When Deterrence Fails Appendix

Peter Schram

Preliminary draft.

April 28, 2020

Abstract

In Part I, I discuss proofs for the model in the paper (Sections 1-4), and solve for a variation of the model in the paper that has A pay costs for investing in t (Section 5). I conclude this part by describing justification for how I define a "deterrence failure" (Section 6). In Part II, I consider scenarios where θ is a continuous random variable rather than a dichotomous random variable. In Section 7, I include a sample model, and in Section 8, I utilize mechanism design to present general results on deterrence failure. 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 deterrence failure.

Contents

I Discussion of the Example Model and Extensions. 3

1 In-Paper Model Equilibrium 3

1.1 Intuition . . . 3
1.2 Equilibrium . . . 6
1.3 A's Behavior within Predictability and Emboldneing Mechanisms . . . 6
1.4 Predictability and Poker . . . 9

2 Proposition 2 Proof 10

3 Proposition 3 Proof 10

4 Proposition 4 Proof 11

5 Observation 2 Proof 13

5.1 When a Welfare Loss Doesn't Occur . . . 15

6 Example Model With Costs from Investing in t 15

6.1 Model Assumptions . . . 15
6.2 Equilibrium Preliminaries, Intuition, and Equilibrium . . . 16
6.3 Results . . . 20
6.4 Proving Proposition 6 . . . 24
6.4.1 Proving $\leftarrow$ . . . 24
6.4.2 Proving $\rightarrow$ . . . 30
6.5 Other Results . . . 37

7 On the Definition of Deterrence Failure 37

II Analysis for Models with a Convex Set of Hassling Levels 38

8 Example Model with θ as a Continuum 38

8.1 Game Form . . . 39
8.2 Equilibrium Type, Assumptions, and Deterrence Failure . . . 40
8.3 Equilibrium and Results . . . 41
8.4 Deriving Proposition 7 Equilibrium . . . 43
8.4.1 Outcome 1: D always hassles . . . 44

8.4.2 Outcomes 2&3: D Sometimes Hassles, Sometimes Goes to War, & D Always Goes to War 44

9 General Hassling Game Analysis 46

9.1 Defining Hassling Games Broadly 46

9.2 Introducing the Direct Mechanism Analysis 48

9.3 General Results on θ 51

9.4 General Results on α 53

9.5 Proof of Proposition 10 57

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

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

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

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

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

$$EU_A = Pr(\underline{\theta}) * (\rho + t - h^*(\underline{\theta}) - \omega_A) + Pr(\bar{\theta}) * \left(\rho + t - h^*(\bar{\theta}) - \omega_A\right)$$

or

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

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

$$EU_A = Pr(\underline{\theta}) * (\rho - \omega_A) + Pr(\bar{\theta}) * \left(\rho + t - h^*(\bar{\theta}) - \omega_A\right)$$

or

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

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

$$EU_A = \rho - \omega_A. \tag{3}$$

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 A 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 investing in the rising technology; therefore, A is only ever constrained by an increased likelihood of war. Therefore, A 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 A incurs more war. Thus, A 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

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

or

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

A's decision results in no war (when A benchmarks their t to make type $\underline{\theta}$ indifferent) or sometimes war (when A benchmarks their t to make type $\bar{\theta}$ indifferent) depending on A's willingness to sometimes risk war to achieve greater levels of rising technology. I can identify A's expected utility from never going to war. Here A selects a t that makes a type $\underline{\theta}$ indifferent between war and hassling. This is $t(\underline{\theta})$ (which I will also sometimes write as $t(\underline{\theta}, \alpha)$ for $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$ when it is useful to do so), which will give A an expected utility

$$EU_A = Pr(\underline{\theta}) * \left(\rho + \omega_D + \omega_A + \frac{F(\alpha, \underline{\theta})}{4} - \frac{F(\alpha, \underline{\theta})}{2} - \omega_A\right) + Pr(\bar{\theta}) * \left(\rho + \omega_D + \omega_A + \frac{F(\alpha, \underline{\theta})}{4} - \frac{F(\alpha, \bar{\theta})}{2} - \omega_A\right),$$

or

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

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

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

I can then identify A's expected utility from provoking types $\underline{\theta}$ to go to war. Here A 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 $t(\bar{\theta})$ (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 A an expected utility

$$EU_A = Pr(\underline{\theta}) * (\rho - \omega_A) + Pr(\bar{\theta}) * \left(\rho + \omega_D - \frac{F(\alpha, \bar{\theta})}{4}\right)$$

or

$$EU_A = p + Pr(\underline{\theta}) * (-\omega_A) + Pr(\bar{\theta}) * \left(\omega_D - \frac{F(\alpha, \bar{\theta})}{4}\right).$$

I can also express D's expected utility.

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

1.2 Equilibrium

In the interest of saving space, I only include aspects not covered in Proposition 1. Below is D's best response in response to t in Stage 3. A type θ parameter α D will set

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

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

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

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

1.3 A's Behavior within Predictability and Emboldneing Mechanisms

The following figures further elaborate on A's decision making in the Predictability and Emboldening mechanisms. In the figures, the x-axis plots values of t (with t values increasing moving left-to-right), and the y-axis plots A's expected utility from the selected t (with values increasing low-to-high). The asterisks denotes the equilibrium level of rising technology and A's corresponding expected utility. These figures are derived using the parameters described in Figures 1 and 2 in the text.

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), A's utility is increasing in t until it reaches $t(\underline{\alpha}, \underline{\theta})$.[1] A'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 investment in rising technology 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 A goes to war with probability $Pr(\underline{\theta})$. For the set of t's such that $t(\underline{\alpha}, \underline{\theta}) < t \leq t(\underline{\alpha}, \bar{\theta})$, A'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

