When Capabilities Backfire Online Appendix

Peter Schram

January 11, 2021

Abstract

In Part I, I first discuss the complete equilibrium and offer a further discussion on the logic of predictability and emboldening (Section 1). I then offer proofs of Proposition 2-4 (Sections 2-4). In Section 5, I discuss a range of issues mentioned in the paper: why I only consider pure strategy equilibria, why I define a capability failure as I do, why these mechanisms are different from tying hands or burning bridges, and when predictability and can arise when the F function exhibits increasing differences. In Section 6, I consider a model where C faces costs to their transgressions. In Section 7 I discuss how the structure of bargaining between C and D in Stages 4 and 5 is fairly benign; to further justify these claims, in this section I include a model where, in the spirit of Gurantz and Hirsch (2017), I make this bargaining a black box, and I find that the results are essentially the same. In Part II, I consider models with convex set of types (θ). In Section 8, I include a sample model to illustrate a capability failure following improvements in hassling capabilities, and in Section 9, I utilize mechanism design to present general results on capability failures. Across all models and specifications, I find substantively similar results: that the "predictability" and "emboldening" mechanisms following improvements in public hassling capabilities (notably, not private hassling capabilities) described in the body of the paper are still the key drivers for a capability failure.

Contents

Part I

Discussion of the Example Model and Extensions.

1 In-Paper Model Equilibrium

1.1 Intuition

Assume that for a selected t, a type θ D optimally does not go to war. This type θ D selects an optimal h^* following

$$h^*(t) \in argmax_{h \in \mathbb{R}_+} \left\{ 1 - \rho - t + h + \kappa_C - \frac{(h)^2}{F(\alpha, \theta)} \right\}.$$

I take first-order conditions with respect to h of the expression above to identify the optimal level of hassling h^*. This unique value is

$$h^* = \frac{F(\alpha, \theta)}{2}.$$

Using h^*, I re-write D's utility in terms of the selected t and parameters α and θ. This is

$$U_D = \begin{cases} 1 - \rho - t + \kappa_C + \frac{F(\alpha,\theta)}{4} & if\, 1 - \rho - t + \kappa_C + \frac{F(\alpha,\theta)}{4} \geq 1 - \rho - \kappa_D, \\ 1 - \rho - \kappa_D & otherwise. \end{cases}$$

The top line is D's utility from hassling, and the bottom line is from going to war. For any t that C chooses, C will produce one of three outcomes: C selects a t that never invokes D to go to war, C selects a t that invokes types $\underline{\theta}$ to go to war, or C selects a t that invokes all types to go to war. When C selects a t such that no types D will go to war, C attains utility

$$EU_C = Pr(\underline{\theta}) * (\rho + t - h^*(\underline{\theta}) - \kappa_C) + Pr(\bar{\theta}) * \big(\rho + t - h^*(\bar{\theta}) - \kappa_C\big)$$

or

$$EU_C = \rho - \kappa_C + t - Pr(\underline{\theta})\frac{F(\alpha, \underline{\theta})}{2} - Pr(\bar{\theta})\frac{F(\alpha, \bar{\theta})}{2}. \tag{1}$$

When C selects a t such that only types $\underline{\theta}$ will go to war, C attains utility

$$EU_C = Pr(\underline{\theta}) * (\rho - \kappa_C) + Pr(\bar{\theta}) * \big(\rho + t - h^*(\bar{\theta}) - \kappa_C\big)$$

or

$$EU_C = \rho - \kappa_C + Pr(\bar{\theta})\left(t - \frac{F(\alpha, \bar{\theta})}{2}\right). \tag{2}$$

When C selects a t such that all types go to war, C attains utility

$$EU_C = \rho - \kappa_C. \tag{3}$$

There are two issues to note across (1)-(3). First, (1) and (2) are always weakly larger than (3); because I have assumed that $P(t^*, h^*) = \rho + t^* - h^* \in (\rho, 1)$, then I know that C can always do weakly better not going to war. Second, expressions (1) and (2) are both strictly increasing in t because there are no costs to transgressing; therefore, C is only ever constrained by an increased likelihood of war. Therefore, C will select a t that will either make a type $\underline{\theta}$ or a type $\bar{\theta}$ indifferent between war and hassling because at these points C incurs more war. Thus, C will select the value t that will make a type $\theta \in \{\underline{\theta}, \bar{\theta}\}$ indifferent between hassling and war, which is defined by condition

$$1 - \rho - t + \kappa_C + \frac{F(\alpha, \theta)}{4} = 1 - \rho - \kappa_D$$

or

$$t(\alpha, \theta) = \kappa_D + \kappa_C + \frac{F(\alpha, \theta)}{4}.$$

C's decision results in no war (when C benchmarks their t to make type $\underline{\theta}$ indifferent) or sometimes war (when C benchmarks their t to make type $\bar{\theta}$ indifferent) depending on C's willingness to sometimes risk war to achieve greater transgressions. I can identify C's expected utility from never going to war. Here C selects a t that makes a type $\underline{\theta}$ indifferent between war and hassling. This is $t(\alpha, \underline{\theta})$, which will give C an expected utility

$$EU_C = Pr(\underline{\theta}) * \left(\rho + \kappa_D + \kappa_C + \frac{F(\alpha, \underline{\theta})}{4} - \frac{F(\alpha, \underline{\theta})}{2} - \kappa_C\right) + Pr(\bar{\theta}) * \left(\rho + \kappa_D + \kappa_C + \frac{F(\alpha, \underline{\theta})}{4} - \frac{F(\alpha, \bar{\theta})}{2} - \kappa_C\right),$$

or

$$EU_C = \rho + Pr(\underline{\theta}) * \left(\kappa_D - \frac{F(\alpha, \underline{\theta})}{4}\right) + Pr(\bar{\theta}) * \left(\kappa_D + \frac{F(\alpha, \underline{\theta})}{4} - \frac{F(\alpha, \bar{\theta})}{2}\right).$$

I can also express D's utility, which differs based on type.

$$\begin{aligned} EU_D(\underline{\theta}) &= 1 - \rho - \kappa_D \\ EU_D(\bar{\theta}) &= 1 - \rho - \kappa_D + \frac{F(\alpha, \bar{\theta})}{4} - \frac{F(\alpha, \underline{\theta})}{4}. \end{aligned}$$

I can then identify C's expected utility from provoking types $\underline{\theta}$ to go to war. Here C selects a t that makes a type $\bar{\theta}$ indifferent between war and hassling (thus provoking types $\underline{\theta}$ to go to war). This is $t(\alpha, \bar{\theta})$ (which I will also sometimes write as $t(\bar{\theta}, \alpha)$ for $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$ when it is useful to do so), which will give C an expected utility

$$EU_C = Pr(\underline{\theta}) * (\rho - \kappa_C) + Pr(\bar{\theta}) * \left(\rho + \kappa_D - \frac{F(\alpha, \bar{\theta})}{4}\right)$$

or

$$EU_C = p + Pr(\underline{\theta}) * (-\kappa_C) + Pr(\bar{\theta}) * \left(\kappa_D - \frac{F(\alpha, \bar{\theta})}{4}\right).$$

I can also express D's expected utility.

$$\begin{aligned} EU_D(\underline{\theta}) &= 1 - \rho - \kappa_D \\ EU_D(\bar{\theta}) &= 1 - \rho - \kappa_D. \end{aligned}$$

1.2 Equilibrium

In the interest of saving space, I only include aspects not covered in Proposition 1. At the stage 2, C's beliefs on θ follow from its given distribution. At stage 5, on-the-path, C can perfectly identify D's type based on the selected h. Off-the-path, I assume for ease that C believes any h, x, or war decision outside of equilibrium play is made by a type $\underline{\theta}$ D.

Below is D's best response in response to t in Stage 3. A type θ parameter α D will set

$$w_D^* = \begin{cases} 0 & if\, \kappa_D + \kappa_C + \frac{F(\alpha,\theta)}{4} \geq t, \\ 1 & otherwise. \end{cases}$$

Regarding D's offer to C in Stage 4, D's best response is $x^* = \rho + t - h - \kappa_C$ for all t and h.

Regarding C's decision to go to war in Stage 5, I assume that, for a fixed t and h,

$$w_A^* = \begin{cases} 0 & if\, \rho + t^* - h^* - \kappa_C \leq x, \, and \\ 1 & otherwise. \end{cases}$$

C's utility varies across cases. When $Q(\alpha) \geq 0$ holds, C has expected utility

$$U_C = \rho + \kappa_D + Pr(\underline{\theta}) \left(-\frac{F(\alpha, \underline{\theta})}{4} \right) + Pr(\bar{\theta}) * \left(\frac{F(\alpha, \underline{\theta})}{4} - \frac{F(\alpha, \bar{\theta})}{2} \right).$$

When $Q(\alpha) < 0$ holds, C has expected utility

$$U_C = \rho + Pr(\underline{\theta}) \left(-\kappa_C \right) + Pr(\bar{\theta}) \left(\kappa_D - \frac{F(\alpha, \bar{\theta})}{4} \right).$$

1.3 C's Behavior within Predictability and Emboldening Mechanisms

The following figures further elaborate on C's decision making in the Predictability and Emboldening mechanisms. These figures are derived using the parameters described in Figures 1 and 2 in the text.

1.3.1 Predictability

Figure 3 illustrates more of what happens in the predictability example (partially described in Figure 1 in the text). When facing a parameter $\underline{\alpha}$ D (top graph), C's utility is increasing in t until it reaches $t(\underline{\alpha}, \underline{\theta})$.[1] C's utility is increasing in this range because, for the increases in t, D will not go to war and each type of D selects a fixed hassling level. Then, for any transgression such that $t > t(\underline{\alpha}, \underline{\theta})$, types $\underline{\theta}$ D will go to war. This creates the discontinuity at $t = t(\underline{\alpha}, \underline{\theta})$, as beyond that point C goes to war with probability $Pr(\underline{\theta})$. For the set of t's such

[1] While I only visualize $t \geq 0.575$, C's utility is also increasing for $0 \leq t \leq 0.575$.

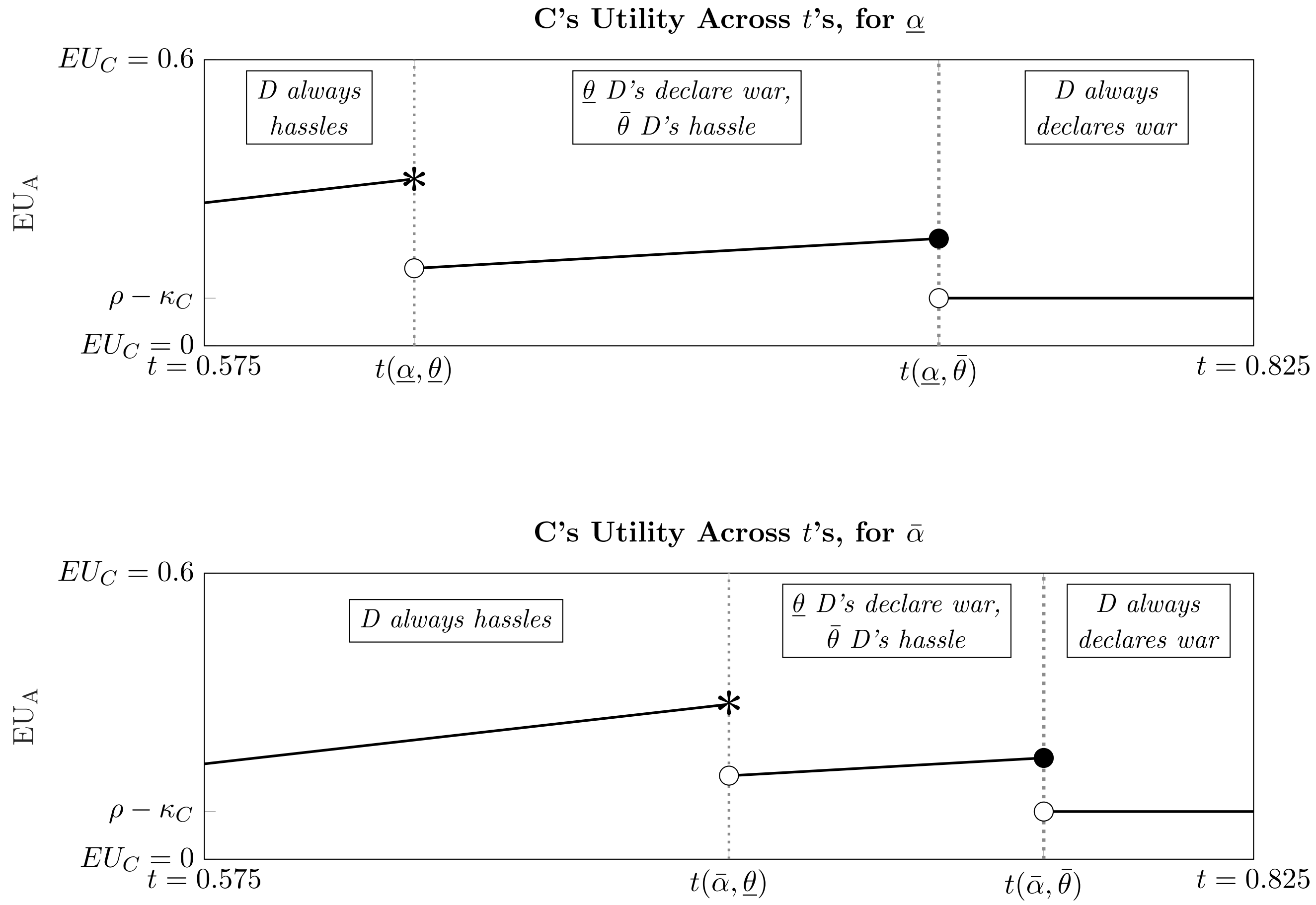

Figure 3. State C's Utilities Across Values of t (Predictability).

The parameter values are identical to those used in the Predictability example in the text. C's selected level of transgression under parameters $\underline{\alpha}$ and $\bar{\alpha}$ are denoted by the asterisks.

that $t(\underline{\alpha}, \underline{\theta}) < t \leq t(\underline{\alpha}, \bar{\theta})$, C's utility is once again increasing in t, then experiences another discontinuity at $t = t(\underline{\alpha}, \bar{\theta})$ as, for any $t > t(\underline{\alpha}, \bar{\theta})$, types $\bar{\theta}$ D's will go to war. Thus, if C selects a $t > t(\underline{\alpha}, \bar{\theta})$, then types $\underline{\theta}$ and types $\bar{\theta}$ D's will go to war, which grants C their wartime expected utility. For parameter $\underline{\alpha}$, selecting $t = t(\underline{\alpha}, \underline{\theta})$ gives C the greatest expected utility. Intuitively, in the text of the paper, I discussed the trade-off between increased transgression benefit and an increased risk of war. Here, C attains the greatest utility from selecting transgression level $t = t(\underline{\alpha}, \underline{\theta})$ and never going to war rather than selecting transgression level $t = t(\underline{\alpha}, \bar{\theta})$ and sometimes going to war.

An identical logic holds for parameter $\bar{\alpha}$.

1.3.2 Emboldening

Figure 4 illustrates more of what happens in the emboldening example (partially described in Figure 2 in the text). In the emboldening example – which, importantly, has different parameter values than the predictability example! – under the $\bar{\alpha}$ parameter, a new dynamic (relative to anything seen for predictability) occurs. Under parameter $\bar{\alpha}$, there is a very large space between $t(\bar{\alpha}, \underline{\theta})$ and $t(\bar{\alpha}, \bar{\theta})$, indicating that there is now a big difference between the selected t that would avoid war altogether ($t(\bar{\alpha}, \underline{\theta})$), and the greatest t that C could select that would only invoke type $\underline{\theta}$ D's to go to war ($t(\bar{\alpha}, \bar{\theta})$). This means that under $\bar{\alpha}$, C has more to gain by risking war with likelihood $Pr(\underline{\theta})$. In the emboldening case for $\underline{\alpha}$, while C's utility is increasing in t for $t(\underline{\alpha}, \underline{\theta}) < t \leq t(\underline{\alpha}, \bar{\theta})$, the greatest t in this set does not produce a utility for A that sufficiently offsets C's costs from having to go to war with type $\underline{\theta}$ D's. Due to the small difference between $t(\underline{\alpha}, \underline{\theta})$ and $t(\underline{\alpha}, \bar{\theta})$), ultimately resulting in C attaining a greater expected utility for selecting $t = t(\underline{\alpha}, \underline{\theta})$ and not going to war relative to selecting $t = t(\underline{\alpha}, \bar{\theta})$ and sometimes risking war. This is no longer the case under parameter $\bar{\alpha}$, as the greater distance between $t = t(\bar{\alpha}, \underline{\theta})$ and $t = t(\bar{\alpha}, \bar{\theta})$ now C can transgress more so long that C is willing to sometimes go to war. Referencing the terminology in the paper, under parameter $\underline{\alpha}$, the trade-off between selecting a greater level transgression and going to war more was not worthwhile; under parameter $\bar{\alpha}$, C can select a much greater level of t, which offers enough upside to make it worthwhile to go to war with types $\underline{\theta}$.

When C is faced with a parameter $\underline{\alpha}$ D, C will avoid war. As shown in the top two panels, C selects $t^* = t(\underline{\alpha}, \underline{\theta})$, which is characterized by the asterisks. As is illustrated, here type $\underline{\theta}$ D's are indifferent between war and hassling, and type $\bar{\theta}$ D's attain a surplus due to their private information. In these top two panels, it is worthwhile highlighting what C does not do; C could have selected $t(\underline{\alpha}, \bar{\theta})$.[2] This "move-not-taken" captures C's trade-off between greater transgression and more war. By selecting a greater t, C would do better when facing

[2] Note that C would never select a t greater than the value that would make a type $\bar{\theta}$ indifferent between hassling and war because this would always result in war, which is strictly worse for C.

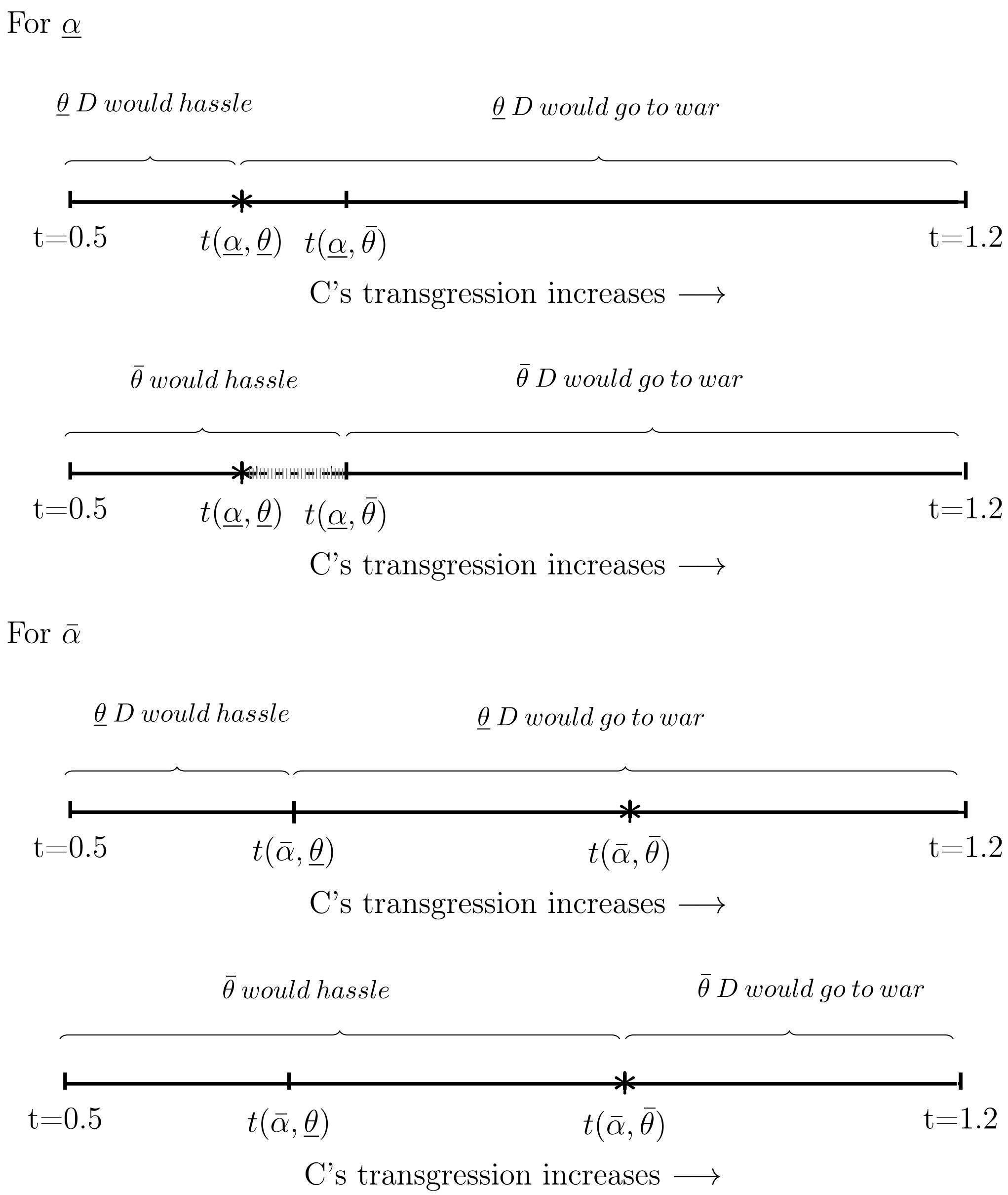
For $\underline{\alpha}$
$\underline{\theta}$ D would hassle
$\underline{\theta}$ D would go to war
t=0.5
$t(\underline{\alpha},\underline{\theta})$
$t(\underline{\alpha},\bar{\theta})$
t=1.2
C's transgression increases ⟶
$\bar{\theta}$ would hassle
$\bar{\theta}$ D would go to war
t=0.5
$t(\underline{\alpha},\underline{\theta})$
$t(\underline{\alpha},\bar{\theta})$
t=1.2
C's transgression increases ⟶
For $\bar{\alpha}$
$\underline{\theta}$ D would hassle
$\underline{\theta}$ D would go to war
t=0.5
$t(\bar{\alpha},\underline{\theta})$
$t(\bar{\alpha},\bar{\theta})$
t=1.2
C's transgression increases ⟶
$\bar{\theta}$ would hassle
$\bar{\theta}$ D would go to war
t=0.5
$t(\bar{\alpha},\underline{\theta})$
$t(\bar{\alpha},\bar{\theta})$
t=1.2
C's transgression increases ⟶

Figure 4. Optimal transgression and D's response (Emboldening).

