Optimal Domestic Constraints in International Crises," *manuscript, Princeton University.*

Banks, J. S. and J. Sobel (1987): "Equilibrium selection in signaling games," *Econometrica*, 647–661.

Camerer, C. F. (2004): "Prospect theory in the wild: Evidence from the field," *Advances in Behavioral Economics*, 148–161.

Davies, G. A. and R. Johns (2013): "Audience costs among the British public: the impact of escalation, crisis type, and prime ministerial rhetoric," *International Studies Quarterly*, 57, 725–737.

Farber, H. S. (2008): "Reference-dependent preferences and labor supply: The case of New York City taxi drivers," *American Economic Review*, 98, 1069–1082.

Farrell, J. and E. Maskin (1989): "Renegotiation in repeated games," *Games and Economic Behavior*, 1, 327–360.

Fearon, J. D. (1994): "Domestic political audiences and the escalation of international disputes." *American Political Science Review*, 88, 577–592.

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

Fehr, E., O. Hart, and C. Zehnder (2011): "Contracts as Reference Points–Experimental Evidence," *American Economic Review*, 101, 493–525.

Grillo, E. (2016): "The hidden cost of raising voters expectations: Reference dependence and politicians credibility," *Journal of Economic Behavior & Organization*, 130, 126 – 143.

Guisinger, A. and A. Smith (2002): "Honest Threats: The Interaction of Reputation and Political Institutions in International Crises," *Journal of Conflict Resolution*, 46, 175–200.

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1991): "Anomalies: The endowment effect, loss aversion, and status quo bias," *Journal of Economic Perspectives*, 5, 193–206.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–91.

——— (1991): "Loss Aversion in Riskless Choice: A Reference-Dependent Model," *Quarterly Journal of Economics*, 106, 1039–61.

Kőszegi, B. and M. Rabin (2006): "A model of reference-dependent preferences," *Quarterly Journal of Economics*, 1133–1165.

——— (2007): "Reference-Dependent Risk Attitudes," *American Economic Review*, 97, 1047–1073.

Kimball, D. C. and S. C. Patterson (1997): "Living up to Expectations: Public Attitudes toward Congress," *Journal of Politics*, 59, 701–728.

Levy, J. S. (1997): "Prospect theory, rational choice, and international relations," *International Studies Quarterly*, 41, 87–112.

Lockwood, B. and J. Rockey (2016): "Negative Voters? Electoral Competition with Loss-Aversion," *manuscript, University of Warwick*.

Martin, L. (2016): "Taxation, loss aversion, and accountability: theory and experimental evidence for taxation's effect on citizen behavior," *manuscript, Duke University*.

Passarelli, F. and G. Tabellini (2017): "Emotions and Political Unrest," *Journal of Political Economy*, 125, 903–946.

Pope, D. G. and M. E. Schweitzer (2011): "Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes," *American Economic Review*, 101, 129–157.

Schultz, K. A. (1999): "Do democratic institutions constrain or inform? Contrasting two institutional perspectives on democracy and war," *International Organization*, 53, 233–266.

——— (2001): "Looking for audience costs," *Journal of Conflict Resolution*, 45, 32–60.

Slantchev, B. L. (2006): "Politicians, the media, and domestic audience costs," *International Studies Quarterly*, 50, 445–477.

Smith, A. (1998): "International crises and domestic politics," *American Political Science Review*, 92, 623–638.

Tomz, M. (2007): "Domestic audience costs in international relations: An experimental approach," *International Organization*, 61, 821–840.

Trager, R. F. and L. Vavreck (2011): "The political costs of crisis bargaining: Presidential rhetoric and the role of party," *American Journal of Political Science*, 55, 526–545.

Waterman, R., H. C. Jenkins-Smith, and C. L. Silva (1999): "The Expectations Gap Thesis: Public Attitudes toward an Incumbent President," *Journal of Politics*, 61, 944–966.

Weeks, J. L. (2008): "Autocratic audience costs: Regime type and signaling resolve," *International Organization*, 62, 35–64.

Appendix

A Proof of Proposition 1

The results for cases (i) and (ii) follow from the discussion in the main text following the statement of Proposition 1. Here we examine case (iii).