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

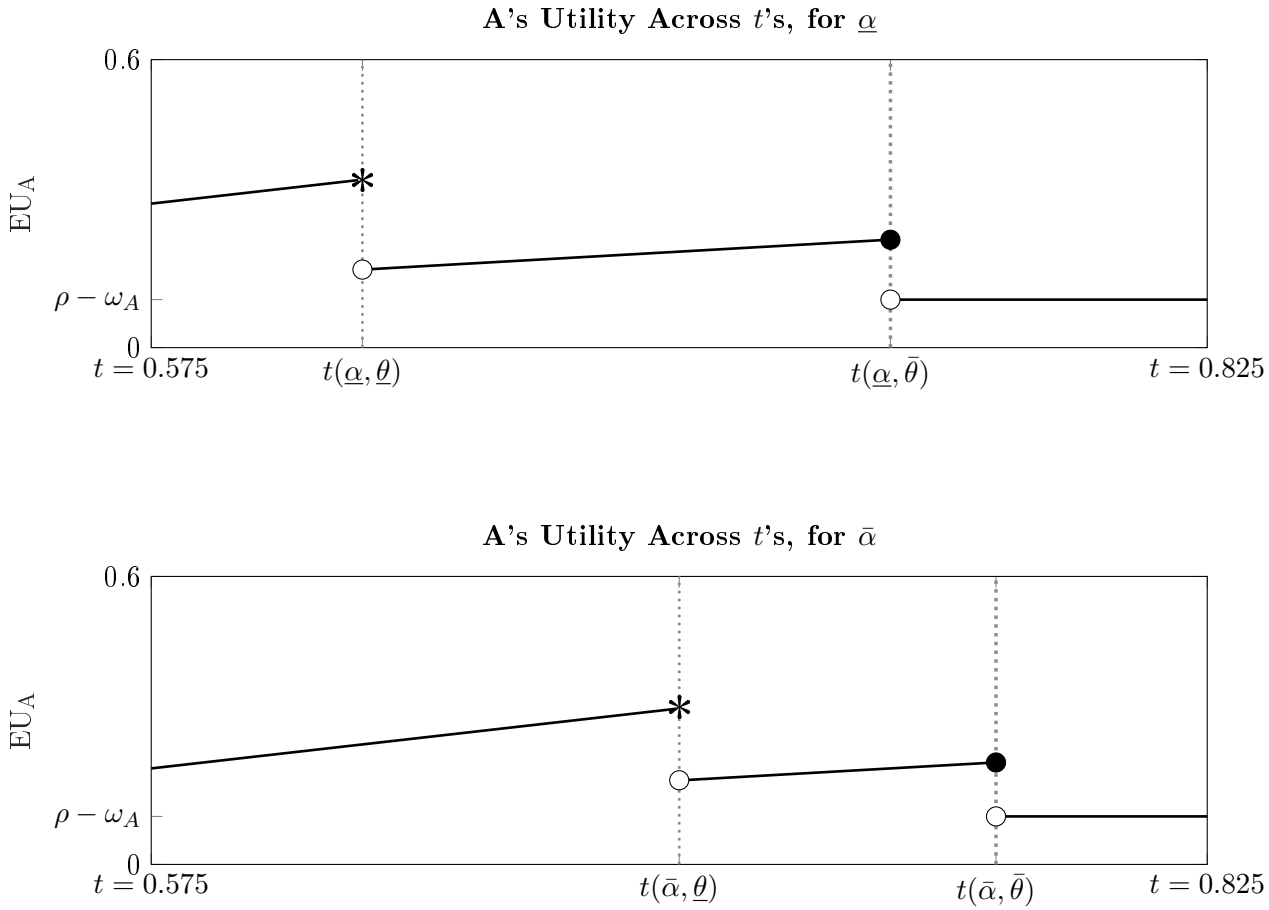

Figure 3. Predictability and State A's Utilities Across Values of t. The parameter values are identical to those used in the Predictability example in the text. A's selected level of investment under parameters $\underline{\alpha}$ and $\bar{\alpha}$ are denoted by the asterisks.

go to war. Thus, if A selects a $t > t(\underline{\alpha}, \bar{\theta})$, then types $\underline{\theta}$ and types $\bar{\theta}$ D's will go to war, which grants A their wartime expected utility. For parameter $\underline{\alpha}$, selecting $t = t(\underline{\alpha}, \underline{\theta})$ gives A the greatest expected utility. Intuitively, in the text of the paper, I discussed the trade-off between increased investment in rising technology and an increased risk of war. Here, A attains the greatest utility from selecting investment level $t = t(\underline{\alpha}, \underline{\theta})$ and never going to war rather than selecting investment level $t = t(\underline{\alpha}, \bar{\theta})$ and sometimes going to war.

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

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 A could select that would only invoke type $\underline{\theta}$ D's to go to war ($t(\bar{\alpha}, \bar{\theta})$). This means that under $\bar{\alpha}$, A has more to gain by risking war with likelihood $Pr(\underline{\theta})$. In the emboldening case for $\underline{\alpha}$, while A's utility is increasing in t for $t(\underline{\alpha}, \underline{\theta}) < t \leq t(\underline{\alpha}, \bar{\theta})$, but the greatest t in this set does not produce a utility for A that sufficiently offsets A's costs from having to go to war with type $\underline{\theta}$ D's. Visually, under $\underline{\alpha}$, A's expected utility is increasing for $t(\underline{\alpha}, \underline{\theta}) < t \leq t(\underline{\alpha}, \bar{\theta})$, but this does not creep up high enough (due to the small difference between $t(\underline{\alpha}, \underline{\theta})$ and $t(\underline{\alpha}, \bar{\theta})$), ultimately resulting in A 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 A can invest much more in rising technology so long that A 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 of rising technology and going to war more was not worthwhile; under parameter $\bar{\alpha}$, A can select a much greater level of t, which offers enough upside to make it worthwhile to go to war with types $\underline{\theta}$.