C's selected transgression under parameters $\underline{\alpha}$ and $\bar{\alpha}$ are denoted by the asterisks. D's response to given t's are bracketed off. The dashed lines represent D's the surplus type $\bar{\theta}$ D's attain from their private information in equilibrium. The parameters are identical to those used in the Emboldening example in the text.

type $\bar{\theta}$ and not give these types as much a surplus. However, selecting $t(\underline{\alpha}, \bar{\theta})$ would provoke type $\underline{\theta}$ D's to go to war. Put in slightly more grounded terms, C's "downside" to selecting a $t(\underline{\alpha}, \bar{\theta})$ is that now all type $\underline{\theta}$ D's will go to war, increasing the likelihood of war from zero to the likelihood that D is type $\underline{\theta}$ (which here is $Pr(\underline{\theta}) = 0.2$). Also here, C's "upside" to selecting $t(\underline{\alpha}, \bar{\theta})$ is that C can select a greater t, as represented by the dashed line; when facing a type $\bar{\theta}$, this gives C a greater utility than C would get from the t that makes a type $\underline{\theta}$ indifferent.[3] Under parameter $\underline{\alpha}$, the upside of switching from $t(\underline{\alpha}, \underline{\theta})$ to $t(\underline{\alpha}, \bar{\theta})$ does not sufficiently outweigh the downside of invoking type $\underline{\theta}$ D's to go to war.

However, under parameter $\bar{\alpha}$, C faces a new trade-off. In the top two panels (for parameter $\underline{\alpha}$), there was not much upside to moving from $t(\underline{\alpha}, \underline{\theta})$ to $t(\underline{\alpha}, \bar{\theta})$ because there is not much difference between these t's. In the bottom two panels (for parameter $\bar{\alpha}$) there is now a large upside to moving from $t(\bar{\alpha}, \underline{\theta})$ to $t(\bar{\alpha}, \bar{\theta})$.[4] Under $\underline{\alpha}$, the small upside of a greater t did not offset the downside of having to go to war with all type $\underline{\theta}$ D's, and so C always avoided war. Under $\bar{\alpha}$, the greater upside sufficiently offset the downside of having to go to war with all type $\underline{\theta}$ D's, and so C now was willing to play more aggressively and sometimes risk war. Here, the improvement in D's public hassling capabilities offered C incentives to pursue a riskier transgression strategy and sometimes invoke war. Thus here, the improvements in α emboldened A in their transgressions.

1.4 Predictability and Poker

"In the whole range of human activities, war most closely resembles a game of cards" Clausewitz (1873, 86).

An alternate way to think about how better capabilities could lead to worse outcomes through the predictability channel is to consider a different game of private information: poker. In

[3] For the selected parameters, when C selects $t = t(\underline{\alpha}, \underline{\theta}) = 0.625$, C attains utility 0.375 when nature sets $\theta = \bar{\theta}$. When C selects $t = t(\underline{\alpha}, \bar{\theta}) = 0.675$, C attains utility 0.425 when nature sets $\theta = \bar{\theta}$.

[4] For the selected parameters, when C selects $t = t(\bar{\alpha}, \underline{\theta}) = 0.65$, C attains utility 0.1 when nature sets $\theta = \bar{\theta}$. When C selects $t = t(\bar{\alpha}, \bar{\theta}) = 0.825$, C attains utility 0.275 when nature sets $\theta = \bar{\theta}$.

a game of Texas hold'em poker, consider the following choice: a player can choose between drawing a pair of tens and this information would be private (only the player would know what cards they had), or they could draw a pair of aces and the player's opponents could see its hand. Faced with this choice, any serious poker player would take the pair of tens. This is not to say that two aces are a worse hand than the two tens: the odds favor the aces. But rather, because the player's opponents know how best to respond to the pair of aces, the player with two aces cannot capitalize on its better hand because they have lost the edge that private information gives them. Instead, a good poker player knows that with a pretty good hand (two tens) and private information, they could likely extract more value from the game.[5]

2 Proposition 2 Proof

Through the top two Predictability Conditions, across $\underline{\alpha}$ and $\bar{\alpha}$, the equilibrium is discussed under Case 1 in Proposition 1. The third condition defines

$$F(\underline{\alpha},\bar{\theta}) - F(\underline{\alpha},\underline{\theta}) > F(\bar{\alpha},\bar{\theta}) - F(\bar{\alpha},\underline{\theta}),$$

which is equivalent to

$$1-\rho-\kappa_D + \frac{F(\underline{\alpha},\bar{\theta})}{4} - \frac{F(\underline{\alpha},\underline{\theta})}{4} > 1-\rho-\kappa_D + \frac{F(\bar{\alpha},\bar{\theta})}{4} - \frac{F(\bar{\alpha},\underline{\theta})}{4}$$

where the left-hand side is type $\bar{\theta}$, parameter $\underline{\alpha}$ D's utility and the right hand side is type $\bar{\theta}$, parameter $\bar{\alpha}$ D's utility. Because across parameters $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$ types $\underline{\theta}$ attain their wartime utility, I am able to say that

$$U_D(\sigma^*(\underline{\theta},\underline{\alpha})) = U_D(\sigma^*(\underline{\theta},\bar{\alpha}))$$

[5] Some of this logic is also discussed in Gartzke (1999).

and

$$U_D(\sigma^*(\bar{\theta}, \underline{\alpha})) > U_D(\sigma^*(\bar{\theta}, \bar{\alpha})),$$

which is a capability failure.

3 Proposition 3 Proof

Through the two Emboldening Conditions, when $\alpha = \underline{\alpha}$, the equilibrium is discussed under Case 1 in Proposition 1, and when $\alpha = \bar{\alpha}$, the equilibrium is discussed under Case 2. Thus, when these two conditions hold, types $\underline{\theta}$ always attain their wartime utility.

Comparing the utilities that a capabilities $(\bar{\alpha}, \bar{\theta})$ attains (right hand side below) to the utility that a capabilities $(\underline{\alpha}, \bar{\theta})$ attain (left hand side below) is

$$1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > 1 - \rho - \kappa_D.$$

Thus, I can say

$$U_D(\sigma^*(\underline{\theta}, \underline{\alpha})) = U_D(\sigma^*(\underline{\theta}, \bar{\alpha}))$$

and

$$U_D(\sigma^*(\bar{\theta}, \underline{\alpha})) > U_D(\sigma^*(\bar{\theta}, \bar{\alpha})),$$

which is a capability failure.

4 Proposition 4 Proof

The “only if” was shown in Propositions 2 and 3.

When there is a capability failure following improvements in α, it must be that

$$U_D(\sigma^*(\underline{\theta}, \underline{\alpha})) \geq U_D(\sigma^*(\underline{\theta}, \bar{\alpha})) \tag{4}$$

and

$$U_D(\sigma^*(\bar{\theta}, \underline{\alpha})) \geq U_D(\sigma^*(\bar{\theta}, \bar{\alpha})), \tag{5}$$

with at least one inequality holding strictly. Proposition 1 shows, for a fixed α, one of two cases are possible: C selects either $t(\underline{\theta}, \alpha) = \kappa_D + \kappa_C + \frac{F(\alpha,\underline{\theta})}{4}$ (type $\underline{\theta}$ is made indifferent between hassling and war) or $t(\bar{\theta}, \alpha) = \kappa_D + \kappa_C + \frac{F(\alpha,\bar{\theta})}{4}$ (type $\bar{\theta}$ is made indifferent between hassling and war). This implies that there are four possible outcomes across $\underline{\alpha}$ and $\bar{\alpha}$, which I will designate by the selected t's: they are $(t(\underline{\alpha}, \underline{\theta}),\ t(\bar{\alpha}, \underline{\theta}))$, $(t(\underline{\alpha}, \bar{\theta}),\ t(\bar{\alpha}, \underline{\theta}))$, $(t(\underline{\alpha}, \underline{\theta}),\ t(\bar{\alpha}, \bar{\theta}))$, or $(t(\underline{\alpha}, \bar{\theta}),\ t(\bar{\alpha}, \bar{\theta}))$.

Outcome 1: $(t(\underline{\alpha}, \underline{\theta}),\ t(\bar{\alpha}, \underline{\theta}))$

When C selects $t(\underline{\alpha}, \underline{\theta})$ and $t(\bar{\alpha}, \underline{\theta})$, it implies that C does weakly better avoiding war for both $\underline{\alpha}$ and $\bar{\alpha}$. For C to behave this way, it must be that $Pr(\underline{\theta})\,(\kappa_C + \kappa_D) + \big(Pr(\bar{\theta}) - Pr(\underline{\theta})\big)\,\frac{F(\alpha,\underline{\theta})}{4} - Pr(\bar{\theta})\frac{F(\alpha,\bar{\theta})}{4} \geq 0$ for $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$. Thus, the first two parts of the Predictability Conditions hold. Furthermore, because there is a capability failure, I can substitute D's utilities from these cases into expressions (4) and (5) above, yielding

$$1 - \rho - \kappa_D \geq 1 - \rho - \kappa_D$$

$$1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} \geq 1 - \rho - \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4}.$$

Because types $\underline{\theta}$ receive the same utility across α's, the bottom inequality holds strictly when there is a capability failure. Re-arranging the bottom condition yields the final Predictability

Condition.

Outcome 2: $(t(\underline{\alpha}, \bar{\theta}),\ t(\bar{\alpha}, \underline{\theta}))$

When C selects $t(\underline{\alpha}, \bar{\theta})$ and $t(\bar{\alpha}, \underline{\theta})$, it implies that C does better sometimes going to war when facing a parameter $\underline{\alpha}$ and weakly better not going to war against a parameter $\bar{\alpha}$. By the conditions of the "if" clause (i.e. if there is a capability failure), this case is ruled out. As shown in Proposition 1, when C selects $t(\underline{\alpha}, \bar{\theta})$ and $t(\bar{\alpha}, \underline{\theta})$, capabilities $\underline{\alpha}$, $\bar{\theta}$ D attains utility $1 - \rho - K_D$ while capabilities $\bar{\alpha}, \bar{\theta}$ D attains utility $1 - \rho - \kappa_D + \frac{F(\alpha, \bar{\theta})}{4} - \frac{F(\alpha, \underline{\theta})}{4}$, which is strictly larger.

Outcome 3. $(t(\underline{\alpha}, \underline{\theta}),\ t(\bar{\alpha}, \bar{\theta}))$

When C selects $t(\underline{\alpha}, \underline{\theta})$ and $t(\bar{\alpha}, \bar{\theta})$, it implies that C does weakly better not going to war against a parameter $\underline{\alpha}$ and better sometimes going to war when facing a parameter $\bar{\alpha}$. For C to behave this way, it must be that $Pr(\underline{\theta})\left(\kappa_C + \kappa_D\right) + \left(Pr(\bar{\theta}) - Pr(\underline{\theta})\right) \frac{F(\underline{\alpha}, \underline{\theta})}{4} - Pr(\bar{\theta}) \frac{F(\underline{\alpha}, \bar{\theta})}{4} \geq 0$ and $Pr(\underline{\theta})\left(\kappa_C + \kappa_D\right) + \left(Pr(\bar{\theta}) - Pr(\underline{\theta})\right) \frac{F(\bar{\alpha}, \underline{\theta})}{4} - Pr(\bar{\theta}) \frac{F(\bar{\alpha}, \bar{\theta})}{4} < 0$. This meets the two Emboldening Conditions. Note that here a capability failure occurs within this case without any further conditions needed; I substitute D's utilities into expressions (4) and (5), yielding

$$1 - \rho - \kappa_D \geq 1 - \rho - \kappa_D$$

and

$$1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} \geq 1 - \rho - \kappa_D,$$

where the bottom inequality always holds strictly.

Outcome 4. $t(\underline{\alpha}, \bar{\theta}),\ t(\bar{\alpha}, \bar{\theta})$

When C selects $t(\underline{\alpha}, \bar{\theta})$ and $t(\bar{\alpha}, \underline{\theta})$, it implies that C does weakly sometimes going to war for both $\underline{\alpha}$ and $\bar{\alpha}$. When there is a capability failure, this case is ruled out. As shown in

Proposition 1, when C selects $t(\underline{\alpha}, \bar{\theta})$ and $t(\bar{\alpha}, \bar{\theta})$, all types and capabilities D attain their wartime utility $1-\rho-K_D$. Because across all hassling capabilities D attains the same utility, there cannot be a capability failure. □

5 Additional Results and Comments

5.1 On Only Considering Pure Strategy Equilibria

This paper only considers pure strategy Perfect Bayesian Nash Equilibria. Natually including mixed strategy equilbria does not rule out the equilibrium in the text, but it does introduce knife-edge cases where one player may mix over their actions in such a manner that does not "break" the equilibrium. For example, it could be that for a given set of parameters, A is indifferent between playing a t that sometimes risks war in Stage 3 and playing a t that induces all type D's to hassle in the Stage 3. In these edge-cases, A could mix over selecting $t(\alpha, \underline{\theta})$ and $t(\alpha, \bar{\theta})$. Alternatively, in the game, when A selects $t(\alpha, \underline{\theta})$, type $\underline{\theta}$ D's are indifferent between declaring war in the third stage and hassling in the third stage. With some small probability, the type $\underline{\theta}$ D's could sometimes go to war without breaking the equilibrium. Ultimately, these cases are not particularly interesting to the analysis and may even produce edge-cases where capability failures can or cannot occur through equilibrium selection among mixed strategy equilibria; for these reasons, they are excluded.

5.2 On the Definition of Capability Failure

In the paper, I formally define capability failure as the following:

Definition*: Improvements in publicly observed hassling capabilities (i.e. moving from $\underline{\alpha}$ to $\bar{\alpha}$ with $\underline{\alpha} < \bar{\alpha}$) produce a* ***capability failure*** *when, $U_D\left(\sigma^*(\theta, \underline{\alpha})\right) \geq U_D\left(\sigma^*(\theta, \bar{\alpha})\right)$ for all $\theta \in \Theta$ and $U_D\left(\sigma^*(\theta, \underline{\alpha})\right) > U_D\left(\sigma^*(\theta, \bar{\alpha})\right)$ for some $\theta \in \Theta$.*

By this definition, a capability failure occurs when all types of D perform weakly worse following an improvement in α, with at least some types doing strictly worse. As an alternative definition, I could have defined a capability failure as D attaining a lower ex-ante expected utility. I select the definition I do because I am interested in cases where improvements in hassling capabilities lead to overall worse outcomes for D. The definition that I adopt is stricter than the alternate-expected-utility definition, and the alternate definition (i.e. the expected utility definition) fails to exclude cases where some types of D attain a better outcome following a capability failure. As an example of this, one plausible outcome in the game is that an improvement in public hassling capability results in 20% less war (which is good for some types of D) but a 50% lower expected utility for some types of D (which is bad for those D's). This is a mixed outcome for D, and this could be classified as a capability failure under the weaker definition, but is never classified as a capability failure under the stricter definition.

Furthermore, it is not clear why calculating D's ex-ante expected utility is the right measure when, in all likelihood, D knows something about their private type upon entering a conflict. In the example above, if D knew they were likely a type that would avoid war following an improvement in public hassling capabilities, D would want the improvement in hassling capabilities to occur despite the weaker definition implying that an improved hassling capabilities would create a capability failure. Thus, to exclude illogical cases where a partially informed D would actually want a capability failure to occur, I adopt the stricter definition of capability failure for my analysis.[6]

5.3 On Burning Bridges and Tying Hands

The mechanisms for emboldening and predictability through improved hassling capabilities (as described in the paper) are different from the logic of burning bridges or tying hands by restricting low-level conflict (or hassling) options. Additionally, the logic of burning bridges or tying hands are not observed in my model because I assume a formal framework where the

[6] The stricter definition of capability failure that I use also satisfies the min-max heuristic.

challenger can select from a range of possible transgressions $t \geq 0$ rather than a dichotomous transgress-or-not framework (something like $t \in \{0, \bar{t}\}$).[7] I go into more detail on that here. First, I describe the logic of burning bridges, as characterized Schelling (1966, pp. 43-44). Then I discuss one simple formalization of this logic though a version of a burning-bridges game characterized in Spaniel (2014, pp. 135-138), where the decision to transgress is a discrete choice. Then I show that adding a broader action set of possible transgressions—using a similar modeling technology to what I have in my model—eliminates the effect within the Spaniel-inspired game. Next, I describe why the logic of burning bridges as presented in Schelling and Spaniel is not present in my model. Then, I describe other interpretations of the concept of burning bridges, and why these other interpretations do not apply to the case that I am considering. Finally, I repeat the process for tying hands in the context of a complete-information, fully parameterized version of Fearon (1997), though in a more abbreviated fashion.

5.3.1 Burning Bridges

Now I describe how the capability failures following improvements in hassling capabilities are not equivalent to failing to adequately "burn one's bridges." Schelling (1966, pp. 43-44) describes the logic of burning bridges elegantly, while also elaborating on how the logic of burning bridges can apply to having multiple policy choices . While the concept is mentioned in several points in his work, to the best of my knowledge, below is the most expansive articulation of the concept:

> *"Often we must maneuver into a position where we no longer have much choice left. This is the old business of burning bridges. If you are faced with an enemy who thinks you would turn and run if he kept advancing, and if the bridge is there to run across, he may keep advancing. He may advance to the point where,*

[7] It is also relevant that in my game "burning bridges" does not reduce the defender's choice set, but rather makes some choices worse.

if you do not run, a clash is automatic. Calculating what is in your long-run interest, you may turn and cross the bridge. At least, he may expect you to. But if you burn the bridge so that you cannot retreat, and in sheer desperation there is nothing you can do but defend yourself, he has a new calculation to make. He cannot count on what you would prefer to do if he were advancing irresistibly; he must decide instead what he ought to do if you were incapable of anything but resisting him.[...]

"This idea of burning bridges—of maneuvering into a position where one clearly cannot yield—conflicts somewhat, at least semantically, with the notion that what we want in our foreign policy is "the initiative." Initiative is good if it means imaginativeness, boldness, new ideas. But the term somewhat disguises the fact that deterrence, particularly deterrence of anything less than mortal assault on the United States, often depends on getting into a position where the initiative is up to the enemy and it is he who has to make the awful decision to proceed to a clash.

"In recent years it has become something of a principle in the Department of Defense that the country should have abundant "options" in its choice of response to enemy moves. The principle is a good one, but so is a contrary principle—that certain options are an embarrassment. The United States government goes to great lengths to reassure allies and to warn Russians that it has eschewed certain options altogether, or to demonstrate that it could not afford them or has placed them out of reach. The commitment process on which all American overseas deterrence depends—and on which all confidence within the alliance depends—is a process of surrendering and destroying options that we might have been expected to find too attractive in an emergency. We not only give them up in exchange for commitments to us by our allies; we give them up on our own account to make

> *our intentions clear to potential enemies. In fact, we do it not just to display our intentions but to adopt those intentions. If deterrence fails it is usually because someone thought he saw an "option" that the American government had failed to dispose of, a loophole that it hadn't closed against itself."*

Schelling here describes value to destroying options that are too attractive in an emergency. In the context of hassling capabilities, hassling (or any low-level conflict options) could represent a more attractive alternative to war; instead of committing to a costly and devastating military response to handle a transgression, a defender could hassle. Thus, by not developing one's own hassling capabilities, the defender could "burn its bridges" and be able to commit to the more devastating response.

How could this be formalized? Figure 1 is one such way, which is adapted from Spaniel (2014, pp. 136-138). The game order is as follows. First, a defender can "Burn" their bridges, or not (∼Burn). Then, a challenger can transgress ($t = 1$) or not ($t = 0$). If the challenger does not transgress, the game ends. Finally, if the challenger transgressed, the defender can choose whether to fight ($f = 1$) or retreat ($f = 0$). D's payoffs are listed first. To put this in the context of hassling, the decision to burn bridges can be thought of as weakening (but not altogether eliminating) one's ability to conduct hassling. When D does not select an all-out war ($f = 1$) but rather selects "retreat" ($f = 0$), so long that D has not burned their bridges, D can still use their hassling/low-level conflict options. In the model, D does worse when retreating when D has crippled their own hassling abilities.

In the original Spaniel model, burning the bridge eliminates the "retreat" option within the "Burn" subgame. As I model it, burning the bridge makes the retreat option produce a lower payoff. Why do I make this change? Practically, there always exists *some* form of hassling that could be implemented. If a state has underdeveloped low-level/hassling capabilities (i.e. they have burned their bridges), then a low-level response is still available, it's just inefficient or, in Schelling's terms, an "embarrassment" of an option. For example, if Israel didn't have

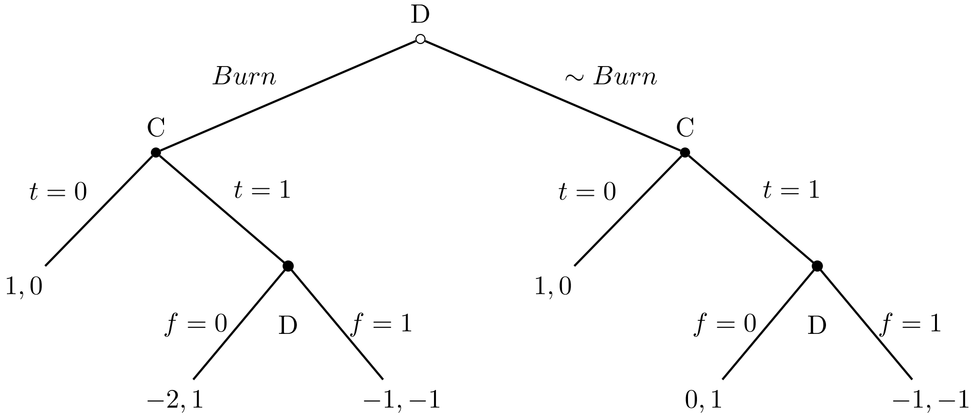

Figure 1: Burning Bridges, Discrete t (adapted from Spaniel)

effective cybercapabilities, it still could have conducted something like Operation Outside the Box, but that option would have likely needed a larger footprint and would have come with a greater chance of a costly or embarrassing failure. Additionally, technically speaking, not reducing the action set in the Figure 1 model is closer to the model in the text while still displaying the logic of burning bridges.[8]

The payoffs here represent the simple logic that by not burning ones bridges (by having more effective hassling options), retreating (hassling) can be too appealing an option. In more grounded terms, the ($\sim$ Burn, $t = 1$, $f = 0$) set of play describes the case where the defender hasn't disabled its appealing hassling options, the challenger transgresses, and the defender retreats/hassles. Intuitively, here the retreating defender will not incur too many costs from retreating because the defender still has their bridges. This could be thought of as the defender using hassling as a face-saving measure which, in the $\sim$ Burn subgame, is preferred to fighting. The (Burn, $t = 1$, $f = 0$) set of play describes the identical case, but where the defender *has* burned their bridges. Now the defender's final choice is between fighting and a messy, ineffective retreat. Without effective hassling options, the defender

[8] Finally, and this is an issue for the model in Figure 2 with $t \geq 0$, if burning bridges is viewed as eliminating the $f = 0$ option altogether, then what should be made of how D would respond to very small transgressions? Additional modeling technology would need to be added to the subgame with burned-bridges to prevent the illogical feature where, following any minuscule transgression, D can commit to war.