In case (iii), the weak and strong types make separating choices at their final decision nodes: the weak type chooses to back down while the strong type chooses war. As a result, we have

$$\mathcal{R}^{\rho}(\mathcal{I}_{ch}) = (1 - \sigma_D)v + \sigma_D \tilde{q} z^{\lambda}. \tag{13}$$

From here, the proof proceeds in two more steps. In the first step, we prove that in any equilibrium the strong type of C challenges with probability 1. In the second step, we provide a characterization of the full equilibrium set by searching for equilibria in three exhaustive cases: the case where D concedes, the case where D resists, and the case where D mixes between conceding and resisting.

Lemma A.1. *In case (iii), the strong type challenges in equilibrium.*

Proof: Suppose, for the sake of contradiction, that there is an equilibrium in which the strong type of C challenges with probability less than 1. If there were such an equilibrium, then the strong type's expected payoff from challenging could not exceed his expected payoff from not challenging, i.e.

$$\begin{aligned} 0 &\geq \sigma_D\big[z_S + \beta\big(z^{\lambda} + \eta(z^{\lambda} - \mathcal{R}^{\rho}(\mathcal{I}_{ch}))\big)\big] + (1 - \sigma_D)\big[v + \beta\big(v + \eta(v - \mathcal{R}^{\rho}(\mathcal{I}_{ch}))\big)\big] \\ &= \sigma_D\big[z_S + \beta(1 + \eta(1 - \tilde{q}))z^{\lambda}\big] + (1 - \sigma_D)(1 + \beta)v \end{aligned} \tag{14}$$

where the second line follows from substituting $\mathcal{R}^{\rho}(\mathcal{I}_{ch})$ from (13). If $z^{\lambda} \geq 0$, the right side of (14) is strictly positive, establishing the contradiction. If $z^{\lambda} < 0$, we have $z_S + \beta(1 + \eta(1 - \tilde{q}))z^{\lambda} > z_S + \beta(1 + \eta)z^{\lambda} > 0$, where the last inequality follows from the fact that we are analyzing a case where $z_S > -a_t$, and by the definition of a_t. Thus, (14) is again positive, establishing the contradiction. □

Since the strong type always challenges in equilibrium, $\tilde{q}$ is pinned down by Bayes rule. In particular,

$$\tilde{q} = \frac{q}{q + \sigma_W(1 - q)}. \tag{15}$$

Then, given that the two types of C separate at their final decision nodes, D chooses to concede only if

$$\tilde{q}(-z_D) + (1-\tilde{q})v \tag{16}$$

is weakly less than 0, in which case D's expected payoff from resisting is weakly lower than her expected payoff from backing down. D chooses to resist only if (16) is weakly greater than 0, and D chooses to mix between conceding and resisting only if it is exactly equal to 0. We now complete the characterization of the equilibrium set, organizing the analysis according to D's equilibrium choices.

Equilibria where D concedes Suppose that D concedes, so $\sigma_D = 0$. Then $\mathcal{R}^\rho(\mathcal{I}_{ch}) = v$ and the payoff to both types of C from challenging is $v + \frac{1}{2} + \beta v$, which is greater than $\frac{1}{2}$, the payoff from not challenging. So both types challenge, and $\tilde{q} = q$. Then, for it to be optimal for D to concede we need (16) to be at least as large as 0 when $\tilde{q} = q$; that is, we need $q \geq v/(v+z_D)$. Thus, if the prior q is above $v/(v+z_D)$ there is an equilibrium in which D concedes, and both types of C challenge. If $q < v/(v+z_D)$, then D has a profitable deviation and there is no equilibrium in which D concedes for sure.

Equilibria where D resists Next, consider the case where D resists, so $\sigma_D = 1$. For D to want to resist we would need (16) to be weakly greater than 0 evaluated when $\tilde{q}$ is given by (15). Thus, we need $\sigma_W \geq qz_D/(1-q)v$. This latter inequality defines a feasible value of σ_W if and only if $q \leq v/(v+z_D)$.

Now suppose that $\eta = 0$. The weak type of C is always indifferent between challenging and not challenging since his expected payoff from challenging is $\frac{1}{2} - \beta\eta\mathcal{R}^\rho(\mathcal{I}_{ch}) = \frac{1}{2}$ and his expected payoff from not challenging is also $\frac{1}{2}$. Therefore, when $q \leq v/(v+z_D)$ and $\eta = 0$, there is a continuum of equilibria in which D resists and the weak type of C challenges with any probability $\sigma_W \geq qz_D/(1-q)v$.