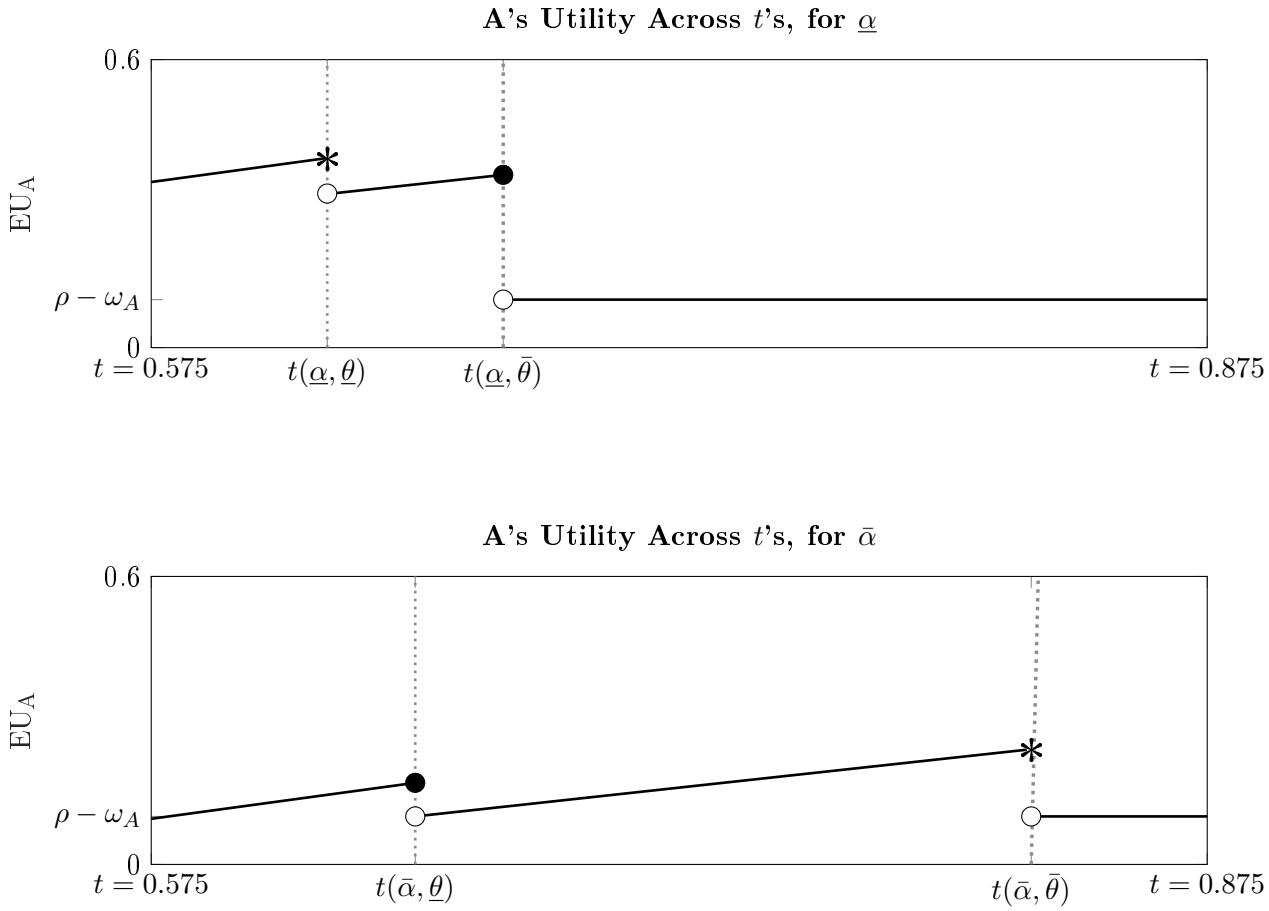

Figure 4. Emboldening and State A's Utilities Across Values of t. The parameter values are identical to those used in the Emboldening example in the text. A's selected level of investment under parameters $\underline{\alpha}$ and $\bar{\alpha}$ are denoted by the asterisks.

1.4 Predictability and Poker

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 a game of Texas hold'em poker, consider the following choice: a player can choose between drawing a pair of queens 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 queens. This is not to say that two aces are a worse hand than the two queens: 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 queens) and private information, they could likely extract more

value from the game.

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, through algebra, is equivalent to

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

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 deterrence 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

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

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 deterrence failure.

4 Proposition 4 Proof

The "only if" was shown in Propositions 2 and 3.

When there is a deterrence 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: A selects either $t(\underline{\theta}, \alpha) = \omega_D + \omega_A + \frac{F(\alpha,\underline{\theta})}{4}$ (type $\underline{\theta}$ is made indifferent between hassling and war) or $t(\bar{\theta}, \alpha) = \omega_D + \omega_A + \frac{F(\alpha,\bar{\theta})}{4}$ (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 $(t(\underline{\theta}, \underline{\alpha}),\ t(\underline{\theta}, \bar{\alpha}))$, $(t(\bar{\theta}, \underline{\alpha}),\ t(\underline{\theta}, \bar{\alpha}))$, $(t(\underline{\theta}, \underline{\alpha}),\ t(\bar{\theta}, \bar{\alpha}))$, or $(t(\bar{\theta}, \underline{\alpha}),\ t(\bar{\theta}, \bar{\alpha}))$.

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

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

these cases into expressions (4) and (5) above, yielding

$$1-\rho-\omega_D \geq 1-\rho-\omega_D$$

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

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

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

When A selects $t(\bar{\theta},\underline{\alpha})$ and $t(\underline{\theta},\bar{\alpha})$, it implies that A 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 deterrence failure), this case is ruled out. As shown in Proposition 1, when A selects $t(\bar{\theta},\underline{\alpha})$ and $t(\underline{\theta},\bar{\alpha})$, capabilities $\underline{\alpha}$, $\bar{\theta}$ D attains utility $1-\rho-C_D$ while capabilities $\bar{\alpha},\bar{\theta}$ D attains utility $1-\rho-\omega_D+\frac{F(\alpha,\bar{\theta})}{4}-\frac{F(\alpha,\underline{\theta})}{4}$, which is strictly larger.

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

When A selects $t(\underline{\theta},\underline{\alpha})$ and $t(\bar{\theta},\bar{\alpha})$, it implies that A 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 A to behave this way, it must be that $Pr(\underline{\theta})\,(\omega_A+\omega_D)+\big(Pr(\bar{\theta})-Pr(\underline{\theta})\big)\,\frac{F(\underline{\alpha},\underline{\theta})}{4}-Pr(\bar{\theta})\frac{F(\underline{\alpha},\bar{\theta})}{4} \geq 0$ and $Pr(\underline{\theta})\,(\omega_A+\omega_D)+\big(Pr(\bar{\theta})-Pr(\underline{\theta})\big)\,\frac{F(\bar{\alpha},\underline{\theta})}{4}-Pr(\bar{\theta})\frac{F(\bar{\alpha},\bar{\theta})}{4}<0$. This meets the two Emboldening Conditions. Note that here a deterrence failure occurs within this case without any further conditions needed; I substitute D's utilities into expressions (4) and (5), yielding

$$1-\rho-\omega_D \geq 1-\rho-\omega_D$$

and

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

where the bottom inequality always holds strictly.

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

When A selects $t(\bar{\theta}, \underline{\alpha})$ and $t(\underline{\theta}, \bar{\alpha})$, it implies that A does weakly sometimes going to war for both $\underline{\alpha}$ and $\bar{\alpha}$. When there is a deterrence failure, this case is ruled out. As shown in Proposition 1, when A selects $t(\bar{\theta}, \underline{\alpha})$ and $t(\bar{\theta}, \bar{\alpha})$, all types and capabilities D attain their wartime utility $1-\rho-C_D$. Because across all hassling capabilities D attains the same utility, there cannot be a deterrence failure.

□

5 Observation 2 Proof

Based on Proposition 4, I can prove Observation 1 by showing that the Predictability and Emboldening Conditions produce a welfare loss. Assume for now the Predictability Conditions hold. When nature selects $\underline{\theta}$, A's and D's utilities are

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

I can compare utilities across $\underline{\alpha}$ and $\bar{\alpha}$. Because $F(\cdot, \underline{\theta})$ is increasing in α, $U_D(\bar{\alpha}, \underline{\theta}) + U_A(\bar{\alpha}, \underline{\theta}) < U_D(\underline{\alpha}, \underline{\theta}) + U_A(\underline{\alpha}, \underline{\theta})$. Thus, welfare is strictly reduced in moving from $\underline{\alpha}$ to $\bar{\alpha}$.