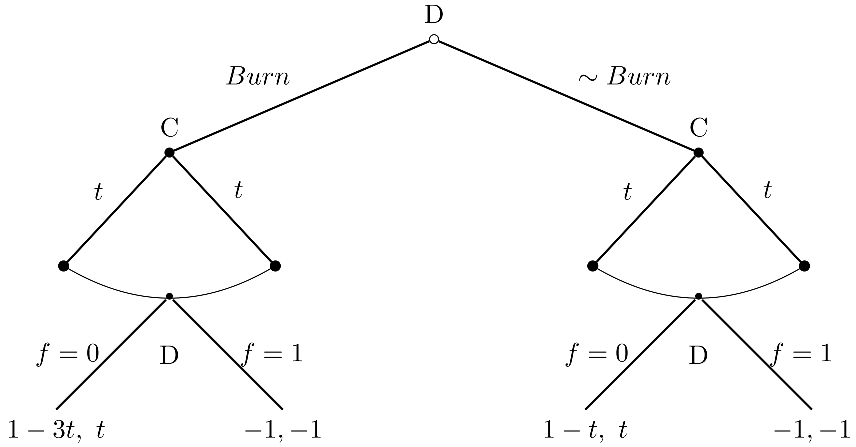

Figure 2: Burning Bridges, $t \geq 0$

cannot save face through hassling and prefers to fight.[9]

Under subgame perfection, the equilibrium play is straightforward. If the defender did not burn their bridges, the challenger would transgress, knowing that the defender could not credibly commit to fighting over fleeing. As a result, the defender prefers to burn bridges, knowing it enhances the credibility of the "fight" threat later on. Thus, in equilibrium, the defender will burn bridges (Burn), and the challenger will not transgress ($t = 0$). The value of burning bridges is borne out here in the complete information environment above. If the defender didn't burn their bridges, they would attain a payoff of 0. By burning bridges, the defender attains a payoff of 1.

Consider now a simple modification to the model in Figure 1 where the challenger no longer selects from $t \in \{0, 1\}$, but rather can select any $t \geq 0$. Practically speaking, in the cases described in the text, the challenger likely has many different transgressions at their disposal rather than just one. The new model is presented in Figure 2.

The model in Figure 2 is faithful to the discrete-choice version payoffs when $t = 0$ and $t = 1$.

[9] For interested readers, to offer a simple formal foundation for these payoffs, treat C and D in a crisis over an asset of size 1. When C does not transgress ($t = 0$), D gets the entirety of the asset (1 util) and C gets nothing (0 utils). When C transgresses ($t = 1$) and D fights ($f = 1$), they go to war, which delivers each of them a final expected payoff of -1 because war is costly. When C transgresses ($t = 1$) and D retreats ($f = 0$), C gets the asset (1 util), while D's payoffs depend on if D had the bridge to flee across (which makes retreating have zero cost) or if D does not have the bridge to flee across (which gives D -2 utils).

The key difference in this model is that C can select any transgressions level $t \geq 0$. As some intuition, $t = 2$ ($t = 0.5$) means that C is undertaking a more aggressive (less aggressive) transgression than $t = 1$, and therefore attains more (less) surplus when D does not fight. Additionally, this model makes clear what bridges do for transgressions: as modeled, when the challenger transgresses and the defender retreats while lacking bridges, the transgressions hurt the defender more, which justifies the further $-2t$ cost term within the "Burn" subgame. As intuition, there is nothing the defender can do to save face without bridges (or hassling options) here, so an unmet transgression hurts more.

In this modified game (with $t \geq 0$), there are no longer any benefits to burning bridges. If D burns their bridges, C will select $t = \frac{2}{3}$, which will make D indifferent between retreating and fighting, and so D will retreat. This gives D a final payoff of -1. If D does not burn their bridges, C will select $t = 2$, which will also result in D retreating, also giving D a final payoff of -1. In other words, when C can calibrate their selected t, they will always select the transgression that makes D indifferent between retreating or fighting because this maximizes C's payoffs. Thus, there is no value to restricting hassling capabilities when the challenger can flexibly choose their transgression level, because the defender will always attain their "fight" utility, which does not change with the presence of bridges.

The discussion in the last paragraph echoes what is presented in Corollary 3. In the model in the paper, when $\underline{\theta} = \bar{\theta}$, then C can always select the calibrated t to give D their wartime payoff. As a result, in a complete information version of my model (in the paper), a capability failure can never occur. For a capabilities failure to occur in my model, private information must play a role, an assumption that is absent from the Schelling (1966) discussion on the topic. Thus, I am considering a different strategic environment for the burning bridges mechanism to function (I consider $t \geq 0$), and the mechanisms I present in the text requires uncertainty over the defender's willingness to hassle. Finally, note that in the discrete-choice game form above in Figure 1, war never occurs when the defender fails to burn their bridges (i.e. have strong hassling capabilities). This highlights that failing to burn bridges is empirically distinct from what is described in emboldening, where improvements in hassling capabilities results in a greater likelihood of war. This empirical distinction is particularly important for the case study on the lead-up to the 2003 Iraq invasion, where it is possible that the U.S.'s improved hassling capabilities altered Saddam's strategic calculus in a way that led to war; in the simple model above, the failure to burn bridges could lead to more low-level conflict, but not an escalation to war.

Of course, I do not want to claim that burning bridges can *never* function in a model with a multiple transgression options and information asymmetry. Take, for example, Baliga *et al.* (2020), which has a different form of information asymmetry that what is considered in my model (Baliga *et al.* assumes it is difficult to identify who attacked). While they do not explicitly consider a model with both multiple transgressions (or their equivalent, "attacks") and multiple defender response options, they could still find the result that a defender who is better at low-level conflict could do worse when the challenger has many transgression options.[10] Why? In Baliga *et al.*, the defender wants to use low-level conflict options when the defender receives an ambiguous signal about who transgressed and wants to prevent an unjustified aggressive retaliation. Unless more modifications are made to their model (beyond expanding the set of possible transgressions), there would still be ambiguity over who conducted the transgression, and the defender could still be temped to scale back their retaliations when they have the low-level response option. In other words, the kind of ambiguity explored in Baliga *et al.* allows the burning bridges mechanism to still function with uncertainty and a dichotomous transgression option.

There are other versions of how burning bridges function too. For example Pecorino (2019)

[10] Baliga *et al.* describes how, when the defender possesses several possible low-level responses, the defender can do worse. Their logic, which is very much in-line with the burning bridges logic, is as follows: "The intuition is that, when a weaker weapon is available, ex post the defender is sometimes tempted to use it rather than the stronger weapon (in particular, when she is uncertain of the identify of the perpetrator). This is bad for ex ante deterrence. The defender can thus benefit from committing in advance to never retaliate with a less destructive weapon."

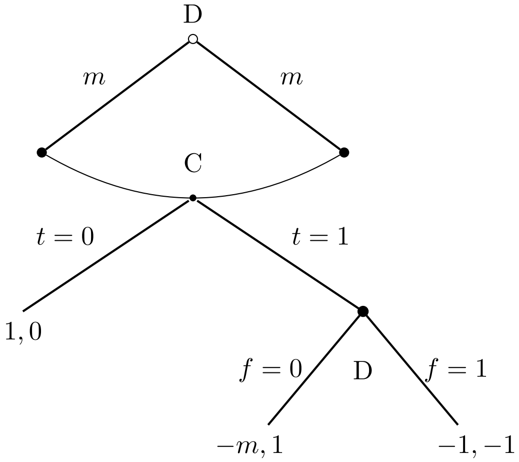

Figure 3: Tying Hands, Discrete t (from Fearon (1997))

presents an instance where, once bridges are burned, the defender works harder to secure wartime victory. This interpretation is common in the literature on entry deterrence (see Tirole (1988, pp. 314-359) for a discussion). Instead of treating "war" as a fixed endeavor, as is done in the models above and in the paper, Pecorino considers war as a contest success function with costly effort, where actors can endogenously make their defeat outcome worse for them by burning bridges. This model does not quite describe the cases of interest here, nor does Pecorino claim it does. Hassling in this paper is treated as an alternative action to war for dealing with a transgressing challenger. In Pecorino, burning bridges effects what happens when a state loses.

5.3.2 Tying Hands

Now I describe how the capability failures following improvements in hassling capabilities are not equivalent to failing to adequately "tie one's hands." Figure 3 presents a parameterized version of the complete-information "tying hands" case as described in Fearon (1997). In it, a defender can first select $m \geq 0$. Next, the challenger can transgress ($t = 1$) or not ($t = 0$). Finally, the defender can respond by fighting ($f = 1$) or retreating ($f = 0$). To empirically ground the moves in the model within that case that I'm considering, a smaller m means that D has more and more productive hassling options available, which makes the

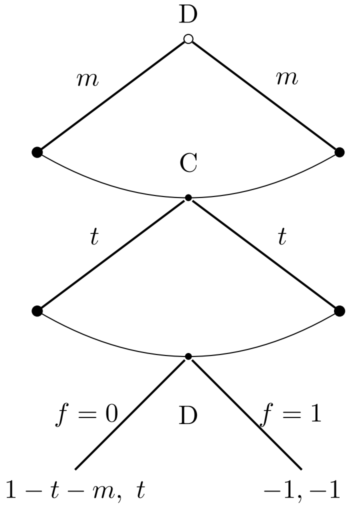

Figure 4: Tying Hands, $t \geq 0$

final decision to not use all-out war (i.e. to select $f = 0$) more appealing.

The structure for this game (Figure 3) is similar to the game in Figure 1, only now there is the endogenous m action which influences the $-m$ term for when the challenger transgresses and the defender retreats. When m is high, the defender has effectively tied their hands against retreating. In terms of hassling conflict, if one effectively ties their hands against using hassling (by, for example, being ineffective at it), then the defender may prefer to resort to going to war when facing a transgression.

When t is discrete (as it is in Figure 3), a class of pure-strategy subgame perfect equilibria exist taking the form where D selects some $m > 1$, which makes retreating worse than war for the defender. Thus, when the challenger observes some $m > 1$, the challenger knows that their transgression would be met with a fight, and the challenger prefers to not transgress.

However, when transgressions can be flexible (with $t \geq 0$), tying hands loses its value. This is the model in Figure 4.

For the model in Figure 4, there are two classes of pure-strategy subgame equilibria. First, whenever D selects an m such that $m \in [0, 2]$, C will select $t = 2 - m$. This selection makes D indifferent between $f = 0$ and $f = 1$, and so D will retreat and attain payoff -1. Second, whenever D selects an m such that $m > 2$, C will select any $t \geq 0$, D will fight and attain payoff -1. To summarize, regardless of the class of subgame perfect equilibria, D attains the same payoff of -1.

I have shown above that the tying hands result (as demonstrated in the model in Figure 3) can be eliminated through the move from a discrete transgression choice to a flexible range of transgression levels. Additionally, the logic of tying hands in Figure 3 is borne out in a complete information game: by keeping hassling options available (i.e. by selecting a low m), the possibility of using the available hassling options ($f = 0$) is too tempting, war ($f = 1$) is undermined as a deterrent threat, and C suffers. Tying hands is different from the predictability or emboldening cases as described in the paper, which, among other reasons, rely on information asymmetry, and, in the case of emboldening, have the feature where better hassling capabilities can lead to more war and not just more hassling.

5.4 When Increasing Differences in F Creates a Capabilities Failure Via Predictability

Proposition 2 describes a capabilities failure following from decreasing differences in the F function (as well as the conditions on Q holding). As I show here, the decreasing differences condition is not a necessary condition to predictability. Let the game form remain the same and consider the new outcome utilities where C's selected transgression affects the costs of hassling. As intuition, more aggressive transgressions may be harder to roll-back:

Scenario	**C's utility**	**D's utility**
D initiates war at Stage 3 (before h and t are realized)	$P(0,0)-\kappa_A$	$1-P(0,0)-\kappa_D$
C initiates war at Stage 5 (after h and t are realized)	$P(t,h)-\kappa_A$	$1-P(t,h)-\kappa_D-\frac{th^2}{F(\alpha,\theta)}$
C accepts at Stage 5 (after h and t are realized)	x	$1-x-\frac{th^2}{F(\alpha,\theta)}$

Table 1: Summarized payoffs for the new Appendix Model.

Solving the game, following the logic from earlier, D's optimization is

$$h^*(t) \in argmax_{h\in\mathbb{R}_+}\left\{1-\rho-t+h+\kappa_A-\frac{h^2t}{F(\alpha,\theta)}\right\}.$$

I take first-order conditions with respect to h of the expression above to identify the optimal level of hassling h^*. This unique value is

$$h^* = \frac{F(\alpha,\theta)}{2t}.$$

I can then say

$$U_D = \begin{cases} 1-\rho-t+\kappa_A+\frac{F(\alpha,\theta)}{4t} & if\, 1-\rho-t+\kappa_A+\frac{F(\alpha,\theta)}{4t} \geq 1-\rho-\kappa_D, \\ 1-\rho-\kappa_D & otherwise. \end{cases}$$

This means I can solve for the t that would make a type θ indifferent between war and hassling.

$$t^2 - t(\kappa_A+\kappa_D) - \frac{F(\alpha,\theta)}{4} = 0.$$

One solution to the quadratic is

$$t = \frac{\kappa_A+\kappa_D+\left((\kappa_A+\kappa_D)^2+F(\alpha,\theta)\right)^{1/2}}{2}.$$

Assume C's costs of war are sufficently high so that C always avoids war. C will select a t benchmarked against the lowest type of D. This will give D utilities

$$\begin{aligned}
U_D(\alpha,\underline{\theta}) =& 1-\rho-\kappa_D \\
U_D(\alpha,\bar{\theta}) =& 1-\rho-\left(\frac{\kappa_A+\kappa_D+\left((\kappa_A+\kappa_D)^2+F(\alpha,\underline{\theta})\right)^{1/2}}{2}\right)+\kappa_A+\frac{F(\alpha,\bar{\theta})}{4}\left(\frac{\kappa_A+\kappa_D+\left((\kappa_A+\kappa_D)^2+F(\alpha,\underline{\theta})\right)^{1/2}}{2}\right)^{-1}
\end{aligned}$$

When $U(\alpha,\bar{\theta})$ is decreasing in α, then D experiences a capabilities failure. One set of parameters where this can occur is $\kappa_D = 0.2$, $\kappa_A = 2$, $p = 0.5$, $Pr(\underline{\theta}) = 0.5$, $F(\underline{\alpha},\underline{\theta}) = 2$, $F(\underline{\alpha},\bar{\theta}) = 3$, $F(\bar{\alpha},\underline{\theta}) = 5$ and $F(\bar{\alpha},\bar{\theta}) = 6.1$. Note that $F(\alpha,\theta)$ here exhibits increasing differences. D's utilities are $U_D(\underline{\alpha},\underline{\theta}) = 0.3$, $U_D(\underline{\alpha},\bar{\theta}) = 0.4038$, $U_D(\bar{\alpha},\underline{\theta}) = 0.3$, and $U_D(\bar{\alpha},\bar{\theta}) = 0.4031$.

6 Example Model With Costs from Investing in t

In this section, I introduce a model with costs. I find, similar to the model in the text, that the only way for a capability failure to arise is through the emboldening or predictability mechanisms.

When D becomes more predictable, there is (weakly) less war across all types, and high type D's (types $\theta > \underline{\theta}$) attain less value from their private information. When C becomes emboldened, C is motivated to select a new, more aggressive t^* that diminishes high type D's utility and there is more war. Both circumstances can still arise here, though the conditions are more expansive than in the case in the text.

6.1 Model Assumptions

I modify the model in the text. The key distinction here is that I assume that selecting transgression $t \in \mathbb{R}_+$ comes with cost $-\kappa_t t^2$ for C. These costs are realized for C whenever war does not occur in Stage 3.

As it was for the model in the text, here I also assume that in equilibrium the final value of the P function and final offer have the properties $P \in (\rho, 1)$ and $x \in (0, 1)$, which prevent any kinks or constraints from driving the results.

Note, from the assumptions above, it rules out the possibility that $t = 0$. When C does not want to transgress, this falls outside the scope of the analysis. It also rules out the knife-edge cases where $t^* = h^*$, which do not matter much for the analysis.

6.2 Equilibrium Preliminaries, Intuition, and Equilibrium

As it was for the model in the text, through the assumptions above, I can describe equilibrium behavior in a relatively straightforward manner. Before discussing how C plays the game, note that D selects h^*, w_D^*, and x^* identical to how D played in the model where C faced no costs.

In the model in the text, C restricted their selected t out of fear of more war with D. This led to C choosing between $t(\alpha, \underline{\theta})$ and $t(\alpha, \bar{\theta})$, the levels of transgression that would make a low-type ($\underline{\theta}$) and a high-type ($\bar{\theta}$) indifferent between hassling and war (respectively). These two values are still relevant, but here C may be constrained by their costs of war. For now assume that C's optimal selection of t is weakly less than $t(\alpha, \underline{\theta})$ – in other words, C wants to avoid war and the cost to the transgression are binding (I discuss other cases next). Then C is selecting a t^* defined by

$$t^* \in argmax \left\{ Pr(\underline{\theta}) * \left(\rho + t - \frac{F(\alpha, \underline{\theta})}{2} - \kappa_C - \kappa_t t^2 \right) + Pr(\bar{\theta}) * \left(\rho + t - \frac{F(\alpha, \bar{\theta})}{2} - \kappa_C - \kappa_t t^2 \right) \right\}.$$

The solution to this optimization is $t^* = \frac{1}{2\kappa_t}$. Now assume C's optimal selection of t such that $t \in (t(\alpha, \underline{\theta}), t(\alpha, \bar{\theta}))$ – in other words, C is willing to sometimes go to war (and select a $t > t(\alpha, \underline{\theta})$), but the costs to the transgression are still binding. Then C is selecting a t^*

defined by

$$t^* \in argmax\left\{Pr(\underline{\theta}) * (\rho - \kappa_C) + Pr(\bar{\theta}) * \left(\rho + t - \frac{F(\alpha,\bar{\theta})}{2} - \kappa_C - \kappa_t t^2\right)\right\}.$$

The solution to this optimization is also $t^* = \frac{1}{2\kappa_t}$. Thus, I define a new constant: $\hat{t} = \frac{1}{2\kappa_t}$ which is, when C's cost constraints bind, how far C is willing to go in their selection of t. Here C will choose between the values $\hat{t}$, $t(\alpha,\underline{\theta})$, and $t(\alpha,\bar{\theta})$. To better characterize C's choice, I define C's utilities as a function of their selected t and, in the case of $\hat{t}$, whether it ever provokes a type $\underline{\theta}$ to go to war (which occurs when $\hat{t}$ is selected and $\hat{t} > t(\alpha,\underline{\theta})$). For a fixed α, these are

$$U_C(\hat{t}, peace) = Pr(\underline{\theta})\left(\rho + \frac{1}{4\kappa_t} - \frac{F(\alpha,\underline{\theta})}{2} - \kappa_C\right) + Pr(\bar{\theta})\left(\rho + \frac{1}{4\kappa_t} - \frac{F(\alpha,\bar{\theta})}{2} - \kappa_C\right),$$

$$U_C(t(\alpha,\underline{\theta})) = -\kappa_t\left(\kappa_D + \kappa_C + \frac{F(\alpha,\underline{\theta})}{4}\right)^2 + Pr(\underline{\theta}) * \left(\rho + \kappa_D - \frac{F(\alpha,\underline{\theta})}{4}\right) + Pr(\bar{\theta}) * \left(\rho + \kappa_D + \frac{F(\alpha,\underline{\theta})}{4} - \frac{F(\alpha,\bar{\theta})}{2}\right),$$

$$U_C(\hat{t}, war) = Pr(\underline{\theta}) * (\rho - \kappa_C) + Pr(\bar{\theta}) * \left(\rho - \frac{F(\alpha,\bar{\theta})}{2} - \kappa_C + \frac{1}{4\kappa_t}\right),$$

$$U_C(t(\alpha,\bar{\theta})) = Pr(\underline{\theta})(\rho - \kappa_C) + Pr(\bar{\theta})\left(\rho + \kappa_D - \frac{F(\alpha,\bar{\theta})}{4}\right) - \kappa_t\left(\kappa_D + \kappa_C + \frac{F(\alpha,\bar{\theta})}{4}\right)^2.$$

The top value is the case where the costs bind, $t^* = \hat{t}$, and $\hat{t} < t(\alpha,\underline{\theta})$; in this case, war never occurs. The second value is when $t^* = t(\alpha,\underline{\theta})$. The third value is when the cost constraints bind, $t^* = \hat{t}$, and $\hat{t} \in (t(\alpha,\underline{\theta}), t(\alpha,\bar{\theta}))$; in this case, war sometimes occurs. The last value is when $t^* = t(\alpha,\bar{\theta})$.

With this in place, I can define the selected equilibria to the game. Proposition 5 describes five cases. These cases rely on (a) where the value $\hat{t}$ lies relative to $t(\alpha,\underline{\theta})$ and $t(\alpha,\bar{\theta})$, and (b) some comparison of C's expected utility across some subset $t \in \{\hat{t}, t(\alpha,\underline{\theta}), t(\alpha,\bar{\theta})\}$. Why does (a) matter? For example, when $t(\alpha,\bar{\theta})$ is greater than $\hat{t}$, then C would never play $t(\alpha,\bar{\theta})$ (with the same logic for $t(\alpha,\underline{\theta}) > \hat{t}$ and C never playing $t(\alpha,\underline{\theta})$). Why does (b) matter? Once (a) narrows the possible range of actions, then (b) compares C's utilities across possible selected actions.