Lastly, consider the case where $\eta > 0$. If $z^\lambda > 0$, then there is no equilibrium where D resists, because if this were the case, the weak type's payoff from not challenging, $\frac{1}{2}$, would exceed his equilibrium payoff from challenging, $\frac{1}{2} - \beta\eta\mathcal{R}^\rho(\mathcal{I}_{ch})$, giving this type a profitable deviation. On the other hand, if $z^\lambda < 0$ then the weak type would want to challenge. Therefore, when $\eta > 0$ there is an equilibrium in which D resists if and only if $z^\lambda < 0$. In this equilibrium, both the weak and strong types challenge.

Equilibria where D mixes Suppose that D mixes between conceding and resisting. To mix, D must be indifferent, so (16) must equal 0. Substituting (15) into this indifference condition gives us

$$0 = \frac{q}{q + \sigma_W(1-q)}(-z_D) + \frac{\sigma_W(1-q)}{q + \sigma_W(1-q)}v \tag{17}$$

This pins down the equilibrium value of σ_W, which is $\sigma_W = qz_D/(1-q)v$. As in the previous case, this condition defines a feasible value for σ_W if and only if $q \leq v/(v+z_D)$. If this condition is satisfied, then $\tilde{q} = v/(v+z_D)$. Otherwise, there is no equilibrium in which D mixes. Thus, suppose $q < v/(v+z_D)$.[14] Since $\sigma_W \in (0,1)$, the weak type of C must also be indifferent between challenging and not challenging, we need

$$0 = (1-\sigma_D)\left(v + \beta\left[v + \eta\left(v - \mathcal{R}^\rho(\mathcal{I}_{ch})\right]\right) + \sigma_D\left(-\beta\eta\mathcal{R}^\rho(\mathcal{I}_{ch})\right) \tag{18}$$

Now we substitute the equilibrium belief $\tilde{q} = v/(v+z_D)$ into $\mathcal{R}^\rho(\mathcal{I}_{ch})$ in (13), and then $\mathcal{R}^\rho(\mathcal{I}_{ch})$ from (13) into (18), and solve for σ_D to get

$$\sigma_D = \frac{(1+\beta)v}{(1+\beta)v + \beta\eta\tilde{q}z^\lambda} = \frac{(1+\beta)(v+z_D)}{(1+\beta)(v+z_D) + \beta\eta z^\lambda} \tag{19}$$

This implies that there is no equilibrium in which D mixes if $z^\lambda < 0$ or $\eta = 0$, but there is such an equilibrium when z^λ and η are both positive.

B Alternative Updating Rules

In the main text, we assumed that the endogenous reference point for voters is updated at two information sets: the one following C's decision not to challenge, and the one following C's decision to challenge. Here, we discuss the equilibrium consequences of alternative modeling choices. We maintain the assumption that in the benchmark model there is no exogenous audience cost, $a = 0$.

The model has three decision points: C's initial decision of whether or not to challenge the territory, D's decision of whether or not to concede, and C's decision of whether to escalate or back down. This means that there are a total of four natural possibilities: the reference point is updated after only the first decision (the case analyzed in the

[14]The case where $q = v/(v+z_D)$ would yield a continuum of equilibria. Since our assumption that z_D is generic rules out this case, we do not characterize the set of equilibria for this case.

main body of the paper), the reference point is never updated (section B.1 below), the reference point is updated after all three decisions (section B.2 below), and the reference point is update after the first and second decisions (section B.3 below).[15]

B.1. The reference point is updated nowhere

If the endogenous reference point is determined (based on rational expectations about equilibrium behavior) at the initial information set and it is *never* updated, then there is no sunk audience cost, $a_s^\rho = 0$. Instead, the tying hands audience cost would be the same as the one we characterized in the main text, $a_t = \beta(1+\eta)z^\lambda$. Then, we can have one of three possible cases:

(i) If $-a_t > z_S > z_W$, then both types of C backs down, D resists and both types of C are indifferent between not challenging and challenging, so each may challenge with any probability.

(ii) If $z_S > z_W > -a_t$, then both types of C choose war, D concedes and both types of C choose to challenge.