And when nature selects $\bar{\theta}$, A's and D's utilities are

$$\begin{aligned}
U_D(\underline{\alpha}, \bar{\theta}) =& 1 - \rho - \omega_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4}, \\
U_A(\underline{\alpha}, \bar{\theta}) =& \rho + \omega_D + \left(\frac{F(\underline{\alpha}, \underline{\theta})}{4} - \frac{F(\underline{\alpha}, \bar{\theta})}{2} \right), \\
U_D(\bar{\alpha}, \bar{\theta}) =& 1 - \rho - \omega_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4}, \\
U_A(\bar{\alpha}, \bar{\theta}) =& \rho + \omega_D + \left(\frac{F(\bar{\alpha}, \underline{\theta})}{4} - \frac{F(\bar{\alpha}, \bar{\theta})}{2} \right).
\end{aligned}$$

Through algebra, the expression $U_D(\bar{\alpha},\bar{\theta})+U_A(\bar{\alpha},\bar{\theta}) \lessgtr U_D(\underline{\alpha},\bar{\theta})+U_A(\underline{\alpha},\bar{\theta})$ can be simplified to $-\frac{F(\bar{\alpha},\bar{\theta})}{4} \lessgtr -\frac{F(\underline{\alpha},\bar{\theta})}{4}$. Because $F(\alpha,\cdot)$ is strictly increasing in α, welfare is strictly reduced in moving from $\underline{\alpha}$ to $\bar{\alpha}$. Thus, regardless of the realized θ, when the Predictability Conditions holds, following an improvement in α, there is a welfare loss.

Now assume the Emboldening Conditions holds. When nature selects $\underline{\theta}$, A's and D's utilities are

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

Through algebra, the expression $U_D(\bar{\alpha},\underline{\theta})+U_A(\bar{\alpha},\underline{\theta}) \lessgtr U_D(\underline{\alpha},\underline{\theta})+U_A(\underline{\alpha},\underline{\theta})$ can be simplified to $-\omega_D-\omega_A \lessgtr -\frac{F(\underline{\alpha},\underline{\theta})}{4}$. By the assumption that $P(t^*,h^*)>\rho$,[2] welfare is strictly reduced in moving from $\underline{\alpha}$ to $\bar{\alpha}$.

And when nature selects $\bar{\theta}$, A's and D's utilities are

$$\begin{aligned}
U_D(\underline{\alpha},\bar{\theta}) =& 1-\rho-\omega_D+\frac{F(\underline{\alpha},\bar{\theta})}{4}-\frac{F(\underline{\alpha},\underline{\theta})}{4},\\
U_A(\underline{\alpha},\bar{\theta}) =& \rho+\omega_D+\left(\frac{F(\underline{\alpha},\underline{\theta})}{4}-\frac{F(\underline{\alpha},\bar{\theta})}{2}\right),\\
U_D(\bar{\alpha},\bar{\theta}) =& 1-\rho-\omega_D,\\
U_A(\bar{\alpha},\bar{\theta}) =& \rho+\omega_D-\frac{F(\bar{\alpha},\bar{\theta})}{4}.
\end{aligned}$$

Through algebra, the expression $U_D(\bar{\alpha},\underline{\theta})+U_A(\bar{\alpha},\underline{\theta}) \lessgtr U_D(\underline{\alpha},\underline{\theta})+U_A(\underline{\alpha},\underline{\theta})$ can be simplified to $-\frac{F(\bar{\alpha},\bar{\theta})}{4} \lessgtr -\frac{F(\underline{\alpha},\bar{\theta})}{4}$. Because $F(\alpha,\cdot)$ is strictly increasing in α, welfare is strictly reduced in moving from $\underline{\alpha}$ to $\bar{\alpha}$. Thus, regardless of the realized θ, when the Emboldening Conditions holds, following an improvement in α, there is a welfare loss.

[2] This assumption amounts to $t^*>h^*$. When D is parameter $\underline{\alpha}$ and type $\underline{\theta}$, then $t^*=\omega_D+\omega_A+\frac{F(\underline{\alpha},\underline{\theta})}{4}$ and $h^*=\frac{F(\underline{\alpha},\underline{\theta})}{2}$.

5.1 When a Welfare Loss Doesn't Occur

Example of improvements in hassling capabilities not producing a welfare loss, consider when $Pr(\underline{\theta})(\omega_A + \omega_D) + \left(Pr(\bar{\theta}) - Pr(\underline{\theta})\right)\frac{F(\underline{\alpha},\underline{\theta})}{4} - Pr(\bar{\theta})\frac{F(\underline{\alpha},\bar{\theta})}{4} < 0$ and $Q = Pr(\underline{\theta})(\omega_A + \omega_D) + \left(Pr(\bar{\theta}) - Pr(\underline{\theta})\right)\frac{F(\bar{\alpha},\underline{\theta})}{4} - Pr(\bar{\theta})\frac{F(\bar{\alpha},\bar{\theta})}{4} > 0$. Under these conditions, A sometimes goes to war under parameter $\underline{\alpha}$ (Case 2 in Proposition 1), and always avoids war under parameter $\bar{\alpha}$ (Case 1 in Proposition 1). When nature selects $\underline{\theta}$, A's and D's utilities are

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

Through algebra, the expression $U_D(\bar{\alpha},\bar{\theta}) + U_A(\bar{\alpha},\bar{\theta}) \lessgtr U_D(\underline{\alpha},\bar{\theta}) + U_A(\underline{\alpha},\bar{\theta})$ can be simplified to $-\frac{F(\bar{\alpha},\underline{\theta})}{4} \lessgtr -\omega_D - \omega_A$. By the assumption that $P(t^*,h^*) > \rho$,[3] welfare is strictly increasing in moving from $\underline{\alpha}$ to $\bar{\alpha}$.

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 deterrence failure to arise is through the emboldening or predictability mechanisms. I generalize these mechanisms to accommodate the broader

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 A becomes emboldened, there is more war, and A 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.

6.1 Model Assumptions

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

[3] This assumption amounts to $t^* > h^*$. When D is parameter $\bar{\alpha}$ and type $\underline{\theta}$, then $t^* = \omega_D + \omega_A + \frac{F(\bar{\alpha},\underline{\theta})}{4}$ and $h^* = \frac{F(\bar{\alpha},\underline{\theta})}{2}$.

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 A does not want to invest in the rising technology, 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 A plays the game, note that D selects h^*, w_D^*, and x^* identical to how D played in the model where A faced no costs.