Proposition 5: *Under the assumptions above, for a fixed $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$, the following actions are part of the perfect Bayesian Nash Equilibrium.*

Case 1. When $\hat{t} \leq t(\alpha, \underline{\theta})$ holds, C avoids war:

- *C's beliefs on θ follow its expected distribution. C selects transgression level*

$$t^* = \frac{1}{2\kappa_t}.$$

For the selected t^*, a type $\theta \in \{\underline{\theta}, \bar{\theta}\}$ D will always hassle, setting $h^* = \frac{F(\alpha,\theta)}{2}$. Each type D has utility

$$U_D(\sigma^*(\underline{\theta}, \alpha)) = 1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\alpha, \underline{\theta})}{4},$$
$$U_D(\sigma^*(\bar{\theta}, \alpha)) = 1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\alpha, \bar{\theta})}{4}.$$

C has expected utility

$$EU_C = Pr(\underline{\theta}) \left(\rho + \frac{1}{4\kappa_t} - \frac{F(\alpha, \underline{\theta})}{2} - \kappa_C\right) + Pr(\bar{\theta}) \left(\rho + \frac{1}{4\kappa_t} - \frac{F(\alpha, \bar{\theta})}{2} - \kappa_C\right).$$

Case 2. When $\hat{t} \in \big(t(\alpha, \underline{\theta}), t(\alpha, \bar{\theta})\big)$ and $U_C(\hat{t}, war) \leq U_C(t(\alpha, \underline{\theta}))$, C avoids war:

- *C's beliefs on θ follows its expected distribution. C selects transgression level*

$$t^* = \kappa_D + \kappa_C + \frac{F(\alpha, \underline{\theta})}{4}.$$

For the selected t^*, a type $\theta \in \{\underline{\theta}, \bar{\theta}\}$ D will always hassle, setting $h^* = \frac{F(\alpha,\theta)}{2}$. Each *type D has utility*

$$
\begin{aligned}
U_D(\sigma^*(\underline{\theta},\alpha)) =& 1-\rho-\kappa_D,\\
U_D(\sigma^*(\bar{\theta},\alpha)) =& 1-\rho-\kappa_D+\frac{F(\alpha,\bar{\theta})}{4}-\frac{F(\alpha,\underline{\theta})}{4}.
\end{aligned}
$$

- *C has expected utility*

$$
EU_C = \rho+\kappa_D-\kappa_t\left(\kappa_D+\kappa_C+\frac{F(\alpha,\underline{\theta})}{4}\right)^2 - Pr(\underline{\theta})\left(\frac{F(\alpha,\underline{\theta})}{4}\right)+Pr(\bar{\theta})\left(\frac{F(\alpha,\underline{\theta})}{4}-\frac{F(\alpha,\bar{\theta})}{2}\right).
$$

Case 3. When $\hat{t}\in\big(t(\alpha,\underline{\theta}),t(\alpha,\bar{\theta})\big)$ and $U_C(\hat{t},war)>U_C(t(\alpha,\underline{\theta}))$, *A Sometimes Risks War:*

- *C’s beliefs on θ follows its expected distribution. C selects transgression level*

$$
t^* = \frac{1}{2\kappa_t}.
$$

- *For the selected t^*, type $\underline{\theta}$ D will go to war and type $\bar{\theta}$ D will hassle, setting $h^*=\frac{F(\alpha,\bar{\theta})}{2}$. Each type D has utility*

$$
\begin{aligned}
U_D(\sigma^*(\underline{\theta},\alpha)) =& 1-\rho-\kappa_D,\\
U_D(\sigma^*(\bar{\theta},\alpha)) =& 1-\rho-\frac{1}{2\kappa_t}+\kappa_C+\frac{F(\alpha,\bar{\theta})}{4}.
\end{aligned}
$$

  C has expected utility

$$
EU_C = Pr(\underline{\theta})*(\rho-\kappa_C)+Pr(\bar{\theta})*\left(p-\frac{F(\alpha,\bar{\theta})}{2}-\kappa_C+\frac{1}{4\kappa_t}\right).
$$

Case 4. *When* $\hat{t}\geq t(\alpha,\bar{\theta})$ and $U_C(t(\alpha,\underline{\theta}))<U_C(t(\alpha,\bar{\theta}))$, *A Sometimes Risks War:*

- *C's beliefs on θ follows its expected distribution. C selects transgression level*

$$t^* = \kappa_D + \kappa_C + \frac{F(\alpha, \bar{\theta})}{4}.$$

For the selected t^*, a type $\underline{\theta}$ D will go to war, and a type $\bar{\theta}$ D will hassle, setting $h^* = \frac{F(\alpha,\bar{\theta})}{2}$. Each type D has utility

$$U_D(\sigma^*(\underline{\theta}, \alpha)) = 1 - \rho - \kappa_D,$$

$$U_D(\sigma^*(\bar{\theta}, \alpha)) = 1 - \rho - \kappa_D.$$

C has expected utility

$$EU_C = Pr(\underline{\theta})\left(\rho - \kappa_C\right) + Pr(\bar{\theta})\left(\rho + \kappa_D - \frac{F(\alpha, \bar{\theta})}{4}\right) - \kappa_t\left(\kappa_D + \kappa_C + \frac{F(\alpha, \bar{\theta})}{4}\right)^2.$$

Case 5. When $\hat{t} \geq t(\alpha, \bar{\theta})$ and $U_C(t(\alpha, \underline{\theta})) \geq U_C(t(\alpha, \bar{\theta}))$, C avoids war:

- *C's beliefs on θ follows its expected distribution. C selects transgression level*

$$t^* = \kappa_D + \kappa_C + \frac{F(\alpha, \underline{\theta})}{4}.$$

For the selected t^*, a type $\theta \in \{\underline{\theta}, \bar{\theta}\}$ D will always hassle, setting $h^* = \frac{F(\alpha,\theta)}{2}$. Each type D has utility

$$U_D(\sigma^*(\underline{\theta}, \alpha)) = 1 - \rho - \kappa_D,$$
$$U_D(\sigma^*(\bar{\theta}, \alpha)) = 1 - \rho - \kappa_D + \frac{F(\alpha, \bar{\theta})}{4} - \frac{F(\alpha, \underline{\theta})}{4}.$$

- *C has expected utility*

$$EU_C = \rho + \kappa_D - \kappa_t \left(\kappa_D + \kappa_C + \frac{F(\alpha, \underline{\theta})}{4}\right)^2 - Pr(\underline{\theta}) \left(\frac{F(\alpha, \underline{\theta})}{4}\right) + Pr(\bar{\theta}) \left(\frac{F(\alpha, \underline{\theta})}{4} - \frac{F(\alpha, \bar{\theta})}{2}\right).$$

Proof: Follows from discussion above.

The above completely characterizes all possible equilibria.

6.3 Results

Recall from earlier the features of C becoming emboldened or D becoming predictable following changes in α. When D becomes more predictable, there is (weakly) less war, and high type D's attain less value from their private information. When C becomes emboldened, there is more war, and C is motivated to select a new, more aggressive t^* that diminishes high type D's utility. Both circumstances can still arise here, though the conditions are more expansive than in the case in the text. To keep the terminology reasonable, I use "E" and "P" to denote emboldening or predictability, and κ_t to denote that the conditions relate to the model with costs. Also, I now refer to $U_C(t, \alpha)$ (for $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$) when it is necessary to clarify what t is being selected.

Definition: *under the* $\boldsymbol{P\kappa_t 1}$ *conditions, the following holds:*

$$\hat{t} \in \left(t(\underline{\alpha}, \underline{\theta}), t(\underline{\alpha}, \bar{\theta})\right),$$

$$\hat{t} \in \left(t(\bar{\alpha}, \underline{\theta}), t(\bar{\alpha}, \bar{\theta})\right),$$

$$U_C(\hat{t}, war, \underline{\alpha}) \leq U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha}),$$

$$U_C(\hat{t}, war, \bar{\alpha}) \leq U_C(t(\bar{\alpha}, \underline{\theta})),$$

$$F(\underline{\alpha}, \bar{\theta}) - F(\underline{\alpha}, \underline{\theta}) > F(\bar{\alpha}, \bar{\theta}) - F(\bar{\alpha}, \underline{\theta}).$$

Definition: *under the* $\boldsymbol{P\kappa_t 2}$ *conditions, the following holds:*

$$\hat{t} \geq t(\underline{\alpha}, \bar{\theta}),$$

$$\hat{t} \in \left(t(\bar{\alpha}, \underline{\theta}), t(\bar{\alpha}, \bar{\theta})\right),$$

$$U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha}) \geq U_C(t(\underline{\alpha}, \bar{\theta}), \underline{\alpha}),$$

$$U_C(\hat{t}, war, \bar{\alpha}) \leq U_C(t(\bar{\alpha}, \underline{\theta}), \bar{\alpha}),$$

$$F(\underline{\alpha}, \bar{\theta}) - F(\underline{\alpha}, \underline{\theta}) > F(\bar{\alpha}, \bar{\theta}) - F(\bar{\alpha}, \underline{\theta}).$$

Definition: *under the* $\boldsymbol{P\kappa_t 3}$ *conditions, the following holds:*

$$\hat{t} \geq t(\underline{\alpha}, \bar{\theta}),$$

$$\hat{t} \geq t(\bar{\alpha}, \bar{\theta}),$$

$$U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha}) \geq U_C(t(\underline{\alpha}, \bar{\theta}), \underline{\alpha}),$$

$$U_C(t(\bar{\alpha}, \underline{\theta}), \bar{\alpha}) \geq U_C(t(\bar{\alpha}, \bar{\theta}), \bar{\alpha}),$$

$$F(\underline{\alpha}, \bar{\theta}) - F(\underline{\alpha}, \underline{\theta}) > F(\bar{\alpha}, \bar{\theta}) - F(\bar{\alpha}, \underline{\theta}).$$

Definition: *under the* $\boldsymbol{P\kappa_t 4}$ *conditions, the following holds:* [11]

$$\hat{t} \in \left(t(\underline{\alpha}, \underline{\theta}), t(\underline{\alpha}, \bar{\theta})\right),$$

$$\hat{t} \in \left(t(\bar{\alpha}, \underline{\theta}), t(\bar{\alpha}, \bar{\theta})\right),$$

$$U_C(\hat{t}, war, \underline{\alpha}) \geq U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha}),$$

$$U_C(t(\bar{\alpha}, \underline{\theta}), \bar{\alpha}) \geq U_C(\hat{t}, war, \bar{\alpha}),$$

$$-\frac{1}{2\kappa_t} + \kappa_C + \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} > \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4}.$$

And I can also define conditions for emboldening.

[11] The last condition follows from $1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\underline{\alpha}, \bar{\theta})}{4} \geq 1 - \rho - \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4}$.

Definition: *under the* $\boldsymbol{E}\kappa_t\boldsymbol{1}$ *conditions, the following holds:*[12]

$$\hat{t} \in \left(t(\underline{\alpha}, \underline{\theta}), t(\underline{\alpha}, \bar{\theta})\right),$$

$$\hat{t} \in \left(t(\bar{\alpha}, \underline{\theta}), t(\bar{\alpha}, \bar{\theta})\right),$$

$$U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha}) \geq U_C(\hat{t}, war, \underline{\alpha}),$$

$$U_C(t(\bar{\alpha}, \underline{\theta}), \bar{\alpha}) \leq U_C(\hat{t}, war, \bar{\alpha}),$$

$$\frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > -\frac{1}{2\kappa_t} + \kappa_C + \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4}.$$

Definition: *under the* $\boldsymbol{E}\kappa_t\boldsymbol{2}$ *conditions, the following holds:*

$$\hat{t} \geq t(\underline{\alpha}, \bar{\theta}),$$

$$\hat{t} \in \left(t(\bar{\alpha}, \underline{\theta}), t(\bar{\alpha}, \bar{\theta})\right),$$

$$U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha}) \geq U_C(t(\underline{\alpha}, \bar{\theta}), \underline{\alpha}),$$

$$U_C(t(\bar{\alpha}, \underline{\theta}), \bar{\alpha}) \leq U_C(\hat{t}, war, \bar{\alpha}),$$

$$\frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > -\frac{1}{2\kappa_t} + \kappa_C + \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4}.$$

[12] The last condition is equivalent to $1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > 1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\bar{\alpha}, \bar{\theta})}{4}$.

Definition: *under the* $\boldsymbol{E\kappa_t 3}$ *conditions, the following holds:*

$$\hat{t} \geq t(\underline{\alpha}, \bar{\theta}),$$

$$\hat{t} \geq t(\bar{\alpha}, \bar{\theta}),$$

$$U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha}) \geq U_C(t(\underline{\alpha}, \bar{\theta}), \underline{\alpha}),$$

$$U_C(t(\bar{\alpha}, \underline{\theta}), \bar{\alpha}) < U_C(t(\bar{\alpha}, \bar{\theta}), \bar{\alpha}).$$

Having introduced these cases, I can present Proposition 6, which is analogous to Proposition 4:

Proposition 6: *Improvements in* α *produce a capability failure if and only if* $\boldsymbol{P\kappa_t 1}$*,* $\boldsymbol{P\kappa_t 2}$*,* $\boldsymbol{P\kappa_t 3}$*,* $\boldsymbol{P\kappa_t 4}$*,* $\boldsymbol{E\kappa_t 1}$*,* $\boldsymbol{E\kappa_t 2}$*, or* $\boldsymbol{E\kappa_t 3}$ *hold.*

6.4 Proving Proposition 6

6.4.1 Proving ←

On $\mathbf{P}\kappa_t$1:

Through the top four $\mathbf{P}\kappa_t\mathbf{1}$ conditions, across $\underline{\alpha}$ and $\bar{\alpha}$, the equilibrium is discussed under Case 2 in Proposition 5. The fifth condition defines

$$F(\underline{\alpha}, \bar{\theta}) - F(\underline{\alpha}, \underline{\theta}) > F(\bar{\alpha}, \bar{\theta}) - F(\bar{\alpha}, \underline{\theta}),$$

which through algebra is equivalent to

$$1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > 1 - \rho - \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4},$$

where the left-hand side is type $\bar{\theta}$, parameter $\underline{\alpha}$ D's utility and the right hand side is type $\bar{\theta}$, parameter $\bar{\alpha}$ D's utility. Because across parameters $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$ types $\underline{\theta}$ attain their wartime utility, I am able to say that

$$U_D(\sigma^*(\underline{\theta}, \underline{\alpha})) = U_D(\sigma^*(\underline{\theta}, \bar{\alpha}))$$

and

$$U_D(\sigma^*(\bar{\theta}, \underline{\alpha})) > U_D(\sigma^*(\bar{\theta}, \bar{\alpha})),$$

which is a capability failure.

An example of parameters that produce this condition are as follows: $F(\underline{\alpha}, \underline{\theta}) = 0.1$, $F(\underline{\alpha}, \bar{\theta}) = 0.43$, $F(\bar{\alpha}, \underline{\theta}) = 0.35$, $F(\bar{\alpha}, \bar{\theta}) = 0.44$, $\kappa_D = 0.3$, $\kappa_C = 0.1$, $\rho = 0.1$, $Pr(\underline{\theta}) = 0.2$, $Pr(\bar{\theta}) = 0.8$, $\kappa_t = 1$.

On Pκ_t2:

Through the top four **P**κ_t**2** conditions, when $\alpha = \underline{\alpha}$, the equilibrium is discussed under Case 5 in Proposition 5, and when $\alpha = \bar{\alpha}$, the equilibrium is discussed under Case 2. The fifth condition defines

$$F(\underline{\alpha}, \bar{\theta}) - F(\underline{\alpha}, \underline{\theta}) > F(\bar{\alpha}, \bar{\theta}) - F(\bar{\alpha}, \underline{\theta}),$$

which through algebra is equivalent to

$$1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > 1 - \rho - \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4},$$

where the left-hand side is type $\bar{\theta}$, parameter $\underline{\alpha}$ D's utility and the right hand side is type $\bar{\theta}$, parameter $\bar{\alpha}$ D's utility. Because across parameters $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$ types $\underline{\theta}$ attain their wartime utility, I am able to say that

$$U_D(\sigma^*(\underline{\theta}, \underline{\alpha})) = U_D(\sigma^*(\underline{\theta}, \bar{\alpha}))$$

and

$$U_D(\sigma^*(\bar{\theta}, \underline{\alpha})) > U_D(\sigma^*(\bar{\theta}, \bar{\alpha})),$$

which is a capability failure.

An example of parameters that produce this condition are as follows: $F(\underline{\alpha}, \underline{\theta}) = 0.1$, $F(\underline{\alpha}, \bar{\theta}) = 0.39$, $F(\bar{\alpha}, \underline{\theta}) = 0.35$, $F(\bar{\alpha}, \bar{\theta}) = 0.44$, $\kappa_D = 0.3$, $\kappa_C = 0.1$, $\rho = 0.1$, $Pr(\underline{\theta}) = 0.2$, $Pr(\bar{\theta}) = 0.8$, $\kappa_t = 1$.

On Pκ_t3:

Through the top four **Pκ_t3** conditions, across $\underline{\alpha}$ and $\bar{\alpha}$, the equilibrium is discussed under Case 5 in Proposition 5. The fifth condition defines

$$F(\underline{\alpha}, \bar{\theta}) - F(\underline{\alpha}, \underline{\theta}) > F(\bar{\alpha}, \bar{\theta}) - F(\bar{\alpha}, \underline{\theta}),$$

which through algebra is equivalent to

$$1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > 1 - \rho - \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4},$$

where the left-hand side is type $\bar{\theta}$, parameter $\underline{\alpha}$ D's utility and the right hand side is type $\bar{\theta}$, parameter $\bar{\alpha}$ D's utility. Because across parameters $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$ types $\underline{\theta}$ attain their wartime utility, I am able to say that

$$U_D(\sigma^*(\underline{\theta},\underline{\alpha})) = U_D(\sigma^*(\underline{\theta},\bar{\alpha}))$$

and

$$U_D(\sigma^*(\bar{\theta},\underline{\alpha})) > U_D(\sigma^*(\bar{\theta},\bar{\alpha})),$$

which is a capability failure.

An example of parameters that produce this condition are as follows: $F(\underline{\alpha},\underline{\theta}) = 0.25$, $F(\underline{\alpha},\bar{\theta}) = 0.4$, $F(\bar{\alpha},\underline{\theta}) = 0.45$, $F(\bar{\alpha},\bar{\theta}) = 0.48$, $\kappa_D = 0.2$, $\kappa_C = 0.1$, $\rho = 0.2$, $Pr(\underline{\theta}) = 0.3$, $Pr(\bar{\theta}) = 0.7$, $\kappa_t = 1$.

Through the top four $\mathbf{P}\kappa_t\mathbf{4}$ conditions, when $\alpha = \underline{\alpha}$, the equilibrium is discussed under Case 3 in Proposition 5, and when $\alpha = \bar{\alpha}$, the equilibrium is discussed under Case 2. The fifth condition defines

$$-\frac{1}{2\kappa_t} + \kappa_C + \kappa_D + \frac{F(\underline{\alpha},\bar{\theta})}{4} > \frac{F(\bar{\alpha},\bar{\theta})}{4} - \frac{F(\bar{\alpha},\underline{\theta})}{4},$$

which through algebra is equivalent to

$$1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\alpha,\bar{\theta})}{4} > 1 - \rho - \kappa_D + \frac{F(\bar{\alpha},\bar{\theta})}{4} - \frac{F(\bar{\alpha},\underline{\theta})}{4},$$

where the left-hand side is type $\bar{\theta}$, parameter $\underline{\alpha}$ D's utility and the right hand side is type $\bar{\theta}$, parameter $\bar{\alpha}$ D's utility. Because across parameters $\alpha \in \{\underline{\alpha},\bar{\alpha}\}$ types $\underline{\theta}$ attain their wartime utility, I am able to say that

$$U_D(\sigma^*(\underline{\theta},\underline{\alpha})) = U_D(\sigma^*(\underline{\theta},\bar{\alpha}))$$

and

$$U_D(\sigma^*(\bar{\theta}, \underline{\alpha})) > U_D(\sigma^*(\bar{\theta}, \bar{\alpha})),$$

which is a capability failure.

An example of parameters that produce this condition are as follows: $F(\underline{\alpha}, \underline{\theta}) = 0.01$, $F(\underline{\alpha}, \bar{\theta}) = 0.26$, $F(\bar{\alpha}, \underline{\theta}) = 0.24$, $F(\bar{\alpha}, \bar{\theta}) = 0.27$, $\kappa_D = 0.25$, $\kappa_C = 0.001$, $\rho = 0.2$, $Pr(\underline{\theta}) = 0.01$, $Pr(\bar{\theta}) = 0.99$, $\kappa_t = 1.6$.]

On Eκ_t1:

Through the top four **Eκ_t1** conditions, when $\alpha = \underline{\alpha}$, the equilibrium is discussed under Case 2 in Proposition 5, and when $\alpha = \bar{\alpha}$, the equilibrium is discussed under Case 3. The fifth condition defines

$$\frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > -\frac{1}{2\kappa_t} + \kappa_C + \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4},$$

which through algebra is equivalent to

$$1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > 1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\bar{\alpha}, \bar{\theta})}{4},$$

where the left-hand side is type $\bar{\theta}$, parameter $\underline{\alpha}$ D's utility and the right hand side is type $\bar{\theta}$, parameter $\bar{\alpha}$ D's utility. Because across parameters $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$ types $\underline{\theta}$ attain their wartime utility, I am able to say that

$$U_D(\sigma^*(\underline{\theta}, \underline{\alpha})) = U_D(\sigma^*(\underline{\theta}, \bar{\alpha}))$$

and

$$U_D(\sigma^*(\bar{\theta},\underline{\alpha})) > U_D(\sigma^*(\bar{\theta},\bar{\alpha})),$$

which is a capability failure.

It is worthwhile mentioning that I have not been able to find parameters that fit the five conditions of $\mathbf{E}\kappa_t\mathbf{1}$, but I also have not been able to prove that these conditions cannot all simultaneously exist. If in fact the five conditions of $\mathbf{E}\kappa_t\mathbf{1}$ cannot all hold simultaneously, then of $\mathbf{E}\kappa_t\mathbf{1}$ cannot create a capability failure, nor should it be considered in the → part of the proof (but it would not negate the key results).

On $\mathbf{E}\kappa_t\mathbf{2}$:

Through the top four $\mathbf{E}\kappa_t\mathbf{2}$ conditions, when $\alpha = \underline{\alpha}$, the equilibrium is discussed under Case 5 in Proposition 5, and when $\alpha = \bar{\alpha}$, the equilibrium is discussed under Case 3. The fifth condition defines

$$\frac{F(\underline{\alpha},\bar{\theta})}{4} - \frac{F(\underline{\alpha},\underline{\theta})}{4} > -\frac{1}{2\kappa_t} + \kappa_C + \kappa_D + \frac{F(\bar{\alpha},\bar{\theta})}{4},$$

which through algebra is equivalent to

$$1-\rho-\kappa_D + \frac{F(\underline{\alpha},\bar{\theta})}{4} - \frac{F(\underline{\alpha},\underline{\theta})}{4} > 1-\rho-\frac{1}{2\kappa_t} + \kappa_C + \frac{F(\bar{\alpha},\bar{\theta})}{4},$$

where the left-hand side is type $\bar{\theta}$, parameter $\underline{\alpha}$ D's utility and the right hand side is type $\bar{\theta}$, parameter $\bar{\alpha}$ D's utility. Because across parameters $\alpha \in \{\underline{\alpha},\bar{\alpha}\}$ types $\underline{\theta}$ attain their wartime utility, I am able to say that

$$U_D(\sigma^*(\underline{\theta},\underline{\alpha})) = U_D(\sigma^*(\underline{\theta},\bar{\alpha}))$$

and

$$U_D(\sigma^*(\bar{\theta},\underline{\alpha})) > U_D(\sigma^*(\bar{\theta},\bar{\alpha})),$$

which is a capability failure.

An example of parameters that produce this condition are as follows: $F(\underline{\alpha},\underline{\theta}) = 0.001$, $F(\underline{\alpha},\bar{\theta}) = 0.005$, $F(\bar{\alpha},\underline{\theta}) = 0.002$, $F(\bar{\alpha},\bar{\theta}) = 0.4$, $\kappa_D = 0.4$, $\kappa_C = 0.01$, $\rho = 0.4$, $Pr(\underline{\theta}) = 0.01$, $Pr(\bar{\theta}) = 0.99$, $\kappa_t = 1$.

On Eκ_t3:

Through the top four **Eκ_t3** conditions, when $\alpha = \underline{\alpha}$, the equilibrium is discussed under Case 5 in Proposition 5, and when $\alpha = \bar{\alpha}$, the equilibrium is discussed under Case 4. Comparing the utilities that a capabilities $(\bar{\alpha},\bar{\theta})$ attains (right hand side below) to the utility that a capabilities $(\underline{\alpha},\bar{\theta})$ attain (left hand side below) is

$$1-\rho-\kappa_D+\frac{F(\underline{\alpha},\bar{\theta})}{4}-\frac{F(\underline{\alpha},\underline{\theta})}{4} > 1-\rho-\kappa_D.$$

Thus, I can say

$$U_D(\sigma^*(\underline{\theta},\underline{\alpha})) = U_D(\sigma^*(\underline{\theta},\bar{\alpha}))$$

and

$$U_D(\sigma^*(\bar{\theta},\underline{\alpha})) > U_D(\sigma^*(\bar{\theta},\bar{\alpha})),$$

which is a capability failure.