(iii) If $z_S > -a_t > z_W$, the strong type of C chooses war, while the weak type chooses to back down. If $q > v/(v + z_D)$ then there is a unique equilibrium in which D concedes and both the strong and weak types of C challenge. If $q < v/(v + z_D)$, then D resists, the strong type of C challenges, and the weak type challenges with any probability weakly larger than $qz_D/(1-q)v$.

B.2. The reference point is updated everywhere

As mentioned in the main text, if the voters' endogenous reference points are updated at *every* information set of the game, including all terminal information sets, then the equilibrium set of the game is the same as in the augmented model with $\eta = 0$.

Since voters update their reference point at every terminal information set, they cannot be pleasantly surprised or disappointed. As a result, in every equilibrium ρ, $a_s^\rho = 0$ and $a_t = \beta z^\lambda$. Equilibrium behavior is then identical to the one we provided in the main text for the specific case in which $\eta = 0$.

[15]Note that for the assumption of endogenous reference-dependent payoffs to play a role in affecting equilibrium behavior, it must be that the endogenous reference point is not updated at every information set. In this case, the equilibrium of the model would be behaviorally identical to the equilibrium of the benchmark model without reference-dependent payoffs.

B.3. The reference point is updated after C's initial choice, and D's choice

Finally, suppose that the endogenous reference point of voters is updated after C's decision of whether or not to challenge, and also after D's decision of whether or not to resist. Then $\mathcal{R}^\rho(\mathcal{I}_d) = 0$, and $\mathcal{R}^\rho(\mathcal{I}_{co}) = v$, where $\mathcal{I}_{co}$ is the information set following D's decision to concede. Also, $\mathcal{R}^\rho(\mathcal{I}_r) = [\tilde{q}\sigma_S^w + (1-\tilde{q})\sigma_W^w]z^\lambda$, where $\mathcal{I}_r$ is the information set following D's decision to resist. The sunk audience cost in any given equilibrium ρ is $a_s^\rho = \beta\eta\mathcal{R}^\rho(\mathcal{I}_r)$, and the tying hands audience cost is again $a_t = \beta(1+\eta)z^\lambda$.

In this case, the equilibria of the game can be pinned down following the same steps we used in the main text. We summarize behavior in the equilibrium set as follows:

(i) If $-a_t > z_S > z_W$, then there is a double continuum of equilibria in which both the strong and weak types of C back down, D resists and each type of C challenges with any probability. Thus, $a_s^\rho = 0$.

(ii) If $z_S > z_W > -a_t$, then there is a unique equilibrium in which both types of C choose war at their final decision nodes, D concedes, and both types of C challenge. In this case, the sunk audience cost is $a_s^\rho = \beta\eta z^\lambda$.

(iii) If $z_S > -a_t > z_W$, then in any equilibrium, the strong type of C chooses war at its final decision node while the weak type backs down. Thus, $a_s^\rho = \beta\eta\tilde{q}z^\lambda$. If $q > v/(v+z_D)$, then both types challenge at the initial decision nodes, D concedes, and $a_s^\rho = \beta\eta q z^\lambda$. Instead, if $q < v/(v+z_D)$, then the strong type challenges at its initial decision node, and we have three subcases:

   (a) If $\eta = 0$, then there is a continuum of equilibria in which D resists, and the weak type of C challenges with any probability $\sigma_W \geq qz_D/(1-q)v$. In all of these equilibria, $a_s^\rho = 0$, so there is no sunk audience cost or benefit.

   (b) If $\eta > 0$ and $z^\lambda < 0$, there is a unique equilibrium in which D resists and the weak type of C challenges. Thus, there is a sunk audience benefit equal to $a_s^\rho = \beta\eta q z^\lambda < 0$.

   (c) If $\eta > 0$ and $z^\lambda > 0$ there is a unique equilibrium in which D resists with probability
$$\sigma_D = \frac{v + z_D}{v + z_D + \beta\eta z^\lambda}$$

and the weak type of C challenges with probability $\sigma_W = qz_D/(1-q)v$. In this case, the sunk audience cost is given by

$$a_s^\rho = \left(\frac{v}{v+z_D}\right)\beta\eta z^\lambda > 0.$$

Thus, behavior in the equilibrium set of the augmented model continues to be analogous to behavior in the equilibrium set of the benchmark model even under the alternative assumption that the endogenous reference points are updated after D's decision as well. It is also straightforward to verify that the comparative statics of the audience costs under this assumption are similar to the comparative statics under the updating assumption made in the main text.