In the model in the text, A restricted their selected t out of fear of more war with D. This led to A choosing between $t(\underline{\theta})$ and $t(\bar{\theta})$, the levels of rising technology 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 A may be constrained by their costs of war. For now assume that A's optimal selection of t is weakly less than $t(\underline{\theta})$ – in other words, A wants to avoid war and the cost to the rising technology are binding (I discuss other cases next). Then A is selecting a t^* defined by

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

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

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

The solution to this optimization is also $t^* = \frac{1}{2\kappa}$. Thus, I define a new constant: $\hat{t} = \frac{1}{2\kappa}$ which is, when A's cost constraints bind, how far A is willing to go in their selection of t. Here A will choose between the values $\hat{t}$, $t(\underline{\theta})$, and $t(\bar{\theta})$. To better characterize A's choice, I define A'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 $\hat{t} > t(\underline{\theta})$). For a fixed α,

these are

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

$$U_A(t(\underline{\theta})) = -\kappa\left(\omega_D + \omega_A + \frac{F(\alpha,\underline{\theta})}{4}\right)^2 + Pr(\underline{\theta}) * \left(\rho + \omega_D - \frac{F(\alpha,\underline{\theta})}{4}\right) + Pr(\bar{\theta}) * \left(\rho + \omega_D + \frac{F(\alpha,\underline{\theta})}{4} - \frac{F(\alpha,\bar{\theta})}{2}\right),$$

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

$$U_A(t(\bar{\theta})) = Pr(\underline{\theta})(\rho - \omega_A) + Pr(\bar{\theta})\left(\rho + \omega_D - \frac{F(\alpha,\bar{\theta})}{4}\right) - \kappa\left(\omega_D + \omega_A + \frac{F(\alpha,\bar{\theta})}{4}\right)^2.$$

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

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 $t(\underline{\theta})$ and $t(\bar{\theta})$, and (b) some comparison of A's expected utility across some subset $t \in \{\hat{t}, t(\underline{\theta}), t(\bar{\theta})\}$. Why does (a) matter? For example, when $t(\bar{\theta})$ is greater than $\hat{t}$, then A would never play $t(\bar{\theta})$ (with the same logic for $t(\underline{\theta}) > \hat{t}$ and A never playing $t(\underline{\theta})$). Why does (b) matter? Once (a) narrows the possible range of actions, then (b) compares A'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 $\hat{t} \leq t(\underline{\theta})$ holds, A avoids war:

- *A's beliefs on θ follow its expected distribution. A selects technology level*

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

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

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

A has expected utility

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

Case 2. When $\hat{t}\in\left(t(\underline{\theta}),t(\bar{\theta})\right)$ and $U_A(\hat{t},war)\leq U_A(t(\underline{\theta}))$, A avoids war:

- *A's beliefs on θ follows its expected distribution. A selects technology level*

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

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

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

- *A has expected utility*

$$EU_A = \rho+\omega_D-\kappa\left(\omega_D+\omega_A+\frac{F(\alpha,\underline{\theta})}{4}\right)^2-Pr(\underline{\theta})\left(\frac{F(\alpha,\underline{\theta})}{4}\right)+Pr(\bar{\theta})\left(\frac{F(\alpha,\underline{\theta})}{4}-\frac{F(\alpha,\bar{\theta})}{2}\right).$$

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

- *A's beliefs on θ follows its expected distribution. A selects technology level*

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

- *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

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

A has expected utility

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

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

- *A's beliefs on θ follows its expected distribution. A selects technology level*

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

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

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

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

A has expected utility

$$EU_A = Pr(\underline{\theta})(\rho-\omega_A)+Pr(\bar{\theta})\left(\rho+\omega_D-\frac{F(\alpha,\bar{\theta})}{4}\right)-\kappa\left(\omega_D+\omega_A+\frac{F(\alpha,\bar{\theta})}{4}\right)^2.$$

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

- *A's beliefs on θ follows its expected distribution. A selects technology level*

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

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

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

- *A has expected utility*

$$EU_A = \rho + \omega_D - \kappa \left(\omega_D + \omega_A + \frac{F(\alpha, \underline{\theta})}{4} \right)^2 - Pr(\underline{\theta}) \left(\frac{F(\alpha, \underline{\theta})}{4} \right) + Pr(\bar{\theta}) \left(\frac{F(\alpha, \underline{\theta})}{4} - \frac{F(\alpha, \bar{\theta})}{2} \right).$$

Proof: Follows from discussion above.

The above completely characterizes all possible equilibria.

6.3 Results

Recall from earlier the features of A 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 A becomes emboldened, there is more war, and A 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 to denote that the conditions relate to the model with costs. Also, I now refer to $t(\theta, \alpha)$ and $U_A(t, \alpha)$ (for $\alpha \in \{\underline{\alpha}, \bar{\alpha}\}$) when it is necessary to clarify.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$-\frac{1}{2\kappa} + \omega_A + \omega_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} > \frac{F(\bar{\alpha}, \bar{\theta})}{4} - \frac{F(\bar{\alpha}, \underline{\theta})}{4}.$$

And I can also define conditions for emboldening.

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

Definition: *under the* ***Eκ1*** *conditions, the following holds:*[5]

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

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

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

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

$$\frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > -\frac{1}{2\kappa} + \omega_A + \omega_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4}.$$

Definition: *under the* ***Eκ2*** *conditions, the following holds:*

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

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

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

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

$$\frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > -\frac{1}{2\kappa} + \omega_A + \omega_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4}.$$

[5] The last condition is equivalent to $1 - \rho - \omega_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > 1 - \rho - \frac{1}{2\kappa} + \omega_A + \frac{F(\bar{\alpha}, \bar{\theta})}{4}$.

Definition: *under the* ***Eκ3*** *conditions, the following holds:*

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

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

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

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

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

Proposition 6: *Improvements in α produce a deterrence failure if and only if* ***Pκ1, Pκ2, Pκ3, Pκ4, Eκ1, Eκ2,*** *or* ***Eκ3*** *hold.*

6.4 Proving Proposition 6

6.4.1 Proving $\leftarrow$

On Pκ1:

Through the top four **Pκ1** 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

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

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 deterrence failure.

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

On Pκ2:

Through the top four **Pκ2** 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

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

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 deterrence failure.

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

On Pκ3:

Through the top four **Pκ3** 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

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

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 deterrence failure.

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

Through the top four **Pκ4** 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

$$-\frac{1}{2\kappa}+\omega_A+\omega_D+\frac{F(\underline{\alpha},\bar{\theta})}{4}>\frac{F(\bar{\alpha},\bar{\theta})}{4}-\frac{F(\bar{\alpha},\underline{\theta})}{4},$$

which through algebra is equivalent to

$$1-\rho-\frac{1}{2\kappa}+\omega_A+\frac{F(\alpha,\bar{\theta})}{4}>1-\rho-\omega_D+\frac{F(\bar{\alpha},\bar{\theta})}{4}-\frac{F(\bar{\alpha},\underline{\theta})}{4},$$

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 deterrence failure.