An example of parameters that produce this condition are as follows: $F(\underline{\alpha},\underline{\theta}) = 0.1$, $F(\underline{\alpha},\bar{\theta}) = 0.2$, $F(\bar{\alpha},\underline{\theta}) = 0.15$, $F(\bar{\alpha},\bar{\theta}) = 0.4$, $\kappa_D = 0.25$, $\kappa_C = 0.05$, $\rho = 0.2$, $Pr(\underline{\theta}) = 0.05$,

$Pr(\bar{\theta}) = 0.95$, $\kappa_t = 1.15$.

6.4.2 Proving $\rightarrow$

When there is a capability failure following improvements in α, it must be that

$$U_D(\sigma^*(\underline{\theta}, \underline{\alpha})) \geq U_D(\sigma^*(\underline{\theta}, \bar{\alpha})) \tag{6}$$

and

$$U_D(\sigma^*(\bar{\theta}, \underline{\alpha})) \geq U_D(\sigma^*(\bar{\theta}, \bar{\alpha})), \tag{7}$$

with at least one inequality holding strictly. Proposition 5 shows, for a fixed α, five cases are possible. Thus, to compare D's utility across $\underline{\alpha}$ and $\bar{\alpha}$, I must consider multiple combinations of cases. I list various outcomes across $\underline{\alpha}$ and $\bar{\alpha}$; I will use the notation "Case i-Case j)" to represent that under $\underline{\alpha}$ the equilibrium is described under Case i (with $i \in \{1, 2, 3, 4, 5\}$) and that under $\bar{\alpha}$ the equilibrium is described under Case j (with $j \in \{1, 2, 3, 4, 5\}$). I am able to limit the number of case combinations that I consider by observing that if $\hat{t} \leq t(\theta, \underline{\alpha})$ for $\theta \in \{\underline{\theta}, \bar{\theta}\}$, then $\hat{t} < t(\theta, \bar{\alpha})$ (because $t(\theta, \alpha)$ is increasing in α). Thus, I do not need to consider the following: Case 1-Case 2, Case 1-Case 3, Case 1-Case 4, Case 1-Case 5, Case 2-Case 4, Case 2-Case 5, Case 3-Case 4, and Case 3-Case 5. Thus, I can go through the following remaining outcomes to either (a) show that if there is a capability failure, then it implies that one of the conditions outlined earlier holds, or (b) there cannot be a capability failure.

Case 1-Case 1:

By the conditions of the cases, $\hat{t} \leq t(\underline{\theta}, \alpha)$ across $\underline{\alpha}$ and $\bar{\alpha}$. C always selects $t = \frac{1}{2\kappa_t}$, and D always hassles. By the conditions of the "if" clause (i.e. if there is a capability failure), this case is ruled out. Because F is increasing in α, for $\theta \in \{\underline{\theta}, \bar{\theta}\}$, the following statement must hold:

$$1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\underline{\alpha}, \theta)}{4} < 1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\bar{\alpha}, \theta)}{4}.$$

This suggests that increases in α always produces better outcomes for D (the left hand side (LHS) and right hand side (RHS) above are D's utility for $\underline{\alpha}$ and $\bar{\alpha}$, respectively).

Case 2-Case 1:

By the conditions of the cases, $\hat{t} \in \big(t(\underline{\alpha}, \underline{\theta}), t(\underline{\alpha}, \bar{\theta})\big)$ and $\hat{t} \leq t(\bar{\alpha}, \underline{\theta})$. C will select $t(\underline{\alpha}, \underline{\theta})$ when $\alpha = \underline{\alpha}$ and $\hat{t}$ when $\alpha = \bar{\alpha}$, and D always hassles. By the conditions of the "if" clause (i.e. if there is a capability failure), this case is ruled out. Because $\hat{t} < t(\bar{\alpha}, \underline{\theta})$ implies $\frac{1}{2\kappa_t} < \kappa_D + \kappa_C + \frac{F(\bar{\alpha}, \underline{\theta})}{4}$, the following statement must hold:

$$1 - \rho - \kappa_D < 1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\bar{\alpha}, \underline{\theta})}{4}.$$

This suggests that increases in α always produces better outcomes for type $\underline{\theta}$ D's (the LHS and RHS above are D's utility for $\underline{\alpha}$ and $\bar{\alpha}$, respectively).

Case 2-Case 2:

By the conditions of the cases, $\hat{t} \in \big(t(\underline{\alpha}, \underline{\theta}), t(\underline{\alpha}, \bar{\theta})\big)$ and $\hat{t} \in \big(t(\bar{\alpha}, \underline{\theta}), t(\bar{\alpha}, \bar{\theta})\big)$, which satisfies the first two conditions of Pκ_t1. C always selects $t^* = t(\underline{\theta}, \alpha)$, and D always hassles. The conditions of the cases imply that $U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha}) \geq U_C(\hat{t}, war, \underline{\alpha})$ and $U_C(t(\bar{\alpha}, \underline{\theta}), \bar{\alpha}) \geq U_C(\hat{t}, war, \bar{\alpha})$, which satisfies the third and forth conditions of Pκ_t1. Furthermore, because there is a capability failure, I can substitute D's utilities from these cases into equations (6) and (7) above, yielding

$$1 - \rho - \kappa_D \geq 1 - \rho - \kappa_D,$$

$$1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} \geq 1 - \rho - \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4}.$$

Because the top holds with equality, the bottom inequality must be strict. Re-writing the bottom inequality yields $F(\underline{\alpha},\bar{\theta}) - F(\underline{\alpha},\underline{\theta}) > F(\bar{\alpha},\bar{\theta}) - F(\bar{\alpha},\underline{\theta})$, which is the final condition in Pκ_t1. Thus, when there is a capability failure in Case 2-Case 2, the conditions defining $P\kappa_t 1$ must hold.

Case 2-Case 3:

By the conditions of the cases, $\hat{t} \in \big(t(\underline{\alpha},\underline{\theta}), t(\underline{\alpha},\bar{\theta})\big)$ and $\hat{t} \in \big(t(\bar{\alpha},\underline{\theta}), t(\bar{\alpha},\bar{\theta})\big)$, which satisfies the first two conditions of Eκ_t1. C will select $t(\underline{\alpha},\underline{\theta})$ when $\alpha = \underline{\alpha}$ and $\hat{t}$ when $\alpha = \bar{\alpha}$, and D always hassles when $\alpha = \underline{\alpha}$ and sometimes goes to war when $\alpha = \bar{\alpha}$. The conditions of the cases imply that $U_C(t(\underline{\alpha},\underline{\theta}),\underline{\alpha}) \geq U_C(\hat{t}, war, \underline{\alpha})$ and $U_C(t(\bar{\alpha},\underline{\theta}),\bar{\alpha}) < U_C(\hat{t}, war, \bar{\alpha})$, which satisfies the third and forth conditions of Eκ_t1. Furthermore, because there is a capability failure, I can substitute D's utilities from these cases into equations (6) and (7) above, yielding

$$1-\rho-\kappa_D \geq 1-\rho-\kappa_D$$

and

$$1-\rho-\kappa_D+\frac{F(\underline{\alpha},\bar{\theta})}{4}-\frac{F(\underline{\alpha},\underline{\theta})}{4} \geq 1-\rho-\frac{1}{2\kappa_t}+\kappa_C+\frac{F(\bar{\alpha},\bar{\theta})}{4}.$$

Because the top holds with equality, the bottom inequality must be strict. Re-writing the bottom inequality yields $\frac{F(\underline{\alpha},\bar{\theta})}{4} - \frac{F(\underline{\alpha},\underline{\theta})}{4} > -\frac{1}{2\kappa_t} + \kappa_C + \kappa_D + \frac{F(\bar{\alpha},\bar{\theta})}{4}$, which is the final condition in Eκ_t1. Thus, when there is a capability failure in Case 2-Case 3, the conditions defining Eκ_t1 must hold.

Case 3-Case 1:

By the conditions of the cases, $\hat{t} \in \big(t(\underline{\alpha},\underline{\theta}), t(\underline{\alpha},\bar{\theta})\big)$ and $\hat{t} \leq t(\bar{\alpha},\underline{\theta})$. C will always select $\hat{t}$, and D sometimes goes to war when $\alpha = \underline{\alpha}$ and always hassles when $\alpha = \bar{\alpha}$. By the conditions of the “if” clause (i.e. if there is a capability failure), this case is ruled out. Because $\hat{t} < t(\bar{\alpha}, \underline{\theta})$ implies $\frac{1}{2\kappa_t} < \kappa_D + \kappa_C + \frac{F(\bar{\alpha},\underline{\theta})}{4}$, the following statement must hold:

$$1 - \rho - \kappa_D < 1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\bar{\alpha}, \underline{\theta})}{4}.$$

This suggests that increases in α always produces better outcomes for type $\underline{\theta}$ D’s (the LHS and RHS above are D’s utility for $\underline{\alpha}$ and $\bar{\alpha}$, respectively).

Case 3-Case 2:

By the conditions of the cases, $\hat{t} \in \big(t(\underline{\alpha}, \underline{\theta}), t(\underline{\alpha}, \bar{\theta})\big)$ and $\hat{t} \in \big(t(\bar{\alpha}, \underline{\theta}), t(\bar{\alpha}, \bar{\theta})\big)$, which satisfies the first two conditions of Pκ_t4. C will select $\hat{t}$ when $\alpha = \underline{\alpha}$ and $t(\bar{\alpha}, \underline{\theta})$ when $\alpha = \bar{\alpha}$, and D sometimes goes to war when $\alpha = \underline{\alpha}$ and always hassles when $\alpha = \bar{\alpha}$. The conditions of the cases imply that $U_C(\hat{t}, war, \underline{\alpha}) \geq U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha})$ and $U_C(t(\bar{\alpha}, \underline{\theta}), \bar{\alpha}) \geq U_C(\hat{t}, war, \bar{\alpha})$, which satisfies the third and forth conditions of Pκ_t4. Furthermore, because there is a capability failure, I can substitute D’s utilities from these cases into equations (6) and (7) above, yielding

$$1 - \rho - \kappa_D \geq 1 - \rho - \kappa_D$$

and

$$1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\underline{\alpha}, \bar{\theta})}{4} \geq 1 - \rho - \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4}$$

Because the top holds with equality, the bottom inequality must be strict. Re-writing the bottom inequality yields $-\frac{1}{2\kappa_t} + \kappa_C + \kappa_D + \frac{F(\underline{\alpha},\bar{\theta})}{4} > \frac{F(\bar{\alpha},\bar{\theta})}{4} - \frac{F(\bar{\alpha},\underline{\theta})}{4}$, which is the final condition in Pκ_t4. Thus, when there is a capability failure in Case 3-Case 2, the conditions defining Pκ_t4 must hold.

Case 3-Case 3:

By the conditions of the cases, $\hat{t} \in \big(t(\underline{\alpha}, \underline{\theta}), t(\underline{\alpha}, \bar{\theta})\big)$ and $\hat{t} \in \big(t(\bar{\alpha}, \underline{\theta}), t(\bar{\alpha}, \bar{\theta})\big)$. C always selects $t = \frac{1}{2\kappa_t}$, and D sometimes goes to war when $\alpha = \underline{\alpha}$ and $\alpha = \bar{\alpha}$. By the conditions of the "if" clause (i.e. if there is a capability failure), this case is ruled out. Because F is increasing in α, the following statement must hold:

$$1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\underline{\alpha}, \bar{\theta})}{4} < 1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\bar{\alpha}, \bar{\theta})}{4}.$$

This suggests that increases in α produces better outcomes for type $\bar{\theta}$ D's (the LHS and RHS above are D's utility for $\underline{\alpha}$ and $\bar{\alpha}$, respectively).

Case 4-Case 1:

By the conditions of the cases, $\hat{t} \geq t(\underline{\alpha}, \bar{\theta})$ and $\hat{t} \leq t(\bar{\alpha}, \underline{\theta})$. C will select $t(\underline{\alpha}, \bar{\theta})$ when $\alpha = \underline{\alpha}$ and $\hat{t}$ when $\alpha = \bar{\alpha}$, and D sometimes goes to war when $\alpha = \underline{\alpha}$ and always hassles when $\alpha = \bar{\alpha}$. By the conditions of the "if" clause (i.e. if there is a capability failure), this case is ruled out. Because $\hat{t} < t(\bar{\alpha}, \underline{\theta})$ implies $\frac{1}{2\kappa_t} < \kappa_D + \kappa_C + \frac{F(\bar{\alpha}, \underline{\theta})}{4}$, the following statement must hold:

$$1 - \rho - \kappa_D < 1 - \rho - \frac{1}{2\kappa_t} + \kappa_C + \frac{F(\bar{\alpha}, \underline{\theta})}{4}.$$

This suggests that increases in α always produces better outcomes for type $\underline{\theta}$ D's (the LHS and RHS above are D's utility for $\underline{\alpha}$ and $\bar{\alpha}$, respectively).

Case 4-Case 2:

By the conditions of the cases, $\hat{t} \geq t(\underline{\alpha}, \bar{\theta})$ and $\hat{t} \in \big(t(\bar{\alpha}, \underline{\theta}), t(\bar{\alpha}, \bar{\theta})\big)$. C will select $t(\underline{\alpha}, \bar{\theta})$ when $\alpha = \underline{\alpha}$ and $t(\bar{\alpha}, \underline{\theta})$ when $\alpha = \bar{\alpha}$, and D sometimes goes to war when $\alpha = \underline{\alpha}$ and always hassles when $\alpha = \bar{\alpha}$. By the conditions of the "if" clause (i.e. if there is a capability failure), this case is ruled out. Because F is increasing in θ, the following statement must hold:

$$1-\rho-\kappa_D < 1-\rho-\kappa_D+\frac{F(\bar{\alpha},\bar{\theta})}{4}-\frac{F(\bar{\alpha},\underline{\theta})}{4}$$

This suggests that increasing in α always produces better outcomes for types $\bar{\theta}$ D (the LHS and RHS above are D's utility for $\underline{\alpha}$ and $\bar{\alpha}$, respectively).

Case 4-Case 3:

By the conditions of the cases, $\hat{t} \geq t(\underline{\alpha},\bar{\theta})$ and $\hat{t} \in \big(t(\bar{\alpha},\underline{\theta}), t(\bar{\alpha},\bar{\theta})\big)$. C will select $t(\underline{\alpha},\bar{\theta})$ when $\alpha=\underline{\alpha}$ and $\hat{t}$ when $\alpha=\bar{\alpha}$, and D sometimes goes to war when $\alpha=\underline{\alpha}$ and when $\alpha=\bar{\alpha}$. By the conditions of the "if" clause (i.e. if there is a capability failure), this case is ruled out. Because $\hat{t} < t(\bar{\alpha},\bar{\theta})$ implies $\frac{1}{2\kappa_t} < \kappa_D+\kappa_C+\frac{F(\bar{\alpha},\bar{\theta})}{4}$, the following statement must hold:

$$1-\rho-\kappa_D < 1-\rho-\frac{1}{2\kappa_t}+\kappa_C+\frac{F(\bar{\alpha},\bar{\theta})}{4}.$$

This suggests that increases in α always produces better outcomes for type $\bar{\theta}$ D's (the LHS and RHS above are D's utility for $\underline{\alpha}$ and $\bar{\alpha}$, respectively).

Case 4-Case 4: All types do the same.

By the conditions of the cases, $\hat{t} \geq t(\underline{\alpha},\bar{\theta})$ and $\hat{t} \geq t(\bar{\alpha},\bar{\theta})$. C will select $t(\underline{\alpha},\bar{\theta})$ when $\alpha=\underline{\alpha}$ and $t(\bar{\alpha},\bar{\theta})$ when $\alpha=\bar{\alpha}$, and D sometimes goes to war when $\alpha=\underline{\alpha}$ and when $\alpha=\bar{\alpha}$. By the conditions of the "if" clause (i.e. if there is a capability failure), this case is ruled out. Because all types $\underline{\theta}$ and $\bar{\theta}$ D attain their wartime utility across $\alpha=\underline{\alpha}$ and $\alpha=\bar{\alpha}$, improvements in α produce no change in D's utility.

Case 4-Case 5:

By the conditions of the cases, $\hat{t} \geq t(\underline{\alpha},\bar{\theta})$ and $\hat{t} \geq t(\bar{\alpha},\bar{\theta})$. C will select $t(\underline{\alpha},\bar{\theta})$ when $\alpha=\underline{\alpha}$ and $t(\bar{\alpha},\underline{\theta})$ when $\alpha=\bar{\alpha}$, and D sometimes goes to war when $\alpha=\underline{\alpha}$ and always hassles when $\alpha=\bar{\alpha}$. By the conditions of the "if" clause (i.e. if there is a capability failure), this case is ruled out. Because F is increasing in θ, the following statement must hold:

$$1-\rho-\kappa_D < 1-\rho-\kappa_D + \frac{F(\bar{\alpha},\bar{\theta})}{4} - \frac{F(\bar{\alpha},\underline{\theta})}{4}$$

This suggests that increasing in α always produces better outcomes for types $\bar{\theta}$ D (the LHS and RHS above are D's utility for $\underline{\alpha}$ and $\bar{\alpha}$, respectively).

Case 5-Case 1:

By the conditions of the cases, $\hat{t} \geq t(\underline{\alpha},\bar{\theta})$ and $\hat{t} \leq t(\bar{\alpha},\underline{\theta})$. C will select $t(\underline{\alpha},\underline{\theta})$ when $\alpha = \underline{\alpha}$ and $\hat{t}$ when $\alpha = \bar{\alpha}$, and D sometimes goes to war when $\alpha = \underline{\alpha}$ and always hassles when $\alpha = \bar{\alpha}$. By the conditions of the "if" clause (i.e. if there is a capability failure), this case is ruled out. Because $\hat{t} < t(\bar{\alpha},\underline{\theta})$ implies $\frac{1}{2\kappa_t} < \kappa_D + \kappa_C + \frac{F(\bar{\alpha},\underline{\theta})}{4}$, the following statement must hold:

$$1-\rho-\kappa_D < 1-\rho-\frac{1}{2\kappa_t}+\kappa_C+\frac{F(\bar{\alpha},\underline{\theta})}{4}.$$

This suggests that increases in α always produces better outcomes for type $\underline{\theta}$ D's (the LHS and RHS above are D's utility for $\underline{\alpha}$ and $\bar{\alpha}$, respectively).

Case 5-Case 2:

By the conditions of the cases, $\hat{t} \geq t(\underline{\alpha},\bar{\theta})$ and $\hat{t} \in \big(t(\bar{\alpha},\underline{\theta}), t(\bar{\alpha},\bar{\theta})\big)$ which satisfies the first two conditions of Pκ_t2. C will select $t(\underline{\alpha},\underline{\theta})$ when $\alpha = \underline{\alpha}$ and $t(\bar{\alpha},\underline{\theta})$ when $\alpha = \bar{\alpha}$, and D sometimes goes to war when $\alpha = \underline{\alpha}$ and always hassles when $\alpha = \bar{\alpha}$. The conditions of the cases imply that $U_C(t(\underline{\alpha},\underline{\theta}),\underline{\alpha}) \geq U_C(t(\underline{\alpha},\bar{\theta}),\underline{\alpha})$ and $U_C(t(\bar{\alpha},\underline{\theta}),\bar{\alpha}) \geq U_C(\hat{t},war,\bar{\alpha})$, which satisfies the third and forth conditions of Pκ_t2. Furthermore, because there is a capability failure, I can substitute D's utilities from these cases into equations (6) and (7) above, yielding

$$
\begin{aligned}
1-\rho-\kappa_D \geq& 1-\rho-\kappa_D, \\
1-\rho-\kappa_D+\frac{F(\underline{\alpha},\bar{\theta})}{4}-\frac{F(\underline{\alpha},\underline{\theta})}{4} \geq& 1-\rho-\kappa_D+\frac{F(\bar{\alpha},\bar{\theta})}{4}-\frac{F(\bar{\alpha},\underline{\theta})}{4}.
\end{aligned}
$$

Because the top holds with equality, the bottom inequality must be strict. Re-writing the bottom inequality yields $F(\underline{\alpha},\bar{\theta})-F(\underline{\alpha},\underline{\theta}) > F(\bar{\alpha},\bar{\theta})-F(\bar{\alpha},\underline{\theta})$, which is the final condition in Pκ_t2. Thus, when there is a capability failure in Case 5-Case 2, the conditions defining $P\kappa_t2$ must hold.

Case 5-Case 3:[13]

By the conditions of the cases, $\hat{t} \geq t(\underline{\alpha},\bar{\theta})$ and $\hat{t} \in \big(t(\bar{\alpha},\underline{\theta}), t(\bar{\alpha},\bar{\theta})\big)$ which satisfies the first two conditions of Eκ_t2. C will select $t(\underline{\alpha},\underline{\theta})$ when $\alpha=\underline{\alpha}$ and $\hat{t}$ when $\alpha=\bar{\alpha}$, and D always hassles when $\alpha=\underline{\alpha}$ and sometimes goes to war when $\alpha=\bar{\alpha}$. The conditions of the cases imply that $U_C(t(\underline{\alpha},\underline{\theta}),\underline{\alpha}) \geq U_C(t(\underline{\alpha},\bar{\theta}),\underline{\alpha})$ and $U_C(t(\bar{\alpha},\underline{\theta}),\bar{\alpha}) < U_C(\hat{t}, war, \bar{\alpha})$, which satisfies the third and forth conditions of Eκ_t2. Furthermore, because there is a capability failure, I can substitute D's utilities from these cases into equations (6) and (7) above, yielding

$$
\begin{aligned}
1-\rho-\kappa_D \geq& 1-\rho-\kappa_D, \\
1-\rho-\kappa_D+\frac{F(\underline{\alpha},\bar{\theta})}{4}-\frac{F(\underline{\alpha},\underline{\theta})}{4} \geq& 1-\rho-\frac{1}{2\kappa_t}+\kappa_C+\frac{F(\bar{\alpha},\bar{\theta})}{4}.
\end{aligned}
$$

Because the top holds with equality, the bottom inequality must be strict. Re-writing the bottom inequality yields $\frac{F(\underline{\alpha},\bar{\theta})}{4}-\frac{F(\underline{\alpha},\underline{\theta})}{4} > -\frac{1}{2\kappa_t}+\kappa_C+\kappa_D+\frac{F(\bar{\alpha},\bar{\theta})}{4}$, which is the final condition in Eκ_t2. Thus, when there is a capability failure in Case 5-Case 3, the conditions defining Eκ_t2 must hold.

Case 5-Case 4:

[13] Note: a capability failure may not be possible here – see above.