C Prospective Voting

In the main text we assumed that voters vote retrospectively. Here we show how our main result generalizes to a setting in which voters vote prospectively: they care about the type of their leader and use the outcome of the crisis bargaining game to make inferences about the incumbent's type. After making these inferences, voters decide whether to re-elect the incumbent or replace him with a challenger.

Suppose that country C is led by an incumbent, who plays the game depicted in Figure 1 of the main text and can be one of two possible types: weak, W, and strong, S. Again, there is no exogenous audience cost so $a = 0$. At the end of the crisis bargaining part of the game, the incumbent runs for reelection against a challenger, whose type is drawn from the same distribution as that of the incumbent. Thus, the challenger is strong with probability $q \in (0, 1)$. The types of the incumbent and the challenger are their own private information.

As in the main text, assume that the outcome of the election is determined by the vote of a unit mass of voters belonging to one of two types: hawks and doves. Voters vote sincerely. The proportion of hawks in the population is equal to λ, so $1 - \lambda$ are doves. The two types of voters differ in their preferences concerning the leader they want in office: a hawk prefers a strong politician, while a dove prefers a weak politician. Hawks gets a payoff equal to 1 if they support a strong politician and payoff of 0 if they support a weak politician. Doves are the reverse: they get a payoff of 1 from supporting a weak politician and 0 from supporting a strong one. Voters choose who to vote for based on the politicians' expected types (and the realization of some random shocks described below), but different voters disagree on which type is desirable.

Voting is probabilistic. Let $\tilde{q}_\omega$ be the probability that voters assign to the incumbent leader being strong when terminal node ω is reached. (Since all actions are uniquely labeled, we will abuse notation by identifying terminal nodes with the action that leads to them.) Absent reference dependence, hawkish voter i votes for the incumbent against the challenger at terminal node ω if and only if $\tilde{q}_\omega + \epsilon_i + \delta \geq q$, where ε_i is a stochastic preference shock to voter i's payoff in favor of the incumbent, and δ is a stochastic aggregate popularity shock in favor of the incumbent that hits all voters (hawks and doves) in the same way.[16] As in the main text, for each voter i, ε_i is drawn uniformly from the interval $\left[-\frac{1}{2\alpha}, \frac{1}{2\alpha}\right]$. The popularity shock δ is drawn uniformly from the interval $\left[-\frac{1}{2\beta}, \frac{1}{2\beta}\right]$. Analogously, in the absence of reference dependence, a dovish voter supports the incumbent against the challenger if and only if

$$(1 - \tilde{q}_\omega) + \epsilon_i + \delta \geq (1 - q).$$

Now, suppose that voters have reference-dependent payoffs. As in the main text, the reference point is determined after the initial decision of the leader on whether to challenge. Therefore, the reference utility of a hawk (resp., dove) is equal to the expected probability with which the leader is strong (resp., weak). Formally, let $\mathbb{E}^\rho[q_\omega \mid \mathcal{I}]$ be the expected probability with which the incumbent is believed to be strong at information set $\mathcal{I}$, where the expectation is taken over the distribution of final outcomes ω that are possible after information set $\mathcal{I}$. Given voters' payoff, this is precisely the hawkish voter's reference point at information set $\mathcal{I}$. A dove voter's reference point at the same information set is simply the complement, $1 - \mathbb{E}^\rho[q_\omega \mid \mathcal{I}]$.

In addition, we also assume that voters are loss averse, namely they are harmed by negative deviations from their reference utility more than they are benefited by equal-size positive deviations, but that their payoffs are piece-wise linear around the reference point. Then, a hawk votes for the incumbent at terminal node ω if and only if:

$$\begin{aligned} q_\omega + \eta q_\omega(1 - \mathbb{E}^\rho[q_\omega \mid \mathcal{I}]) + \eta\ell(1 - q_\omega)(0 - \mathbb{E}^\rho[q_\omega \mid \mathcal{I}]) + \epsilon_i + \delta \geq \\ q + \eta q(1 - \mathbb{E}^\rho[q_\omega \mid \mathcal{I}]) + \eta\ell(1 - q)(0 - \mathbb{E}^\rho[q_\omega \mid \mathcal{I}]), \end{aligned}$$