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

On Eκ1:

Through the top four **Eκ1** 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

$$\frac{F(\underline{\alpha},\bar{\theta})}{4}-\frac{F(\underline{\alpha},\underline{\theta})}{4}>-\frac{1}{2\kappa}+\omega_A+\omega_D+\frac{F(\bar{\alpha},\bar{\theta})}{4},$$

which through algebra is equivalent to

$$1-\rho-\omega_D+\frac{F(\underline{\alpha},\bar{\theta})}{4}-\frac{F(\underline{\alpha},\underline{\theta})}{4}>1-\rho-\frac{1}{2\kappa}+\omega_A+\frac{F(\bar{\alpha},\bar{\theta})}{4},$$

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 deterrence failure.

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

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

Through the top four $\mathbf{E}\kappa\mathbf{2}$ 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

$$\frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > -\frac{1}{2\kappa} + \omega_A + \omega_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4},$$

which through algebra is equivalent to

$$1 - \rho - \omega_D + \frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > 1 - \rho - \frac{1}{2\kappa} + \omega_A + \frac{F(\bar{\alpha}, \bar{\theta})}{4},$$

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 deterrence failure.

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

On Eκ3:

Through the top four **Eκ3** 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

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

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 deterrence failure.

An example of parameters that produce this condition are as follows: $F(\underline{\alpha}, \underline{\theta}) = 0.1$, $F(\underline{\alpha}, \bar{\theta}) = 0.2$, $F(\bar{\alpha}, \underline{\theta}) = 0.15$, $F(\bar{\alpha}, \bar{\theta}) = 0.4$, $\omega_D = 0.25$, $\omega_A = 0.05$, $\rho = 0.2$, $Pr(\underline{\theta}) = 0.05$, $Pr(\bar{\theta}) = 0.95$, $\kappa = 1.15$.

6.4.2 Proving $\rightarrow$

When there is a deterrence 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 deterrence failure, then it implies that one of the conditions outlined earlier holds, or (b) there cannot be a deterrence failure.

Case 1-Case 1:

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

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

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, $\hat{t} \in \big(t(\underline{\theta},\underline{\alpha}), t(\bar{\theta},\underline{\alpha})\big)$ and $\hat{t} \leq t(\underline{\theta},\bar{\alpha})$. A will select $t(\underline{\theta},\underline{\alpha})$ 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 deterrence failure), this case is ruled out. Because $\hat{t} < t(\underline{\theta},\bar{\alpha})$ implies $\frac{1}{2\kappa} < \omega_D + \omega_A + \frac{F(\bar{\alpha},\underline{\theta})}{4}$, the following statement must hold:

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

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, $\hat{t} \in \big(t(\underline{\theta},\underline{\alpha}), t(\bar{\theta},\underline{\alpha})\big)$ and $\hat{t} \in \big(t(\underline{\theta},\bar{\alpha}), t(\bar{\theta},\bar{\alpha})\big)$, which satisfies the first two conditions of Pκ1. A always selects $t^* = t(\underline{\theta},\alpha)$, and D always hassles. The conditions of the cases imply that $U_A(t(\underline{\theta},\underline{\alpha}),\underline{\alpha}) \geq U_A(\hat{t}, war, \underline{\alpha})$ and $U_A(t(\underline{\theta},\bar{\alpha}),\bar{\alpha}) \geq U_A(\hat{t}, war, \bar{\alpha})$, which satisfies the third and forth conditions of Pκ1. Furthermore, because there is a deterrence failure, I can substitute D's utilities from these cases into equations (6) and (7) above, yielding

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

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 Pκ1. Thus, when there is a deterrence failure in Case 2-Case 2, the conditions defining $P\kappa 1$ must hold.

Case 2-Case 3:

By the conditions of the cases, $\hat{t} \in \big(t(\underline{\theta},\underline{\alpha}), t(\bar{\theta},\underline{\alpha})\big)$ and $\hat{t} \in \big(t(\underline{\theta},\bar{\alpha}), t(\bar{\theta},\bar{\alpha})\big)$, which satisfies the first two conditions of Eκ1. A will select $t(\underline{\theta},\underline{\alpha})$ 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 $U_A(t(\underline{\theta},\underline{\alpha}),\underline{\alpha}) \geq U_A(\hat{t}, war, \underline{\alpha})$ and $U_A(t(\underline{\theta},\bar{\alpha}),\bar{\alpha}) < U_A(\hat{t}, war, \bar{\alpha})$, which satisfies the third and forth conditions of Eκ1. Furthermore, because there is a deterrence failure, I can substitute D's utilities from these cases into equations (6) and (7) above,

yielding

$$1-\rho-\omega_D \geq 1-\rho-\omega_D$$

and

$$1-\rho-\omega_D+\frac{F(\underline{\alpha},\bar{\theta})}{4}-\frac{F(\underline{\alpha},\underline{\theta})}{4} \geq 1-\rho-\frac{1}{2\kappa}+\omega_A+\frac{F(\bar{\alpha},\bar{\theta})}{4}.$$

Because the top holds with equality, the bottom inequality must be strict. Re-writing the bottom inequality yields $\frac{F(\underline{\alpha},\bar{\theta})}{4}-\frac{F(\underline{\alpha},\underline{\theta})}{4} > -\frac{1}{2\kappa}+\omega_A+\omega_D+\frac{F(\bar{\alpha},\bar{\theta})}{4}$, which is the final condition in Eκ1. Thus, when there is a deterrence failure in Case 2-Case 3, the conditions defining Eκ1 must hold.

Case 3-Case 1:

By the conditions of the cases, $\hat{t} \in \big(t(\underline{\theta},\underline{\alpha}), t(\bar{\theta},\underline{\alpha})\big)$ and $\hat{t} \leq t(\underline{\theta},\bar{\alpha})$. A 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 deterrence failure), this case is ruled out. Because $\hat{t} < t(\underline{\theta},\bar{\alpha})$ implies $\frac{1}{2\kappa} < \omega_D+\omega_A+\frac{F(\bar{\alpha},\underline{\theta})}{4}$, the following statement must hold:

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

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, $\hat{t} \in \big(t(\underline{\theta},\underline{\alpha}), t(\bar{\theta},\underline{\alpha})\big)$ and $\hat{t} \in \big(t(\underline{\theta},\bar{\alpha}), t(\bar{\theta},\bar{\alpha})\big)$, which satisfies the first two conditions of Pκ4. A will select $\hat{t}$ when $\alpha=\underline{\alpha}$ and $t(\underline{\theta},\bar{\alpha})$ 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 $U_A(\hat{t},war,\underline{\alpha}) \geq U_A(t(\underline{\theta},\underline{\alpha}),\underline{\alpha})$ and $U_A(t(\underline{\theta},\bar{\alpha}),\bar{\alpha}) \geq U_A(\hat{t},war,\bar{\alpha})$, which satisfies the third and forth conditions of Pκ4. Furthermore, because there is a deterrence failure, I can substitute D's utilities from these cases into equations (6) and (7) above, yielding

$$1-\rho-\omega_D \geq 1-\rho-\omega_D$$

and

$$1-\rho-\frac{1}{2\kappa}+\omega_A+\frac{F(\underline{\alpha},\bar{\theta})}{4}\geq 1-\rho-\omega_D+\frac{F(\bar{\alpha},\bar{\theta})}{4}-\frac{F(\bar{\alpha},\underline{\theta})}{4}$$

Because the top holds with equality, the bottom inequality must be strict. Re-writing the bottom inequality yields $-\frac{1}{2\kappa}+\omega_A+\omega_D+\frac{F(\underline{\alpha},\bar{\theta})}{4}>\frac{F(\bar{\alpha},\bar{\theta})}{4}-\frac{F(\bar{\alpha},\underline{\theta})}{4}$, which is the final condition in Pκ4. Thus, when there is a deterrence failure in Case 3-Case 2, the conditions defining Pκ4 must hold.