By the conditions of the cases, $\hat{t} \geq t(\underline{\alpha}, \bar{\theta})$ and $\hat{t} \geq t(\bar{\alpha}, \bar{\theta})$ which satisfies the first two conditions of Eκ_t3. C will select $t(\underline{\alpha}, \underline{\theta})$ when $\alpha = \underline{\alpha}$ and $t(\bar{\alpha}, \bar{\theta})$ when $\alpha = \bar{\alpha}$, and D always hassles when $\alpha = \underline{\alpha}$ and sometimes goes to war when $\alpha = \bar{\alpha}$. The conditions of the cases imply that $U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha}) \geq U_C(t(\underline{\alpha}, \bar{\theta}), \underline{\alpha})$ and $U_C(t(\bar{\alpha}, \underline{\theta}), \bar{\alpha}) < U_C(t(\bar{\alpha}, \bar{\theta}), \bar{\alpha})$, which satisfies the third and forth conditions of Eκ_t3. Note that here a capability failure occurs within this case without any further conditions needed; I substitute D's utilities into expressions (6) and (7), yielding

$$1 - \rho - \kappa_D \geq 1 - \rho - \kappa_D$$

and

$$1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} \geq 1 - \rho - \kappa_D,$$

where the bottom inequality always holds strictly.

Case 5-Case 5:

By the conditions of the cases, $\hat{t} \geq t(\underline{\alpha}, \bar{\theta})$ and $\hat{t} \geq t(\bar{\alpha}, \bar{\theta})$ which satisfies the first two conditions of Pκ_t3. C will select $t(\underline{\alpha}, \underline{\theta})$ when $\alpha = \underline{\alpha}$ and $t(\bar{\alpha}, \underline{\theta})$ when $\alpha = \bar{\alpha}$, and D always hassles when $\alpha = \underline{\alpha}$ and sometimes goes to war when $\alpha = \bar{\alpha}$. The conditions of the cases imply that $U_C(t(\underline{\alpha}, \underline{\theta}), \underline{\alpha}) \geq U_C(t(\underline{\alpha}, \bar{\theta}), \underline{\alpha})$ and $U_C(t(\bar{\alpha}, \underline{\theta}), \bar{\alpha}) \geq U_C(t(\bar{\alpha}, \bar{\theta}), \bar{\alpha})$, which satisfies the third and forth conditions of Pκ_t3. Furthermore, because there is a capability failure, I can substitute D's utilities from these cases into equations (6) and (7) above, yielding

$$\begin{aligned} 1 - \rho - \kappa_D &\geq 1 - \rho - \kappa_D, \\ 1 - \rho - \kappa_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} &\geq 1 - \rho - \kappa_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4}. \end{aligned}$$

Because the top holds with equality, the bottom inequality must be strict. Re-writing the bottom inequality yields $F(\underline{\alpha}, \bar{\theta}) - F(\underline{\alpha}, \underline{\theta}) > F(\bar{\alpha}, \bar{\theta}) - F(\bar{\alpha}, \underline{\theta})$, which is the final condition in Pκ_t3. Thus, when there is a capability failure in Case 5-Case 5, the conditions defining $P\kappa_t3$ must hold.

6.5 Other Results

In the text, I found that increases in θ always produce better results for D. Here, examining all five cases in Proposition 5, a similar result holds. Thus, once again, a capability failure can arise from improved public hassling capabilities, but cannot arise from improved private hassling capabilities.

7 Addressing Concerns Regarding the Specific Bargaining Protocol

One critique of the model in the text is that the bargaining protocol is specific. While this concern is natural—after all, the structure of international bargaining could take many possible forms—this section demonstrates that if we treat bargaining as a "black-box" in the spirit of Gurantz and Hirsch (2017), we derive similar results. Additionally, for a further and more detailed consideration of this topic, please refer to Section 9, which conducts a game-form free analysis, demonstrating that the predictability and emboldening mechanisms are pervasive within a general class of games with any number of possible bargaining protocols.

7.1 The Model and Assumptions

Two states, a challenger (C) and a defender (D) are in a crisis. The following model is one example of a flexible-response crisis bargaining game. The game form is as follows.

1. Define $\{\underline{\theta}, \bar{\theta}\} = \Theta$ with $\underline{\theta} < \bar{\theta}$. Nature sets $\theta = \underline{\theta}$ with probability $Pr(\underline{\theta})$ and sets $\theta = \bar{\theta}$ with probability $1 - Pr(\underline{\theta})$. D knows nature's selection θ, but A does not.

2. State C selects transgression $t \in \mathcal{T} = \mathbb{R}_{\geq 0}$.

3. State D can either go to war by setting $w_D = 1$ or not go to war by setting $w_D = 0$ and countering the transgression by selecting $h \in \mathcal{H} = \mathbb{R}_{\geq 0}$ (with $h = 0$ implying that D "accepts" the transgression). When D goes to war, States C and D recieve wartime payoffs $U_C = W_C$ and $U_D = W_D$ (respectively). When D does not go to war, C and D recieve $U_C = X_C + t - h$ and $U_D = X_D - t + h - \frac{h^2}{F(\alpha,\theta)}$.

We assume that $W_C < X_C$ and $W_D < X_D$, thus implying that an entirely peaceful outcome ($t = 0$ and $h = 0$) is preferred to war, but it is possible for C to select so great a t that D prefers going to war. For simplicity, we assume that t and h shift the political outcome linearly, that C faces no costs to transgressing, and that D incurs a simple strictly increasing cost term; the game-form free analysis below makes none of these simplifying assumptions. Finally, we assume $F(\alpha, \theta)$ is increasing in both α and θ.

7.2 Equilibrium

The logic here follows identically from the model in the text. I add "BB" to indicate this is for the "black box" model.

Proposition 1BB: *The following is the Perfect Pure-Strategy Bayesian Nash Equilibrium. Letting* $S(\alpha) = Pr(\underline{\theta}) \left(X_C + X_D - W_C - \frac{F(\alpha,\underline{\theta})}{4} \right) - W_D - (1 - Pr(\underline{\theta})) \left(\frac{F(\alpha,\bar{\theta})}{4} - \frac{F(\alpha,\underline{\theta})}{4} \right)$*, there are two cases:*

- ***Case 1.*** *$S(\alpha) \geq 0$ holds, C selects transgression level $t^* = t(\alpha, \underline{\theta})$, which results both types of D choosing to not go to war, setting $h^* = \frac{F(\alpha,\theta)}{2}$ for all $\theta \in \{\underline{\theta}, \bar{\theta}\}$. $D(\alpha, \underline{\theta})$ attains utility $U_D(\sigma^*(\alpha, \underline{\theta})) = W_D$, and $D(\alpha, \bar{\theta})$ attains utility $U_D(\sigma^*(\alpha, \bar{\theta})) = W_D - \frac{F(\alpha,\underline{\theta})}{4} + \frac{F(\alpha,\bar{\theta})}{4}$.*

- ***Case 2.*** *When $S(\alpha) < 0$ holds, C selects transgression level $t^* = t(\alpha, \bar{\theta})$, which results* *in $D(\alpha, \underline{\theta})$ declaring war and $D(\alpha, \bar{\theta})$ not going to war and setting $h^* = \frac{F(\alpha,\bar{\theta})}{2}$. Both $D(\alpha, \underline{\theta})$ and $D(\alpha, \bar{\theta})$ attain their wartime utility, or $U_D(\sigma^*(\alpha, \theta)) = W_D$.*

Across all cases, $D(\theta)$'s best response function is to hassle if $t^* \leq t(\theta)$ and to go to war otherwise. C's beliefs follow from the initial probability distribution of types.

7.3 Propositions

Having defined $S(\alpha)$ above, I define Propositions 2BB-4BB here. Note that the proofs for these propositions and their proofs are identical to what is shown for the non-black box model, only when $S(\alpha)$ (and how it is defined) is substituted for $Q(\alpha)$ and W_D is substituted for $1 - \rho - \kappa_D$. For that reason, I exclude these proofs.

Proposition 2BB (Predictability): *Under the BB-Predictability-Conditions, C avoids war across parameters $\underline{\alpha}$ and $\bar{\alpha}$; formally, $S(\underline{\alpha}) \geq 0$ (Condition 1) and $S(\bar{\alpha}) \geq 0$ (Condition 2). Additionally, D's private information plays a diminished role under parameter $\bar{\alpha}$ relative to parameter $\underline{\alpha}$; formally, $(F(\underline{\alpha}, \bar{\theta}) - F(\underline{\alpha}, \underline{\theta}) > F(\bar{\alpha}, \bar{\theta}) - F(\bar{\alpha}, \underline{\theta})$ (Condition 3). When these BB-Predictability-Conditions hold, then improvements from $\underline{\alpha}$ to $\bar{\alpha}$ produce a capability failure.*

Proposition 3BB (Emboldening): *Under the BB-Emboldening-Conditions, C avoids war under parameter $\underline{\alpha}$ and goes to war under parameter $\bar{\alpha}$; formally, $S(\underline{\alpha}) \geq 0$ (Condition 1) and $S(\bar{\alpha}) < 0$ (Condition 2). When these BB-Emboldening-Conditions hold, then improvements from $\underline{\alpha}$ to $\bar{\alpha}$ produce a capability failure.*

Proposition 4BB: *Improvements in α produce a capability failure if and only if the BB-Emboldening- or BB-Predictability-Conditions hold.*

7.4 Deriving the Equilibrium

Assume that for a selected t, a type θ D optimally does not go to war. This type θ D selects an optimal h^* following

$$h^*(t) \in argmax_{h\in\mathbb{R}_+} \left\{ X_D - t + h - \frac{(h)^2}{F(\alpha,\theta)} \right\}.$$

I take first-order conditions with respect to h of the expression above to identify the optimal level of hassling h^*. This unique value is $h^* = \frac{F(\theta)}{2}$. Using h^*, we re-write D's utility in terms of the selected t and parameters α and θ. This is

$$U_D = \begin{cases} X_D - t + \frac{F(\alpha,\theta)}{4} & if \; X_D - t + \frac{F(\alpha,\theta)}{4} \geq W_D, \\ W_D & otherwise. \end{cases}$$

Using D's utility above, the transgression level that makes $D(\alpha, \theta)$ indifferent between war and hassling is

$$t(\alpha,\theta) = X_D - W_D + \frac{F(\alpha,\theta)}{4}.$$

I now express C's utility. C's utility from selecting $t(\alpha, \underline{\theta})$ (which never provokes D to escalate to war) is

$$EU_C(t(\alpha,\underline{\theta})) = X_C + t(\alpha,\underline{\theta}) - Pr(\underline{\theta})\frac{F(\alpha,\underline{\theta})}{2} - (1 - Pr(\underline{\theta}))\frac{F(\alpha,\bar{\theta})}{2},$$

or

$$EU_C(t(\alpha,\underline{\theta})) = X_C + X_D - W_D - Pr(\underline{\theta})\left(\frac{F(\alpha,\underline{\theta})}{4}\right) + (1 - Pr\underline{\theta}))\left(\frac{F(\alpha,\underline{\theta})}{4} - \frac{F(\alpha,\bar{\theta})}{2}\right).$$

Similarly, we can say

$$EU_C(t(\alpha, \bar{\theta})) = Pr(\underline{\theta})W_C + (1 - Pr(\underline{\theta})) \left(X_C + X_D - W_D - \frac{F(\alpha, \bar{\theta})}{4} \right).$$

Using these expressions, C will select $t(\alpha, \underline{\theta})$ when

$$Pr(\underline{\theta}) \left(X_C + X_D - W_C - W_D - \frac{F(\alpha, \underline{\theta})}{4} \right) - (1 - Pr(\underline{\theta})) \left(\frac{F(\alpha, \bar{\theta})}{4} - \frac{F(\alpha, \underline{\theta})}{4} \right) \geq 0,$$

and $t(\alpha, \bar{\theta})$ otherwise. I define $S(\alpha) = Pr(\underline{\theta}) \left(X_C + X_D - W_C - W_D - \frac{F(\alpha, \underline{\theta})}{4} \right) - (1 - Pr(\underline{\theta})) \left(\frac{F(\alpha, \bar{\theta})}{4} - \frac{F(\alpha, \underline{\theta})}{4} \right)$. I can now define D's utility. D will attain their wartime utility in all cases other than when $S(\alpha) \geq 0$ and D is type $\bar{\theta}$. This D attains

$$U_D(\alpha, \bar{\theta}) = X_D - X_D + W_D - \frac{F(\underline{\theta})}{4} + \frac{F(\bar{\theta})}{4},$$

or

$$U_D(\alpha, \bar{\theta}) = W_D - \frac{F(\underline{\theta})}{4} + \frac{F(\bar{\theta})}{4}.$$

Part II

Analysis for Models with a Convex Type Space

8 Example Model with θ as a Convex Set and Black-Box Bargaining

This model is slightly different from the models in the text as here D also has a convex set of types $\theta \in [\underline{\theta}, \bar{\theta}] \subset \mathbb{R}_+$. In order to accommodate this assumption, I need to use a more defined cost function for D than what I did in the text. And, in better defining D's costs, the model below only lends itself to the predictability mechanism; a different set of functional form assumptions can produce the emboldening mechanism.

8.1 Game Form

As it was earlier, States C and D are in a modified, crisis bargaining game without explicit bargaining. Specifically, I treat any bargaining at the back end of the model as a black-box. This can accommodate a wide range of possible bargained outcomes.

1. Nature selects $\theta \in [\underline{\theta}, \bar{\theta}] \subset \mathbb{R}_+$, which designates D's type. Realizations of θ are uniformly distributed.

2. State C selects transgression $t \geq \underline{t} > 0$.[14]

3. State D can either go to "war" by setting $w = 1$ or not go to war (setting $w = 0$) and selecting some level of hassling $h \in \mathcal{H}$, where $\mathcal{H}$ is a closed set with lower bound $\underline{h} > 0$,

[14] It is possible to let $t = 0$, though this adds technical considerations that do not add value to the model.

and where D selecting $h = \underline{h}$ can be thought of as "accepting." Going to war produces payoffs W_A and W_D for states C and D (respectively). Choosing some level of hassling $h \in \mathcal{H}$ (i.e. accepting or hassling) yields payoffs $X_A + \frac{t}{h}$ and $X_D - \frac{t}{h} - \frac{h}{\alpha+\theta}$ for states C and D (respectively).

C's and D's utility functions can be written as $U_C(t; w, h, \alpha)$ and $U_D(w, h; t, \alpha, \theta)$, where, when war does not occur, take values

$$\begin{aligned} U_C(t; 0, h, \alpha) =& X_A + \frac{t}{h} \\ U_D(0, h; t, \alpha, \theta) =& X_D - \frac{t}{h} - \frac{h}{\alpha + \theta}, \end{aligned}$$

and when war does occur take values

$$\begin{aligned} U_C(t; 1, h, \alpha) =& W_A \\ U_D(1, h; t, \alpha, \theta) =& W_D. \end{aligned}$$

To summarize the expected utilities, the expressions $X_A + \frac{t}{h}$ and $X_D - \frac{t}{h} - \frac{h}{\alpha+\theta}$ are the values that C and D receive by not going to war and letting the transgressions and hassling come to fruition. $\frac{h}{\alpha+\theta}$ are the costs that D faces from hassling.

Putting aside the costs from hassling and transgressing, the $X_A + \frac{t}{h}$ and $X_D - \frac{t}{h}$ terms represent bargained outcomes to the game. Essentially, I am treating the non-war outcomes as a black-box, where, without transgressing ($t = 0$), states would reach some expected political settlement X_A and X_D, and where the selected levels of transgression and hassling shift this settlement. I assume that if C does not transgress *at all* (setting $t = 0$), both states prefer not going to war, or $X_A > W_A$ and $X_D - \frac{h}{\alpha+\theta}$ for all θ and α; this implies that war is not inevitable, but rather, if C selects too great a transgression t, D may find it optimal to go to war with C before the transgression is realized.

Unlike the model in the text, here I treat any bargaining after the third-stage as a black box. The benefit of doing so is that this framework can accommodate a wide range of possible bargaining protocols or possible outcomes to some game form.[15] Ultimately, adopting this framework allows me to focus on the most relevant parts of the game: if the transgression is allowed to come to fruition, it will strengthen A, D can degrade this transgression through hassling, and D can prevent the power shift by going to war today.

The actions taken in the game depend the observed and unobserved components of D's capabilities. A strategy for State D in the game is a mapping from its capabilities to its action space, or $\sigma_D : (\mathcal{A}, \Theta) \rightarrow \{war,\ not\, war\} \times \mathcal{H}$. Because State C does not know the value of θ, C's strategy is a mapping from the known parameter α, or $\sigma_A : \mathcal{A} \rightarrow \mathcal{T}$. I let σ denote a pair of strategies or $\sigma = (\sigma_A, \sigma_D)$, and I employ the equilibrium concept Bayesian Nash Equilibrium (Meyerson 1985), where a strategy profile $\sigma^* = (\sigma_A, \sigma_D)$ constitutes a Bayesian equilibrium if $\sigma_D(\alpha, \theta)$ is a best response to σ_A, and $\sigma_A(\alpha)$ is a best response to σ_D based on the known type and the common probability prior $f(\cdot)$. For ease, I limit myself to pure strategy equilibria.

8.2 Equilibrium Type, Assumptions, and Capability Failure

The actions taken in the game depend the observed and unobserved components of D's capabilities. A strategy for State D in the game is a mapping from its capabilities to its action space, or $\sigma_D : (\mathcal{A}, \Theta) \rightarrow \{war,\ not\, war\} \times \mathcal{H}$. Because State C does not know the value of θ, C's strategy is a mapping from the known parameter α, or $\sigma_A : \mathcal{A} \rightarrow \mathcal{T}$. I let σ denote a pair of strategies or $\sigma = (\sigma_A, \sigma_D)$, and I employ the equilibrium concept Bayesian Nash Equilibrium (Meyerson 1985), where a strategy profile $\sigma^* = (\sigma_A, \sigma_D)$ constitutes a Bayesian equilibrium if $\sigma_D(\alpha, \theta)$ is a best response to σ_A, and $\sigma_A(\alpha)$ is a best response to σ_D based on the known type and the common probability prior $f(\cdot)$. I let h^*, w^*, and t^* denote

[15] For example, this framework easily accommodates the ultimatum bargaining framework adopted in the text with D making the final offer, but also would accommodate C making the final offer or a randomizing procedure that gives C or D the final offer (so long that the final realized $P \in (\rho, 1)$ and $x \in (0, 1)$.

equilibrium actions. For ease, I limit myself to pure strategy equilibria.

At this point, it is possible to formalize a "capability failure" for D. I claim that improvements in public hassling capabilities result in a capability failure when a change from α to α' with $\alpha < \alpha'$ results in a lower utility for D for all possible types θ.

I make two assumptions to simplify the analysis. I assume $\underline{t} > \underline{h}(\alpha+\underline{\theta})^{-1/2}$, and $\frac{(X_D-W_D)^2(\alpha+\underline{\theta})}{4} \geq \underline{t}$. This simplifies the analysis below by ruling out cases where C can select a range of t's that induce D to accept. While including this could benefit the model by offering a more complete analysis of parameter spaces, doing so offers no value to the key results (that the predictability mechanism is the only way a capability failure can arise here).

Definition: Improvements in hassling capabilities result in a **capability failure** when, for $\alpha < \alpha'$, $U_D\left(w^*, h^*; t^*, \theta, \alpha'\right) \leq U_D\left(w^*, h^*; t^*, \theta, \alpha\right)$ for all $\theta \in \Theta$.

8.3 Equilibrium and Results

There are two possible equilibria outcomes: one where C never risks war, and one where C sometimes risks war. Note that from the limitations on $\underline{t}$, C will never try to get C to "accept." What determines this is C's payoffs, which are fairly complicated here. To express the equilibrium simply, I define $U_C(\hat{t}) = X_A + \frac{2\left(\frac{(X_D-W_D)^2(\alpha+\underline{\theta})}{4}\right)^{1/2}}{\bar{\theta}-\underline{\theta}}\left((\alpha+\bar{\theta})^{1/2} - (\alpha+\underline{\theta})^{1/2}\right)$ and, defining $\tilde{t} = \frac{(\alpha+\bar{\theta})(X_D-W_D)^4}{(4(W_A-X_A)-2(X_D-W_D))^2}$ and $U_C(\tilde{t}) = \left(X_A\bar{\theta} - W_A\underline{\theta} + (W_A - X_A)\left(4\tilde{t}\left(X_D - W_D\right)^{-2} - \alpha\right) + 2(\tilde{t})^{1/2}\left(\alpha+\bar{\theta}\right)^{1/2} - 2\tilde{t}\left(X_D - W_D\right)^{-1}\right)\frac{1}{\bar{\theta}-\underline{\theta}}$.

Proposition 7:

Case 1: Under the assumptions above and $U_C(\hat{t}) \geq U_C(\tilde{t})$, the following actions are part of the perfect Bayesian Nash Equilibrium.

- *C selects the most aggressive transgression that does not induce war, or*

$$t^* = \frac{\left(X_D - W_D\right)^2 \left(\alpha + \underline{\theta}\right)}{4}.$$

- *Fix a selected level of transgression t. For value $\tilde{\theta} = \frac{4t}{(X_D - W_D)^2} - \alpha$, if D is type $\theta < \tilde{\theta}$ then they declare war (set $w = 1$), and if D is type $\theta \geq \tilde{\theta}$ then they select hassling level $h^* = t^{1/2}(\alpha + \theta)^{1/2}$ and set $w = 0$.*

- *For D, each type $\theta \in [\underline{\theta}, \theta]$ has expected utility*

$$U_D(\theta, \alpha) = X_D - \frac{\left(X_D - W_D\right)\left(\alpha + \underline{\theta}\right)^{\frac{1}{2}}}{(\alpha + \theta)^{1/2}}.$$

Case 2: *Under the assumptions above and $U_C(\hat{t}) < U_C(\tilde{t})$, the following actions are part of the perfect Bayesian Nash Equilibrium.*

- *C selects the most aggressive transgression that does not induce war, or*

$$t^* = \frac{\left(\alpha + \bar{\theta}\right)\left(X_D - W_D\right)^4}{\left(4(W_A - X_A) - 2\left(X_D - W_D\right)\right)^2}.$$

- *Fix a selected level of transgression t. For value $\tilde{\theta} = \frac{4t}{(X_D - W_D)^2} - \alpha$, if D is type $\theta < \tilde{\theta}$ then they declare war (set $w = 1$), and if D is type $\theta \geq \tilde{\theta}$ then they select hassling level $h^* = t^{1/2}(\alpha + \theta)^{1/2}$ and set $w = 0$.*

- *For D, each type $\theta \in [\underline{\theta}, \theta]$ has expected utility*

$$U^D = \begin{cases} X_D - \frac{(\alpha+\bar{\theta})^{1/2}}{(\alpha+\theta)^{1/2}} \left(\frac{(X_D - W_D)^2}{|2(W_A - X_A) - (X_D - W_D)|}\right) & if\, t^* \leq \frac{(K_C + K_D)^2(\alpha+\theta)}{4}, \\ W_D & otherwise. \end{cases}$$

The following observations describe how improvements in D's known hassling capabilities α

affect what occurs in equilibrium.