[16] We break ties assuming that, whenever indifferent, the voter votes for the incumbent. We make the same assumption throughout this section, but our analysis does not hinge on it.

where $\eta > 0$ captures the importance of psychological payoffs as opposed to consumption ones and $\ell > 1$ captures the degree of loss aversion, i.e. the extent to which losses loom larger than gains in the mind of the voters. Notice that at terminal node ω, the voter is uncertain about the types of the incumbent and of the challenger. As a result, they keep into account that they may experience a gain if the politician they support turns out to be strong (which happens with probability q_ω for the incumbent and q for the challenger) or to a loss if the politician they support turns out to be weak (which happens with complementary probabilities. Similarly, a dove votes for the incumbent at terminal node ω if and only if:

$$(1 - q_\omega) + \eta\ell q_\omega(0 - (1 - \mathbb{E}^\rho[q_\omega \mid \mathcal{I}])) + \eta(1 - q_\omega)(1 - (1 - \mathbb{E}^\rho[q_\omega \mid \mathcal{I}])) + \epsilon_i + \delta \geq$$
$$(1 - q) + \eta\ell q(0 - (1 - \mathbb{E}^\rho[q_\omega \mid \mathcal{I}])) + \eta(1 - q)(1 - (1 - \mathbb{E}^\rho[q_\omega \mid \mathcal{I}])),$$

Thus, the vote shares the incumbent gets among hawkish and dovish voter are respectively given by:

$$\frac{1}{2} + \alpha(q_\omega - q)\left[1 + \eta + \eta(\ell - 1)\mathbb{E}^\rho[q_\omega \mid \mathcal{I}]\right)] + \alpha\delta$$
$$\frac{1}{2} - \alpha(q_\omega - q)\left[1 + \eta\ell - \eta(\ell - 1)\mathbb{E}^\rho[q_\omega \mid \mathcal{I}]\right)] + \alpha\delta$$

Combining these two vote shares and exploiting the distributional assumption on δ, we find that the probability with which the incumbent is reelected at terminal node ω is:

$$V(\omega \mid \mathbb{E}^\rho[q_\omega \mid \mathcal{I}]) = \frac{1}{2} + \psi(q_\omega - q)[\lambda(1 + \eta) - (1 - \lambda)(1 + \eta\ell) + \eta(\ell - 1)\mathbb{E}^\rho[q_\omega \mid \mathcal{I}]]. \tag{20}$$

Thus, the winning probability of the incumbent is endogenous insofar as $\mathbb{E}^\rho[\tilde{q}_\omega \mid \mathcal{I}]$ is determined by equilibrium behavior, ρ. Furthermore, whenever σ_W and σ_S are neither both equal to 1 nor both equal to 0, we can apply Bayes rule to conclude that:

$$\mathbb{E}^\rho[\tilde{q}_\omega \mid \mathcal{I}_d] = \frac{q(1 - \sigma_S)}{q(1 - \sigma_S) + (1 - q)(1 - \sigma_W)};$$
$$\mathbb{E}^\rho[\tilde{q}_\omega \mid \mathcal{I}_{ch}] = \frac{q\sigma_S}{q\sigma_S + (1 - q)\sigma_W}.$$

In the remaining cases, one of the two reference points is determined by out-of-equilibrium beliefs. To pin down reference points in these cases, we invoke the D1 criterion adapted in the natural way to our setting (see e.g. Banks and Sobel, 1987).

As in the main text, for any ρ, the continuation value to C's leader from choosing to challenge is (weakly) higher for the strong type than for the weak type. Thus,

$$\sigma_S \geq \sigma_W \qquad \text{and} \qquad \mathbb{E}^{\rho}[\tilde{q}_\omega \mid \mathcal{I}_{ch}] \geq \mathbb{E}^{\rho}[\tilde{q}_\omega \mid \mathcal{I}_d].$$

Moreover, by looking at the decision on whether to back down or go to war, we have one of three possible cases: (i) both types choose war with probability 1 (in which case $\tilde{q}_{war} = q$ and the D1 criterion would yield $\tilde{q}_{back\ down} = 0$), (ii) both types choose to back down with probability 1 (in which case $\tilde{q}_{back\ down} = q$ and the D1 criterion would yield $\tilde{q}_{war} = 1$), and (iii) the two types separate with the strong type choosing war and the weak type choosing to back down (in which case $\tilde{q}_{war} = 1$ and $\tilde{q}_{back\ down} = 0$).