Case 3-Case 3:

By the conditions of the cases, $\hat{t}\in\big(t(\underline{\theta},\underline{\alpha}),t(\bar{\theta},\underline{\alpha})\big)$ and $\hat{t}\in\big(t(\underline{\theta},\bar{\alpha}),t(\bar{\theta},\bar{\alpha})\big)$. A always selects $t=\frac{1}{2\kappa}$, 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 deterrence failure), this case is ruled out. Because F is increasing in α, the following statement must hold:

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

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, $\hat{t}\geq t(\bar{\theta},\underline{\alpha})$ and $\hat{t}\leq t(\underline{\theta},\bar{\alpha})$. A will select $t(\bar{\theta},\underline{\alpha})$ 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 deterrence failure), this case is ruled out. Because $\hat{t}<t(\underline{\theta},\bar{\alpha})$ implies $\frac{1}{2\kappa}<\omega_D+\omega_A+\frac{F(\bar{\alpha},\underline{\theta})}{4}$, the following statement must hold:

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

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, $\hat{t}\geq t(\bar{\theta},\underline{\alpha})$ and $\hat{t}\in\big(t(\underline{\theta},\bar{\alpha}),t(\bar{\theta},\bar{\alpha})\big)$. A will select $t(\bar{\theta},\underline{\alpha})$

when $\alpha = \underline{\alpha}$ and $t(\underline{\theta}, \bar{\alpha})$ 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 deterrence failure), this case is ruled out. Because F is increasing in θ, the following statement must hold:

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

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, $\hat{t} \geq t(\bar{\theta}, \underline{\alpha})$ and $\hat{t} \in \big(t(\underline{\theta}, \bar{\alpha}), t(\bar{\theta}, \bar{\alpha})\big)$. A will select $t(\bar{\theta}, \underline{\alpha})$ 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 deterrence failure), this case is ruled out. Because $\hat{t} < t(\bar{\theta}, \bar{\alpha})$ implies $\frac{1}{2\kappa} < \omega_D + \omega_A + \frac{F(\bar{\alpha}, \bar{\theta})}{4}$, the following statement must hold:

$$1 - \rho - \omega_D < 1 - \rho - \frac{1}{2\kappa} + \omega_A + \frac{F(\bar{\alpha}, \bar{\theta})}{4}.$$

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, $\hat{t} \geq t(\bar{\theta}, \underline{\alpha})$ and $\hat{t} \geq t(\bar{\theta}, \bar{\alpha})$. A will select $t(\bar{\theta}, \underline{\alpha})$ when $\alpha = \underline{\alpha}$ and $t(\bar{\theta}, \bar{\alpha})$ 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 deterrence 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, $\hat{t} \geq t(\bar{\theta}, \underline{\alpha})$ and $\hat{t} \geq t(\bar{\theta}, \bar{\alpha})$. A will select $t(\bar{\theta}, \underline{\alpha})$ when $\alpha = \underline{\alpha}$ and $t(\underline{\theta}, \bar{\alpha})$ 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 deterrence failure), this

case is ruled out. Because F is increasing in θ, the following statement must hold:

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

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, $\hat{t} \geq t(\bar{\theta},\underline{\alpha})$ and $\hat{t} \leq t(\underline{\theta},\bar{\alpha})$. A will select $t(\underline{\theta},\underline{\alpha})$ 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 deterrence failure), this case is ruled out. Because $\hat{t} < t(\underline{\theta},\bar{\alpha})$ implies $\frac{1}{2\kappa} < \omega_D+\omega_A+\frac{F(\bar{\alpha},\underline{\theta})}{4}$, the following statement must hold:

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

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, $\hat{t} \geq t(\bar{\theta},\underline{\alpha})$ and $\hat{t} \in \big(t(\underline{\theta},\bar{\alpha}), t(\bar{\theta},\bar{\alpha})\big)$ which satisfies the first two conditions of Pκ2. A will select $t(\underline{\theta},\underline{\alpha})$ when $\alpha=\underline{\alpha}$ and $t(\underline{\theta},\bar{\alpha})$ 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 $U_A(t(\underline{\theta},\underline{\alpha}),\underline{\alpha}) \geq U_A(t(\bar{\theta},\underline{\alpha}),\underline{\alpha})$ and $U_A(t(\underline{\theta},\bar{\alpha}),\bar{\alpha}) \geq U_A(\hat{t},war,\bar{\alpha})$, which satisfies the third and forth conditions of Pκ2. Furthermore, because there is a deterrence failure, I can substitute D's utilities from these cases into equations (6) and (7) above, yielding

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

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 Pκ2. Thus, when there is a deterrence failure in Case 5-Case 2, the conditions defining

$P\kappa 2$ must hold.

Case 5-Case 3:[6]

By the conditions of the cases, $\hat{t} \geq t(\bar{\theta}, \underline{\alpha})$ and $\hat{t} \in \big(t(\underline{\theta}, \bar{\alpha}), t(\bar{\theta}, \bar{\alpha})\big)$ which satisfies the first two conditions of Eκ2. A will select $t(\underline{\theta}, \underline{\alpha})$ 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 $U_A(t(\underline{\theta}, \underline{\alpha}), \underline{\alpha}) \geq U_A(t(\bar{\theta}, \underline{\alpha}), \underline{\alpha})$ and $U_A(t(\underline{\theta}, \bar{\alpha}), \bar{\alpha}) < U_A(\hat{t}, war, \bar{\alpha})$, which satisfies the third and forth conditions of Eκ2. Furthermore, because there is a deterrence failure, I can substitute D's utilities from these cases into equations (6) and (7) above, yielding

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

Because the top holds with equality, the bottom inequality must be strict. Re-writing the bottom inequality yields $\frac{F(\underline{\alpha}, \bar{\theta})}{4} - \frac{F(\underline{\alpha}, \underline{\theta})}{4} > -\frac{1}{2\kappa} + \omega_A + \omega_D + \frac{F(\bar{\alpha}, \bar{\theta})}{4}$, which is the final condition in Eκ2. Thus, when there is a deterrence failure in Case 5-Case 3, the conditions defining Eκ2 must hold.