Observation: *Consider an α, α' where $\alpha < \alpha'$, and both values satisfy the conditions of Case 1. D experiences a capability failure following a shift from α to α'.*

The Observation above can be derived by taking first order conditions on U_D.[16] In this equilibrium, C is selecting a level of transgression that would make a type $\underline{\theta}$ D indifferent between war and some level of hassling. Thus, as D's hassling capabilities improve, D will hassle more, and C will select a greater transgression. However, in aggregate, these changes lead to worse outcomes for D because the change from α to α' makes D more predictable. As α increases, it shrinks the spread of h^* across types $[\underline{\theta}, \bar{\theta}]$. Thus, when C selects a t that makes a type $\underline{\theta}$ indifferent between war and hassling (as it is optimal for C to do under the conditions above), this selected t is closer to the transgression that would make a type $\theta \in (\underline{\theta}, \bar{\theta}]$ indifferent between war and a level of hassling, thus granting C more of the bargaining surplus. Within this model and this case, improvements in α decrease the relevance of the private parameter and make D more predictable, resulting in a capability failure for D.

8.4 Deriving Proposition 7 Equilibrium

When D selects a hassling level, D's decision will be based on the selected t and parameters α and θ. Any type θ D that hassles faces concave optimization problem

$$h^* \in argmax_{h \geq \underline{h}} \left\{ X_D - \frac{t}{h} - \frac{h}{(\alpha+\theta)} \right\},$$

[16] First order conditions here are $\frac{d}{d\alpha} U^D = -\frac{1}{2} \left(X_D - W_D\right) (\alpha + \underline{\theta})^{-1/2} (\alpha + \theta)^{-1/2} \left(1 - (\alpha + \underline{\theta})(\alpha + \theta)^{-1}\right)$.

I take first-order conditions with respect to h to identify the optimal level of hassling h^*. So long that $h^* \geq \underline{h}$ for a selected t, this yields

$$h^* = t^{1/2}(\alpha + \theta)^{1/2}.$$

Using h^*, I re-write D's utility in terms of the selected t and parameters α and θ, also allowing for war. This is

$$U^D = \begin{cases} X_D - \frac{2t^{1/2}}{(\alpha+\theta)^{1/2}} & if\, t \leq \frac{(X_D - W_D)^2(\alpha+\theta)}{4}, \\ W_D & otherwise. \end{cases}$$

I proceed as follows. C will select a level of transgression that induces D to hassle or declare war, or some combination of these outcomes. To determine what t C selects, I compare C's utilities across possible outcomes. Upon determining C's optimal strategy, I can run comparative statics on this equilibrium.

8.4.1 Outcome 1: D always hassles

Consider when C selects a transgression t such that $t \in [\underline{t}, \frac{(X_D - W_D)^2(\alpha+\underline{\theta})}{4}]$. For this selected transgression, D will not declare war. With D always hassling, for a t that falls in the range above, C optimizes

$$t \in argmax \left\{ \int_{\underline{\theta}}^{\bar{\theta}} \left(X_A + \frac{t}{h^*} \right) f(\theta) d\theta \right\},$$

which can be re-written as

$$U_C(t) = X_A + \frac{2t^{1/2}}{\bar{\theta} - \underline{\theta}} \left((\alpha + \bar{\theta})^{1/2} - (\alpha + \underline{\theta})^{1/2} \right).$$

Note that this is strictly increasing in t; therefore, within this range, C will select the largest $t = \frac{(X_D - W_D)^2(\alpha+\underline{\theta})}{4}$, which is the t that would make a type $\underline{\theta}$ D indifferent between war and hassling.

I can define $\hat{t} = \frac{(X_D - W_D)^2(\alpha + \underline{\theta})}{4}$ and C's utility here as

$$U_C = X_A + \frac{2\hat{t}^{1/2}}{\bar{\theta} - \underline{\theta}} \left((\alpha + \bar{\theta})^{1/2} - (\alpha + \underline{\theta})^{1/2} \right).$$

Because here D always hassles, for any type $\theta \in \Theta$, D's utility is

$$U^D = X_D - (X_D - W_D)(\alpha + \underline{\theta})^{1/2}(\alpha + \theta)^{-1/2}.$$

Taking first order conditions with respect to α yields

$$\frac{d}{d\alpha} U^D = -\frac{1}{2}(X_D - W_D)(\alpha + \underline{\theta})^{-1/2}(\alpha + \theta)^{-1/2}\left(1 - (\alpha + \underline{\theta})(\alpha + \theta)^{-1}\right).$$

Because $(X_D - W_D) > 0$ and $(\alpha + \underline{\theta})(\alpha + \theta)^{-1} < 1$, then this expression is decreasing in α.

8.4.2 Outcomes 2&3: D Sometimes Hassles, Sometimes Goes to War, & D Always Goes to War

Consider the case where C selects a $t \leq \frac{(X_D - W_D)^2(\alpha + \bar{\theta})}{4}$ and $t \geq \frac{(X_D - W_D)^2(\alpha + \underline{\theta})}{4}$. For this selected transgression, D will either hassle or declare war. For a fixed t that falls in this range, I can define a type $\tilde{\theta}$ that represents the cut-point type that is indifferent between hassling and declaring war; for any $\theta \geq \tilde{\theta}$, D will hassle, and for any $\theta < \tilde{\theta}$, D will declare war. I derive this condition comparing the war payoff to the hassling payoff (both for a fixed t). This is $\tilde{\theta} = \frac{4t}{(X_D - W_D)^2} - \alpha$.

For a t that falls in the range above, C has optimization problem

$$t \in argmax \left\{ \int_{\frac{4t}{(X_D - W_D)^2} - \alpha}^{\bar{\theta}} \left(X_A + \frac{t^{1/2}}{(\alpha + \theta)^{1/2}} \right) \frac{1}{\bar{\theta} - \underline{\theta}} d\theta + \int_{\underline{\theta}}^{\frac{4t}{(X_D - W_D)^2} - \alpha} (W_A) \frac{1}{\bar{\theta} - \underline{\theta}} d\theta \right\}.$$

The term $\int_{\frac{4t}{(X_D-W_D)^2}-\alpha}^{\bar{\theta}}(X_A)\frac{1}{\bar{\theta}-\underline{\theta}}d\theta$ becomes $X_A\frac{\bar{\theta}-4t(X_D-W_D)^{-2}-\alpha}{\bar{\theta}-\underline{\theta}}$.

The term $\int_{\frac{4t}{(X_D-W_D)^2}-\alpha}^{\bar{\theta}}\left(\frac{t^{1/2}}{(\alpha+\theta)^{1/2}}\right)\frac{1}{\bar{\theta}-\underline{\theta}}d\theta$ becomes $\frac{2t^{1/2}\left(\alpha+\bar{\theta}\right)^{1/2}-2(t)(X_D-W_D)^{-1}}{\bar{\theta}-\underline{\theta}}$

The term $\int_{\underline{\theta}}^{\frac{4t}{(X_D-W_D)^2}-\alpha}(W_A)\frac{1}{\bar{\theta}-\underline{\theta}}d\theta$ becomes $W_A * \frac{4t(X_D-W_D)^{-2}-\alpha-\underline{\theta}}{\bar{\theta}-\underline{\theta}}$. The term

Following integration, I can express the above as

$$t \in argmax\left\{\left(X_A\bar{\theta}-W_A\underline{\theta}+(W_A-X_A)\left(4t\left(X_D-W_D\right)^{-2}-\alpha\right)+2t^{1/2}\left(\alpha+\bar{\theta}\right)^{1/2}-2t\left(X_D-W_D\right)^{-1}\right)\frac{1}{\bar{\theta}-\underline{\theta}}\right\}.$$

I can take first order conditions of the above with respect to t, which yields

$$\frac{\partial U_C}{\partial t}=\frac{(\alpha+\bar{\theta})^{1/2}}{t^{\frac{1}{2}}}-\frac{4(W_A-X_A)}{\left(X_D-W_D\right)^2}-\frac{2}{\left(X_D-W_D\right)}$$

or

$$\frac{\partial U_C}{\partial t}=\frac{(\alpha+\bar{\theta})^{1/2}}{t^{\frac{1}{2}}}-\frac{4(W_A-X_A)-2\left(X_D-W_D\right)}{\left(X_D-W_D\right)^2}.$$

It is possible to solve for t. Setting the equation equal to zero and solving yields

$$t^*=\frac{(\alpha+\bar{\theta})\left(X_D-W_D\right)^4}{\left(4(W_A-X_A)-2\left(X_D-W_D\right)\right)^2}.$$

For ease, I define $\tilde{t}=\frac{(\alpha+\bar{\theta})(X_D-W_D)^4}{(4(W_A-X_A)-2(X_D-W_D))^2}$ and C's utility from this case as

$$U_C(\tilde{t})=\left(X_A\bar{\theta}-W_A\underline{\theta}+(W_A-X_A)\left(4\tilde{t}\left(X_D-W_D\right)^{-2}-\alpha\right)+2(\tilde{t})^{1/2}\left(\alpha+\bar{\theta}\right)^{1/2}-2\tilde{t}\left(X_D-W_D\right)^{-1}\right)\frac{1}{\bar{\theta}-\underline{\theta}}.$$

I can also define D's utility. For the types of D not going to war

$$U_D = X_D - \frac{2}{(\alpha+\theta)^{1/2}} \left(\frac{(\alpha+\bar{\theta})^{1/2} \left(X_D - W_D\right)^2}{|4(W_A - X_A) - 2\left(X_D - W_D\right)|} \right)$$

$$= X_D - \frac{(\alpha+\bar{\theta})^{1/2}}{(\alpha+\theta)^{1/2}} \left(\frac{\left(X_D - W_D\right)^2}{|2(W_A - X_A) - \left(X_D - W_D\right)|} \right)$$

Taking first order conditions yields

$$\frac{d}{d\alpha} U_D = -\frac{1}{2} \left(\frac{\left(X_D - W_D\right)^2}{|2(W_A - X_A) - \left(X_D - W_D\right)|} \right) \left(\alpha + \bar{\theta}\right)^{-1/2} (\alpha+\theta)^{-1/2} \left(1 - (\alpha+\bar{\theta})(\alpha+\theta)^{-1}\right).$$

Because $\left(\frac{(X_D - W_D)^2}{|2(W_A - X_A) - (X_D - W_D)|}\right) > 0$ and $(\alpha+\bar{\theta})(\alpha+\theta)^{-1} > 1$, then this expression is increasing in α.

9 General Hassling Game Analysis

In this section, I present a general results for hassling games that (a) can have a convex set of hassling options, and (b) that have a more general game form. Every game in the text falls under this framing.

9.1 Defining Hassling Games Broadly

Here I characterize a broad class of "hassling games." In the spirit of the discussion in the body of the text, this general class of games attempts to speak to the following dynamic: State C wants to transgress to improve their bargaining position, but, because θ is private, C does not know how much they can transgress before State D starts hassling to degrade C's transgression, or before D goes to war to prevent C's transgression from coming to fruition.

As I characterize them, hassling games have the following ingredients.

At the onset, nature assigns $\theta \in \Theta$, which designates D's type (respectively). The realization of θ with parameter $\alpha \in \mathcal{A}$ constitutes D's hassling capabilities. D knows nature's selection θ, but C does not. Both actors observe α and know probability distribution $f(\cdot)$ over the set Θ, where $f(\theta) > 0$ for all $\theta \in \Theta$. As a note, because the parameter α ultimately is important for analysis, I will include references to parameter α in settings where it would not typically be included.

I denote the game form G to describe a hassling game. As is standard, the game form defines a set of actions and an outcome function g that maps actions to outcomes. I describe four sets of actions. State C selects some level of transgression $t \in \mathcal{T} \subset \mathbb{R}$. State D selects some level of hassling $h \in \mathcal{H} \subset \mathbb{R}$. And both states each select some vector of actions $a_A \in A_A$ and $a_D \in A_D$ that will determine how any sort of bargaining will play out and if the players will go to war.

The outcome function $g(t, h, a_A, a_D)$ maps to two kinds of outcomes: one outcome where actors go to war, and another outcome where war is not initially declared, the selected transgression and hassling are allowed to be realized, and then states enter some crisis bargaining that is influenced by the hassling and transgression levels. For ease, I decompose outcome function g. When states go to war, war occurs before the hassling or transgression come to fruition, and states receive their wartime payoffs W_A and W_D.[17] When states do not go to war, they receive payoffs $V^g_A(h, t, a_A, a_D) - K^g_C(h, t)$ and $V^g_D(h, t, a_A, a_D) - K^g_D(h, \alpha, \theta)$. The V^g functions define, for a specific outcome function g, the value each state attains through future bargaining that is influenced by the transgression and hassling. To capture the dynamics of t and h described earlier, V^g_A is increasing in t and decreasing in h, while V^g_D is increasing in h and decreasing in t (in all cases weakly).[18] The C^g functions denote

[17] This assumption is distinct from most applications of mechanism design in international relations where a state's private type affects their wartime payoff (see Banks 1991 and Fey and Ramsay 2011).

[18] One way to interpret the V function is that, after the transgression and hassling comes to fruition, enter a new round of bargaining, where t and h affect C's and D's future bargaining positions. This is discussed

the costs each state attains from hassling and the transgression. State C incurs costs from both hassling and transgressing, making K_C weakly increasing in h and t. State D incurs costs from hassling, making K_D weakly increasing in h, and K_D is strictly decreasing in the parameter and type (α and θ, respectively).[19] Finally, $\pi^g : (A_A, A_D) \to \{0, 1\}$ is a function that maps from actions a_A and a_D to the decision to go to war. The expected utilities from a given action profile are

$$u^g_A(a_A, t, a_D, h) = (1 - \pi^g(a_A, a_D))W_A + \pi^g(a_A, a_D)\left(V^g_A(h, t, a_A, a_D) - K^g_C(t, h)\right) \tag{8}$$

and

$$u^g_D(a_D, h, a_A, t \,|\, \alpha, \theta) = (1 - \pi^g(a_A, a_D))W_D + \pi^g(a_A, a_D)\left(V^g_D(h, t, a_A, a_D) - K^g_D(h, \alpha, \theta)\right). \tag{9}$$

The actions taken in the game depend the observed and unobserved components of D's capabilities. A strategy for State D in the game is a mapping from their capabilities to their action space, or $\sigma_D : (\mathcal{A}, \Theta) \to (A_D, \mathcal{H})$. Because State C does not know the value of θ, C's strategy is as mapping only from the known parameter α, or $\sigma_A : (\mathcal{A}) \to (A_A, t)$. Thus together, a strategy profile (σ_A, σ_D) and capabilities (α, θ) generate a likelihood of war $(\pi^g(\sigma_A(\alpha), \sigma_D(\alpha, \theta))$ and a payoff from not going to war $(V^g_D(\sigma_A(\alpha), \sigma_D(\alpha, \theta)) - K^g_D(\sigma_D(\alpha, \theta), \alpha, \theta)$ and $V^g_A(\sigma_A(\alpha), \sigma_D(\alpha, \theta)) - K^g_C(\sigma_A(\alpha), \sigma_D(\alpha, \theta)))$. I let σ denote a pair of strategies or $\sigma = (\sigma_A, \sigma_D)$, and I employ the equilibrium concept Bayesian Nash Equilibrium (Meyerson 1985), where a strategy profile $\sigma^* = (\sigma_A, \sigma_D)$ constitutes a Bayesian equilibrium if $\sigma_D(\alpha, \theta)$ is a best response to σ_A, and $\sigma_A(\alpha)$ is a best response to σ_D based on the known parameter and the common probability prior $f(\cdot)$. I limit my analysis to pure strategy equilibria. I define for a given equilibrium σ^* to a given game form G, $U_D(\theta, \alpha)$ and $U_C(\theta, \alpha)$ are

in Example 2 below.

[19] This is a departure from several previously considered models that had K_D weakly decreasing in α and θ (specifically when $h = \underline{h}$, the minimum of set $\mathcal{H}$).

the expected utilities, or

$$U_D(\theta,\alpha) = u_D^g\left(\sigma^*(\alpha,\theta) \,|\, \alpha,\theta\right) \tag{10}$$

and

$$U_C(\alpha) = \int_\Theta u_A^g\left(\sigma^*(\alpha,\tau)\right) f(\tau) d\tau.$$

With this defined, I can introduce the condition of interest: when a capability failure occurs.

9.2 Introducing the Direct Mechanism Analysis

In the paper and previous sections, I discuss specific equilibria to specific game forms. Using the sparse structure that I defined above, I can discuss the general properties of Bayesian equilibria to any game that follows that structure. For other papers utilizing mechanism design in similar conflict settings, see Banks (1990), Fey and Ramsay (2011), or Fey and Kenkel (2019).

In the previous subsection – and in all the examples and general analysis in the text – I described, for some game form G and some Bayesian equilibrium σ^*, how capabilities (α and θ) affect strategic play which affect outcomes and utilities through the functions π^g, V_A^g, V_D^g, K_C^g and K_D^g. Instead of this "indirect" mechanism – "indirect" because type operates indirectly on outcomes via the game form and strategies – it is possible to consider a "direct" mechanism where state D reports their type private θ, and this type directly informs outcomes and payoffs (D does not need to report α because this is common knowledge). This direct mechanism can be thought of as State D reporting their type θ to a neutral intermediary, and then the neutral intermediary selects the actions within game form that State C and State D would have selected through their equilibrium σ^*. Focusing on D, instead of working through indirect outcome functions and strategies to the hassling game $\pi^g(\sigma^*(\alpha,\theta))$, $V_D^g(\sigma^*(\alpha,\theta))$, and $K_D^g(\sigma^*(\alpha,\theta),\alpha,\theta)$, the new direct mechanism considers outcome functions $\pi(\tilde{\theta},\alpha)$, $V_D(\tilde{\theta},\alpha)$ and $K_D(\tilde{\theta},\theta,\alpha)$, which are functions of the reported type $\tilde{\theta}\in\Theta$, the true unobserved (by A) type $\theta\in\Theta$, and the commonly observed parameter $\alpha\in\mathcal{A}$. When the outcomes and payoffs of the components of the direct mechanism equal the respective components of the indirect mechanism, then the direct mechanism is an "equivalent" direct mechanism (i.e. $\pi(\tilde{\theta},\alpha)=\pi^g(\sigma^*(\alpha,\theta))$, $V_D(\tilde{\theta},\alpha)=V_D^g(\sigma^*(\alpha,\theta))$, $K_D(\tilde{\theta}|\theta,\alpha)=K_D^g(\sigma^*(\alpha,\theta),\alpha,\theta)$, and so on).

In the direct mechanism, State D's action is sending $\tilde{\theta}\in\Theta$, which is a message of their type. One could imagine that when faced with some direct mechanism, a type θ D may be incentivised to lie about their type (i.e. send $\tilde{\theta}\neq\theta$), as it could grant D a greater payoff than truthfully reporting their type (i.e. send $\tilde{\theta}=\theta$). The following definition and result pertain to this issue.

Definition: *A direct mechanism is "Incentive Compatible" if and only if it is a Bayesian Nash Equilibrium for State D to truthfully report their type. Formally, for all* $\theta,\tilde{\theta}\in\Theta$,

$$(1-\pi(\theta|\alpha))\,W_D+\pi(\theta|\alpha)\,(V_D(\theta|\alpha)-K_D(\theta|\theta,\alpha))\geq\Big(1-\pi(\tilde{\theta}|\alpha)\Big)\,W_D+\pi(\tilde{\theta}|\alpha)\,\Big(V_D(\tilde{\theta}|\alpha)-K_D(\tilde{\theta}|\theta,\alpha)\Big).$$

The logic of incentive compatibility is as follows: if D can attain a greater utility by misreporting their type (by reporting some $\tilde{\theta}\neq\theta$), then the mechanism is not incentive compatible. Incentive compatible mechanisms are important due to the Revelation Principle.

Lemma 1 *(Revelation Principle (Myerson, 1979)): If σ^* is a Bayesian Nash equilibrium of game form G, then there exists an incentive compatible direct mechanism yielding the same outcome.*

The Revelation Principle is useful here. Instead of focusing on the properties of the set of Bayesian Nash equilibria of any game form G, it is possible to instead consider the properties of the set of equivalent incentive compatible direct mechanisms. So, if a property is shown for the set of all incentive compatible direct mechanisms, then this property will hold for the set of all Bayesian Nash equilibria. This allows me to avoid discussions on a chain of actions describing how D decided to go to war or how any kind of bargaining plays out after hassling and transgression are realized, and instead directly considers how the factors that this paper is concerned with – hassling capabilities – affects relevant outcomes – how well the state that developed the hassling capabilities does.

Before moving forward with the analysis, I introduce the following assumptions.

General Analysis Assumptions: *For hassling game G, I assume that $K_D^g(h, \theta, \alpha)$ is differentiable in θ for all $h \in H$ and $\alpha \in \mathcal{A}$, and that the derivative of K_D^g is uniformly bounded, or that, for all α and h, $|\frac{\partial}{\partial \theta} K_D^g(h, \theta, \alpha)| \leq M < \infty$.*

Voluntary Agreements (Individual Rationality) Assumption *: I assume there exists an action profile a'_D such that either $\pi(a'_D, a_A) = 0$ for all $a_A \in A_A$, or $V_D(h, t, a_A, a'_D) - K_D(h, \alpha, \theta) \geq W_D$ for all $h \in H$, $t \in T$, $a_A \in A_A$, $\theta \in \Theta$, and $\alpha \in \mathcal{A}$.*

It is worthwhile highlighting how broad these assumptions are. On the General Analysis Assumption, I am making no restrictions on the set of h and α, nor I am not making any of the standard concavity assumptions, differentiability assumptions in h, or assumptions on V_D^g. On the Voluntary Agreements Assumption, I am simply assuming that D has the ability to go to war to prevent an outcomes where D attains less utility than what they could get from going to war; in other words, there is no way that D can become "stuck" in a non-war outcome that is worse for them than war.

Finally, using the new notation of the direct mechanism, I can characterize when a capability failure occurs following improvements in hassling capabilities.

Definition:*In crisis bargaining game G with equilibrium σ^*, D experiences a* ***capability failure*** *following a shift from α to α' when $U_D(\theta, \alpha') \leq U_D(\theta, \alpha)$ for all $\theta \in \Theta$. Additionally, D experiences a* ***capability failure*** *following a shift from θ to θ' when $U_D(\theta', \alpha) \leq U_D(\theta, \alpha)$*

for all $\alpha \in \mathcal{A}$.

9.3 General Results on θ

In this section, I show how changes in D's private hassling capabilities affect D's outcome. In this section, I assume that, for a given $\alpha \in \mathcal{A}$, there exists a Bayesian Nash equilibrium. Given the assumed existence, I first establish Lemma 2, which demonstrates that the likelihood of war is weakly decreasing in θ. I then use Lemma 2 to show Proposition 8, which shows that D's utility must be weakly increasing in θ. This result shows that there cannot be the case where D's private hassling capabilities improves and D does worse. In other words, improvements in D's private hassling capabilities cannot produce a capability failure; this finding confirms that the model-specific results above and in the paper are not just quirks of the selected functional forms or game forms, but a condition that must hold in hassling games.

I start with Lemma 2, which establishes the general properties of $\pi(\theta)$.