In particular, the first case arises if the fraction of hawks in the population is sufficiently high, the second if it is sufficiently low and the third one arises if the fraction of hawks in the population is neither too high nor too low. In this last case, strong types challenge for sure at the initial node, while weak leaders randomize between challenging and not challenging. As a result,

$$\mathbb{E}^{\rho}[q_\omega \mid \mathcal{I}_d] = 0 \qquad \text{and} \qquad \mathbb{E}^{\rho}[q_\omega \mid \mathcal{I}_{ch}] = q/(q + (1-q)\sigma_W)$$

Reasoning as in the main text, we can further conclude that σ_W must leave the leader of country D indifferent between conceding and resisting; thus,

$$\sigma_W = qz_D/(1-q)v \qquad \text{and} \qquad \mathbb{E}^{\rho}[q_\omega \mid \mathcal{I}_{ch}] = v/(v + z_D).$$ [17]

Thus, if the fraction of hawkish voters is neither too high, nor too low,

$$0 = \mathbb{E}^{\rho}[q_\omega \mid \mathcal{I}_d] < \mathbb{E}^{\rho}[q_\omega \mid \mathcal{I}_{ch}] = v/(v + z_D)$$

[17] Instead, if both types of C go to war, D concedes with probability 1 and the D1 criterion would select the equilibrium in which both types of C choose to challenge. And, if a leader does not challenge, then she is believed to be weak. In this case $\mathbb{E}^{\rho}[q_\omega \mid \mathcal{I}_d] = 0$, and $\mathbb{E}^{\rho}[q_\omega \mid \mathcal{I}_{ch}] = q$. Finally, if both types of C choose to back down, then D would resist and there would be a continuum of equilibria in which both types choose to challenge with the same probability. In this case, $\mathbb{E}^{\rho}[q_\omega \mid \mathcal{I}_d] = \mathbb{E}^{\rho}[q_\omega \mid \mathcal{I}_{ch}] = q$.

and (20) is lower if the leader backs down after a challenge than if she does not challenge at all provided that $\ell > 1$. In other words, the model with prospective voters still generates an audience cost due to the joint effect of reference dependence and loss aversion. In the presence of these two behavioral biases, we can define the sunk audience cost and the tying-hands audience cost as:

$$a_s^\rho = \beta\eta(\ell - 1)q\mathbb{E}^\rho[q_\omega \mid \mathcal{I}_{ch}]$$
$$a_t^\rho = \beta\big[\lambda(1+\eta) - (1-\lambda)(1+\eta\ell) + \eta(\ell-1)\mathbb{E}^\rho[q_\omega \mid \mathcal{I}_{ch}]\big]$$

(Recall from our discussion in Section 2.4 that the sunk audience cost is the payoff difference C's leader would get on top of what she would get in the workhorse model between backing down after a challenge and not challenging at the beginning of the game, while the tying-hands audience cost is the payoff difference between war and backing down.)

So even with prospective voters the sunk audience cost can be positive. Moreover, the cost is still increasing in the reference point, as measured by the probability of the leader being strong. Also notice that a_s^ρ is increasing in the weight put on the psychological component of the voters' utility function, η, in the degree of loss aversion, ℓ, and in the ex-ante probability with which politician are strong, q. Intuitively, if any of these parameters increases, then when hawks see their leader challenging the opponent, they will put high probability on her being strong and will be disappointed if she then backs down.

Furthermore, unlike the case of retrospective voting, the tying-hands audience cost (or benefit) is now an equilibrium quantity as well, since it depends on $\mathbb{E}^\rho[q_\omega \mid \mathcal{I}_{ch}]$. This happens because voters evaluate the election of the challenger also with respect to the reference point. Furthermore, the tying-hands audience cost is increasing in λ and can turn into a tying-hands audience benefit if λ is low enough.

Finally, if the fraction of hawks is either very high or very low, we can define the two types of audience costs (or audience benefits) in a similar way by using the reference points given in footnote 17.