Lemma 2: *Consider the set of hassling games G . Assume that there exists a pure strategy Bayesian Nash Equilibrium for parameter $\alpha \in \mathcal{A}$ and for all types $\theta \in \times$. For $\theta \in \Theta$, $\theta' \in \Theta$, if $\theta < \theta'$, then $\pi(\theta') \geq \pi(\theta)$.*

Proof by contrapostive: Assume that there exists some $\theta \in \Theta$ and $\theta' \in \Theta$ such that $\pi(\theta') < \pi(\theta)$. Using incentive compatibility, I can say

$$\pi(\theta')\left(V_D(\theta'|\alpha) - K_D(\theta'|\theta',\alpha)\right) + \left(1-\pi(\theta')\right)W_D \geq \pi(\theta)\left(V_D(\theta|\alpha) - K_D(\theta|\theta',\alpha)\right) + \left(1-\pi(\theta)\right)W_D, \tag{11}$$

and

$$\pi(\theta)\left(V_D(\theta|\alpha) - K_D(\theta|\theta,\alpha)\right) + \left(1-\pi(\theta)\right)W_D \geq \pi(\theta')\left(V_D(\theta'|\alpha) - K_D(\theta'|\theta,\alpha)\right) + \left(1-\pi(\theta')\right)W_D. \tag{12}$$

Because I only consider pure strategies, this implies $\pi(\theta) = 1$ and $\pi(\theta') = 0$. Re-writing the incentive compatibility conditions yields

$$W_D \geq (V_D(\theta|\alpha) - K_D(\theta|\theta', \alpha)), \tag{13}$$

and

$$V_D(\theta|\alpha) - K_D(\theta|\theta, \alpha) \geq W_D. \tag{14}$$

Combining and simplifying the inequalities above yields

$$K_D(\theta|\theta', \alpha) \geq K_D(\theta|\theta, \alpha).$$

Given K_D is strictly decreasing in θ, this final condition implies that $\theta' \leq \theta$. $\square$

Having established Lemma 2, I can demonstrate that improvements in private hassling capabilities (moving from θ to θ' with $\theta < \theta'$) always produces better results for D.

Proposition 8: *Consider the set of hassling games G with Voluntary Agreements. Assume that there exists a pure strategy Bayesian Nash Equilibrium for parameter $\alpha \in \mathcal{A}$ and for all types $\theta \in \times$. For $\theta \in \Theta$, $\theta' \in \Theta$, if $\theta < \theta'$, then $U(\theta') \geq U(\theta)$.*

Proof by contrapostive: Suppose $U_D(\theta', \alpha) < U_D(\theta, \alpha)$. This implies

$$\pi(\theta)\left(V_D(\theta|\alpha) - K_D(\theta|\theta, \alpha)\right) + (1 - \pi(\theta))\, W_D > \pi(\theta')\left(V_D(\theta'|\alpha) - K_D(\theta'|\theta', \alpha)\right) + (1 - \pi(\theta'))\, W_D.$$

I can simplify the above by considering the different values that π can take. I proceed by cases. By Lemma 2, I only need to consider the three cases listed below.

Case 1: $\pi(\theta') = \pi(\theta) = 0$.

In this case, the "if" clause implies

$$W_D > W_D,$$

which is not possible.

Case 2: $\pi(\theta') = 1$ *and* $\pi(\theta) = 0$.

In this case, the "if" clause implies

$$W_D > V_D(\theta'|\alpha) - K_D(\theta'|\theta', \alpha),$$

which violates the Voluntary Agreements assumption.

Case 3: $\pi(\theta') = 1$ *and* $\pi(\theta) = 1$.

In this case, the "if" clause implies

$$V_D(\theta|\alpha) - K_D(\theta|\theta, \alpha) > V_D(\theta'|\alpha) - K_D(\theta'|\theta', \alpha).$$

Using incentive compatibility, I can say

$$V_D(\theta|\alpha) - K_D(\theta|\theta, \alpha) > V_D(\theta|\alpha) - K_D(\theta|\theta', \alpha).$$

Simplifying the above gives

$$K_D(\theta|\theta', \alpha) > K_D(\theta|\theta, \alpha).$$

Because because $K_D(\tilde{\theta}|\theta, \alpha)$ *is decreasing in true type* θ*, the above implies that* $\theta' < \theta$. $\square$.

9.4 General Results on α

In this section, I can define the necessary conditions for a capability failure following a shifts in α. This is somewhat more difficult to accomplish because changes in α can produce shifts in how C plays the game. While before I was able to derive how shifts in θ affect D's utility without actually characterizing D's utility (or using the General Analysis Assumptions), here I cannot do the same, and must add additional structure and utilize the Milgrom and Segal (2002) envelope theorem result.

At this point, I introduce two ways that C's choices in the hassling game G can influence the final outcomes of D's utility. These two ways characterize how C plays the game and offers some simplification to the conditions for a capability failure below. Because I will be working with the direct mechanism for the rest of the game, I provide the following definitions in terms of the direct mechanism.

Definition: *In crisis bargaining game G with equilibrium σ^*, C is "**unconstrained**" in parameter $\alpha \in \mathcal{A}$ when $U_D(\underline{\theta}, \alpha) = W_D$.*

C is unconstrained when C is not sufficiently deterred from transgressing. As I have defined type, type $\underline{\theta}$ incurs the greatest costs from hassling, and thus, for a given t, is the least willing to hassle and the most willing to go to war. When C is "unconstrained," it implies that C either selects a transgression that induces a D with capabilities $(\underline{\theta}, \alpha)$ to go to war (as it was in Case 2 in Proposition 1), or selects a transgression that induces D's with capabilities $(\underline{\theta}, \alpha)$ to optimally select a level of hassling that makes them indifferent between war and non-war outcomes (as it was in Case 1 in Proposition 1).[20] When would C be "constrained"? In the example where C faces costs to selecting t, sometimes C was constrained because their costs from transgressing were too high to justify selecting a t that gave type $\underline{\theta}$ D their wartime payoff.

[20] Also note that voluntary agreements prevents $U_D(\underline{\theta}, \alpha) < W_D$.

Definition: *In crisis bargaining game G with equilibrium σ^*, C "**avoids war**" if, for all $\theta \in \Theta$, $\pi(\theta, \alpha) = 1$.*

When C "avoids war," C is not willing to select a transgression t that will ever result in D responding with war. If, for α and α', $\pi(\theta, \alpha) = \pi(\theta, \alpha') = 1$ for all $\theta \in \Theta$, then C is not emboldened by a shift in D's hassling capabilities.

***Proposition** 9: Consider the set of hassling games G . Let $\alpha, \alpha' \in \mathcal{A}$ with $\alpha' > \alpha$, and let $\sigma^*(\alpha)$ and $\sigma^*(\alpha')$ denote the pure strategy Bayesian Nash equilibria to these games under parameters α and α'. If the General Analysis Assumptions hold, then $U_D(\theta, \alpha)$ is absolutely continuous and differentible almost everywhere in θ, and can be expressed as $U_D(\theta, \alpha) = U_D(\underline{\theta}, \alpha) + \int_{\underline{\theta}}^{\theta} \frac{\partial}{\partial \theta} \left(-\pi(\tau|\alpha) K_D(\tau|\tau, \alpha) d\tau\right)$.*[21] *Additionally, the following conditions hold:*

(i.) If C "avoids war", is "unconstrained" in α and α', and

$$\int_{\underline{\theta}}^{\theta} \left(\frac{\partial K_D}{\partial \theta} (\tau|\tau, \alpha') - \frac{\partial K_D}{\partial \theta} (\tau|\tau, \alpha) \right) d\tau \geq 0,$$

in the direct mechanism, then improvements in public hassling capabilities produce a capability failure.

(ii.) if C is "unconstrained" in α and α' and

$$\int_{\underline{\theta}}^{\theta} \pi(\tau|\alpha') \left(\frac{\partial K_D}{\partial \theta} (\tau|\tau, \alpha') \right) d\tau - \int_{\underline{\theta}}^{\theta} \pi(\tau|\alpha) \left(\frac{\partial K_D}{\partial \theta} (\tau|\tau, \alpha) \right) d\tau \geq 0$$

in the direct mechanism, then improvements in public hassling capabilities produce a capability failure.

[21] *Referencing earlier definitions, $\pi(\theta, \alpha) = \pi^g(\sigma^*(\theta, \alpha))$, $K_D(\theta, \theta, \alpha) = K_D^g(\sigma^*(\theta, \alpha), \sigma, \alpha)$, $\pi(\theta, \alpha') = \pi^g(\sigma^*(\theta, \alpha'))$, and $K_D(\theta, \theta, \alpha') = K_D^g(\sigma^*(\theta, \alpha'), \sigma, \alpha')$.*

(iii.) If

$$U_D(\underline{\theta}, \alpha) - U_D(\underline{\theta}, \alpha') + \int_{\underline{\theta}}^{\theta} \pi(\tau|\alpha') \left(\frac{\partial K_D}{\partial \theta} (\tau|\tau, \alpha') \right) d\tau - \int_{\underline{\theta}}^{\theta} \pi(\tau|\alpha) \left(\frac{\partial K_D}{\partial \theta} (\tau|\tau, \alpha) \right) d\tau \geq 0$$

in the direct mechanism, then improvements in public hassling capabilities produce a capability failure.

Proof: See next subsection.

The next few paragraphs discuss the logic of Proposition 9. The structure of Proposition 9 is framed to describe the least general case first, and then move into more general cases. The critical take-away from Proposition 9 is the relationship between improved hassling capabilities and capability failure is determined by three factors: whether improved hassling capabilities results in D becoming more predictable, C becoming emboldened, or C becoming constrained.

When C is unconstrained and avoid wars (Case i. in Proposition 9), without war ever occurring, increases in hassling capabilities can diminish D's surplus by reducing the impact of D's private information, essentially making D more predictable. Modifying the expression in Proposition 3 in the text, within this case it is possible to represent a type θ parameter α D's utility as

$$U_D(\theta, \alpha) = W_D - \int_{\underline{\theta}}^{\theta} \frac{\partial K_D}{\partial \theta} (\tau|\tau, \alpha) \, d\tau.$$

The W_D term represents the costs from going to war. The expression $\int_{\underline{\theta}}^{\theta} \frac{\partial K_D}{\partial \theta} (\tau|\tau, \alpha) \, d\tau$ is the partial derivative of D's cost function with respect to their true type integrated over their true type when their reported type equals their true type. While this is a complicated expression, there are some ways to make this conceptually simpler. As one example, assume that when faced with any type $\alpha \in \mathcal{A}$, C plays the same t; this is how it was in the model in the text. Also within this example, assume that D "accepting" is playing $\underline{h}$, D hassling is playing h', and these two actions constitute the entire set $\mathcal{H}$. If this is the case, the $\int_{\underline{\theta}}^{\theta} \frac{\partial K_D}{\partial \theta}(\tau|\tau,\alpha)\, d\tau$ expression can be simplified to $K_D(h'|\theta,\alpha) - K_D(h'|\underline{\theta},\alpha)$, making the full for a capability failure $K_D(h'|\theta,\alpha') - K_D(h'|\underline{\theta},\alpha') - (K_D(h'|\theta,\alpha) - K_D(h'|\underline{\theta},\alpha)) \geq 0$; essentially, this condition most directly shows that when there is a larger difference in the effect of costs for types θ and $\underline{\theta}$ for parameter α relative to the difference in costs between types for parameter α', this produce a capability failure.[22] So while here there are more moving parts to this simpler expression, the integral equation captures these additional moving parts, including the following: (a) the partial effect of D's true type θ on their selected h; (b) how different types of D's play the game differently and how this changes C's selected t; both of which intuitively could be thought of as "predictability."

When C is unconstrained but does not always avoid war (Case ii. in Proposition 9), now whether C is emboldened to go to war by a change in α plays a role. Modifying the expression in Proposition 9, within this case it is possible to represent a type θ parameter α D's utility as

$$U_D(\theta,\alpha) = W_D - \int_{\underline{\theta}}^{\theta} \pi(\tau|\alpha) \frac{\partial K_D}{\partial \theta}(\tau|\tau,\alpha)\, d\tau.$$

Outside of D's wartime utility, D's expected utility now consists of the $\int_{\underline{\theta}}^{\theta} \pi(\tau,\alpha) \frac{\partial K_D}{\partial \theta}(\tau|\tau,\alpha)\, d\tau$ term. In addition to accounting for the spread in costs across private types, this expression now also takes into account whether a type θ goes to war ($\pi(\theta|\alpha) = 0$) or not. When C is emboldened, it is not just that D must go to war (and receive their wartime payoff) more; all type D's that are not willing to go to war also incur losses because the spread of possible hassling responses are truncated, meaning that C's greater selected transgression is also eating into D's bargaining surplus.

Finally, when C is constrained (Case iii.), the utility that type $\underline{\theta}$ D attains becomes a salient factor. In the model in the text, C selected a transgression where type $\underline{\theta}$ either went to war

[22] It is useful to note that because K_D is decreasing in θ, $K_D(h'|\theta,\alpha) - K_D(h'|\underline{\theta},\alpha) < 0$.

(Proposition 1, Case 2) or hassled and was indifferent between war and hassling (Proposition 1, Case 1). However, sometimes the costs that C incurs from hassling or from transgressing (or both) makes C hold back on their level of transgression, which can produce a value for $U_D(\underline{\theta}, \cdot) \neq W_D$. Alternatively, this case describes what happens when changes in D's hassling capabilities also alter D's wartime capabilities. For example, if $U_D(\underline{\theta}, \alpha) < U_D(\underline{\theta}, \alpha')$, then a capability failure can never occur, because types $\underline{\theta}$ always do better following improvements in hassling capabilities. Alternatively, if $U_D(\underline{\theta}, \alpha) > U_D(\underline{\theta}, \alpha')$, then capability failures can still occur, but not necessarily (it will be contingent upon the remaining terms).

As a final note, it is worthwhile mentioning that capability failures can arise without the changes in α interacting with θ in new and different ways. Within a possible hassling game, there can be multiple equilibria (though this was ruled out in the analysis above). The analysis on changes in θ was made within the context of a single equilibrium because C could not observe the (private) θ term and change into a different equilibrium. If there are multiple equilibria, then changes from a low-to-high α could essentially function as a signal to switch from one equilibrium to another. This means that there can be functional forms that could create a capabilities failure for D, even if they do not appear, at first glance, to have the "right" interactions between α and θ.

9.5 Proof of Proposition 9

What follows here is a derivation of Milgrom and Segal (2002) envelope theorem results.[23] Lemma 1 implies that for hassling game G and equilibrium σ^*, there exists an incentive compatible direct mechanism yielding the same outcome. For D, the direct mechanism is given by $\pi(\theta|\alpha) = \pi^g(\sigma^*(\theta, \alpha))$, $V_D(\theta'|\alpha) = V_D^g(\sigma^*(\alpha, \theta))$ and $K_D(\theta|\theta, \alpha) = K_D^g(\sigma^*(\theta, \alpha)|\theta, \alpha)$. I define $\Phi(\theta'|\theta, \alpha)$ as the utility a type (θ, α) D receives from announcing type θ',or

$$\Phi_D(\theta'|\theta, \alpha) = (1 - \pi(\theta'|\alpha))W_D + \pi(\theta'|\alpha)\left(V_D(\theta'|\alpha) - K_D(\theta'|\theta, \alpha)\right).$$

[23] A close version of this was provided in Ilya Segal's outstanding notes on mechanism design.

With this notation, I can re-write D's utility from equation (10) in the notation of the direct mechanism with truthful revelation occurring, or

$$U_D(\theta,\alpha) =\Phi_D(\theta|\theta,\alpha) = (1-\pi(\theta|\alpha))W_D + \pi(\theta|\alpha)\left(V_D(\theta|\alpha) - K_D(\theta|\theta,\alpha)\right).$$

I can define the incentive compatibility conditions as $\Phi_D(\theta|\theta,\alpha) \geq \Phi(\theta'|\theta,\alpha)$ for all $\theta, \theta' \in \Theta$, or

$$\theta \in argmax_{\hat{\theta}\in\Theta}\, \Phi_D(\hat{\theta}|\theta,\alpha).$$

Logically, when incentive compatibility holds, $\Phi_D(\theta'|\theta,\alpha) \leq U_D(\theta,\alpha)$ for any $\theta' \neq \theta$. This means I can re-write the incentive compatibility constraint as

$$\theta' \in argmax_{\theta\in\Theta} \left\{\Phi_D(\theta'|\theta,\alpha) - U_D(\theta,\alpha)\right\}. \tag{15}$$

Because $K_D(\theta'|\theta,\alpha)$ (and therefore $\Phi_D(\theta'|\theta,\alpha)$) is differentiable in θ, for any point $\theta' \in \Theta$ where $\frac{d}{d\theta}U_D(\theta',\alpha)$ exists (I will discuss existence soon), differentiating equation 15 with respect to θ at $\theta = \theta'$ yields

$$\left[\frac{\partial}{\partial\theta}\Phi_D(\theta'|\theta,\alpha) - \frac{d}{d\theta}U_D(\theta,\alpha)\right]|_{\theta=\theta'} = 0,$$

or

$$\frac{d}{d\theta}U_D(\theta',\alpha) =\frac{\partial}{\partial\theta}\Phi_D(\theta'|\theta',\alpha) = \frac{\partial}{\partial\theta}\left(-\pi(\theta'|\alpha)K_D(\theta'|\theta',\alpha)\right).$$

The logic here is that the total derivative of D's utility with respect to their true type is equal to the direct effect of D's true type on the cost function, without needing to consider the indirect effects of true type θ on reported type θ'. Additionally (and I will use this in the next part of the proof), when U_D is differentiable in type, I can see that D's utility is weakly increasing in type because, by assumption, $\frac{\partial}{\partial\theta}K_D \leq 0$.

However, I haven't defined if function $U_D(\theta, \alpha)$ is ever differentiable in θ. By the definition of incentive compatibility, I can say

$$U_D(\theta', \alpha) - U_D(\theta, \alpha) \leq \Phi_D(\theta'|\theta', \alpha) - \Phi_D(\theta'|\theta, \alpha)$$

or

$$U_D(\theta', \alpha) - U_D(\theta, \alpha) \leq \pi(\theta'|\alpha)\left(-K_D(\theta'|\theta', \alpha) + K_D(\theta'|\theta, \alpha)\right) \tag{16}$$

And similarly

$$U_D(\theta', \alpha) - U_D(\theta, \alpha) \geq \Phi_D(\theta|\theta', \alpha) - \Phi_D(\theta|\theta, \alpha) = \pi(\theta|\alpha)\left(-K_D(\theta|\theta', \alpha) + K_D(\theta|\theta, \alpha)\right) \tag{17}$$

Case 1: Assume $\theta' > \theta$. This implies $K_D(\theta'|\theta', \alpha) \leq K_D(\theta'|\theta, \alpha)$.

Under the condition that $|\frac{\partial}{\partial\theta}K_D(\theta'|\theta, \alpha)| \leq M$ and that $\pi(\theta'|\alpha) \in [0, 1]$, I can say

$$U_D(\theta', \alpha) - U_D(\theta, \alpha) \leq \pi(\theta'|\alpha)\left(-K_D(\theta'|\theta', \alpha) + K_D(\theta'|\theta, \alpha)\right) \leq \left(-K_D(\theta'|\theta', \alpha) + K_D(\theta'|\theta, \alpha)\right) \leq M(\theta' - \theta).$$

The first inequality holds by the IC constraints above. The second holds because because $\pi(\theta', \alpha) \in \{0, 1\}$ and $-K_D(\theta'|\theta', \alpha) + K_D(\theta'|\theta, \alpha)$ is positive. The third holds because $|\frac{\partial}{\partial\theta}K_D(\theta'|\theta, \alpha)| \leq M$ implies (given that $\frac{\partial}{\partial\theta}K_D$ is negative) $-\frac{\partial}{\partial\theta}K_D(\theta'|\theta, \alpha) \leq M$.

Through symmetry, I can say $U_D(\theta, \alpha) - U_D(\theta', \alpha) \geq M(\theta - \theta')$

Case 2: Assume $\theta' < \theta$. This implies $K_D(\theta|\theta, \alpha) \leq K_D(\theta|\theta', \alpha)$. From the second IC constraint, because $\pi(\theta, \alpha) \in \{0, 1\}$ and $-K_D(\theta|\theta', \alpha) + K_D(\theta|\theta, \alpha)$ is negative, and because

$|\frac{\partial}{\partial\theta}K_D(\theta'|\theta,\alpha)| \leq M$

$U_D(\theta,\alpha) - U_D(\theta',\alpha) \geq \pi(\theta|\alpha)\left(-K_D(\theta|\theta',\alpha) + K_D(\theta|\theta,\alpha)\right) \geq \left(-K_D(\theta|\theta',\alpha) + K_D(\theta|\theta,\alpha)\right) \geq M(\theta' - \theta)$

And by symmetry $U_D(\theta,\alpha) - U_D(\theta',\alpha) \leq M(-\theta' + \theta)$.

Overall, for any $\theta, \theta' \in \Theta$, it holds that $|U_D(\theta',\alpha) - U_D(\theta,\alpha)| \leq M|\theta' - \theta|$. Thus, I can say that $U_D(\theta,\alpha)$ is Lipshitz continuous in θ, which implies that $U_D(\theta,\alpha)$ is absolutely continuous in θ and differentiable almost everywhere in θ. Therefore, I can express U_D as the integral of its derivative, which is

$$U_D(\theta,\alpha) = U_D(\underline{\theta},\alpha) + \int_{\underline{\theta}}^{\theta} \frac{\partial}{\partial\theta}\left(-\pi(\tau|\alpha)K_D(\tau|\tau,\alpha)d\tau\right). \tag{18}$$

I can then derive the conditions above through algebra. □

References

Baliga, S., De Mesquita, E. B. and Wolitzky, A. (2020) Deterrence with imperfect attribution, *American Political Science Review*, **114**, 1155–1178. 5.3.1, 10

Banks, J. S. (1990) Equilibrium behavior in crisis bargaining games, *American Journal of Political Science*, pp. 599–614. 9.2

Clausewitz, C. v. (1873) *On War*, vol. 1, London, N. Trübner & Company. 1.4

Fearon, J. D. (1997) Signaling foreign policy interests tying hands versus sinking costs, *Journal of Conflict Resolution*, **41**, 68–90. 5.3, 3, 5.3.2

Fey, M. and Kenkel, B. (2019) Is an ultimatum the last word on crisis bargaining? 9.2

Fey, M. and Ramsay, K. W. (2011) Uncertainty and incentives in crisis bargaining: Game-free analysis of international conflict, *American Journal of Political Science*, **55**, 149–169. 9.2

Gartzke, E. (1999) War is in the error term, *International Organization*, pp. 567–587. 5

Gurantz, R. and Hirsch, A. V. (2017) Fear, appeasement, and the effectiveness of deterrence, *The Journal of Politics*, **79**, 1041–1056. (document), 7

Milgrom, P. and Segal, I. (2002) Envelope theorems for arbitrary choice sets, *Econometrica*, **70**, 583–601. 9.4, 9.5

Myerson, R. B. (1979) Incentive compatibility and the bargaining problem, *Econometrica*, pp. 61–73. 9.2

Pecorino, P. (2019) Bridge burning and escape routes, *Public Choice*, pp. 1–16. 5.3.1

Schelling, T. C. (1966) *Arms and influence*, Yale University Press. 5.3, 5.3.1, 5.3.1

Spaniel, W. (2014) *Game theory 101: the complete textbook*, CreateSpace. 5.3, 5.3.1, 1

Tirole, J. (1988) *The theory of industrial organization*, MIT press. 5.3.1