Case 5-Case 4:

By the conditions of the cases, $\hat{t} \geq t(\bar{\theta}, \underline{\alpha})$ and $\hat{t} \geq t(\bar{\theta}, \bar{\alpha})$ which satisfies the first two conditions of Eκ3. A will select $t(\underline{\theta}, \underline{\alpha})$ when $\alpha = \underline{\alpha}$ and $t(\bar{\theta}, \bar{\alpha})$ 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 $U_A(t(\underline{\theta}, \underline{\alpha}), \underline{\alpha}) \geq U_A(t(\bar{\theta}, \underline{\alpha}), \underline{\alpha})$ and $U_A(t(\underline{\theta}, \bar{\alpha}), \bar{\alpha}) < U_A(t(\bar{\theta}, \bar{\alpha}), \bar{\alpha})$, which satisfies the third and forth conditions of Eκ3. Note that here a deterrence failure occurs within this case without any further conditions needed; I substitute D's utilities into expressions (6) and (7), yielding

$$1 - \rho - \omega_D \geq 1 - \rho - \omega_D$$

and

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

[6]Note: a deterrence failure may not be possible here – see above.

where the bottom inequality always holds strictly.

Case 5-Case 5:

By the conditions of the cases, $\hat{t} \geq t(\bar{\theta}, \underline{\alpha})$ and $\hat{t} \geq t(\bar{\theta}, \bar{\alpha})$ which satisfies the first two conditions of Pκ3. A will select $t(\underline{\theta}, \underline{\alpha})$ when $\alpha = \underline{\alpha}$ and $t(\underline{\theta}, \bar{\alpha})$ 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 $U_A(t(\underline{\theta}, \underline{\alpha}), \underline{\alpha}) \geq U_A(t(\bar{\theta}, \underline{\alpha}), \underline{\alpha})$ and $U_A(t(\underline{\theta}, \bar{\alpha}), \bar{\alpha}) \geq U_A(t(\bar{\theta}, \bar{\alpha}), \bar{\alpha})$, which satisfies the third and forth conditions of Pκ3. Furthermore, because there is a deterrence failure, I can substitute D's utilities from these cases into equations (6) and (7) above, yielding

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

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 Pκ3. Thus, when there is a deterrence failure in Case 5-Case 5, the conditions defining $P\kappa 3$ 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 deterrence failure can arise from improved public hassling capabilities, but cannot arise from improved private hassling capabilities.

7 On the Definition of Deterrence Failure

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

***Definition**: Improvements in publicly observed hassling capabilities produce a **deterrence 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 deterrence 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 alter-

native definition, I could have defined a deterrence 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 D attain a better outcome following a deterrence 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 deterrence failure under the weaker definition, but is never classified as a deterrence 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 improving hassling capabilities would create a deterrence failure. Thus, to exclude cases where D would actually want a deterrence failure to occur, I adopt the stricter definition of deterrence failure for my analysis.[7]

Part II

Analysis for Models with a Convex Set of Hassling Levels

8 Example Model with θ as a Continuum

This model is slightly different from the models in the text as here D also has a continuum 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

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

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 A 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 A selects arming technology level $t \geq \underline{t} > 0$.[8]

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 $\underline{h} > 0$, 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 A and D (respectively). Choosing some level of hassling $h \in \mathcal{H}$ (i.e. accepting or hassling) yields payoffs $X_A + \frac{t}{h}$[9] and $X_D - \frac{t}{h} - C_D(h, \theta, \alpha)$ for states A and D (respectively).

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

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

and when war does occur take values

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

To summarize the expected utilities, the expressions $X_A + \frac{t}{h}$ and $X_D - \frac{t}{h} - C_D(h, \theta, \alpha)$ are the values that A and D receive by not going to war and letting the investments in rising technologies and hassling come to fruition. $C_D(h, \theta, \alpha)$ are the costs that D faces from hassling. To capture that hassling is costly, I assume C_D is increasing and strictly convex in h.

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

[9] Results for a model where A invests in costs is available upon request.

To capture that greater values of parameters θ and α imply lower costs of hassling, I assume C_D is differentiable and weakly decreasing in α and θ for all h.

Putting aside the costs from hassling and investing, 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 investing in the rising technology ($t = 0$), states would reach some expected political settlement X_A and X_D, and where the selected levels of rising technology and hassling shift this settlement. I assume that if A does not invest in the rising technology (setting $t = 0$), both states prefer not going to war, or $X_A > W_A$ and $X_D - C_D(\underline{h}, \theta, \alpha) > W_D$ for all θ and α; this implies that war is not inevitable, but rather, if A selects too great an investment in rising technology t, D may find it optimal to go to war with A before the investment in rising technology 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.[10] Ultimately, adopting this framework allows me to focus on the most relevant parts of the game: if rising technology is allowed to come to fruition, it will strengthen A, D can degrade this rising technology 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 A does not know the value of θ, A'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 Deterrence 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

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

space, or $\sigma_D : (\mathcal{A}, \Theta) \rightarrow \{war, \, not \, war\} \times \mathcal{H}$. Because State A does not know the value of θ, A'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 equilibrium actions. For ease, I limit myself to pure strategy equilibria.

At this point, it is possible to formalize a "deterrence failure" for D. I claim that improvements in public hassling capabilities result in a deterrence 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 A 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 deterrence failure can arise here).

Definition: Improvements in hassling capabilities result in a **deterrence 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 A never risks war, and one where A sometimes risks war. Note that from the limitations on $\underline{t}$, A will never try to get A to "accept." What determines this is A's payoffs, which are fairly complicated here. To express the equilibrium simply, I define $U_A(\hat{t}) = X_A + \frac{2\left(\frac{(X_D - W_D)^2(\alpha + \underline{\theta})}{4}\right)^{1/2}}{\bar{\theta} - \underline{\theta}} \left((\alpha + \bar{\theta})^{1/2} - (\alpha + \underline{\theta})^{1/2}\right)$ 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
$U_A(\tilde{t}) = \left(X_A\bar{\theta} - W_A\underline{\theta} + (W_A - X_A)\left(4\tilde{t}\left(X_D - W_D\right)^{-2} - \alpha\right) + 2(\tilde{t})^{1/2}\left(\alpha + \bar{\theta}\right)^{1/2} - 2\tilde{t}\left(X_D - W_D\right)^{-1}\right)\frac{1}{\bar{\theta} - \underline{\theta}}$.

Proposition 7:

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

- *A selects the largest technology level that does not induce war, or*

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

- *Fix a selected level of rising technology 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*

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

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

- *A selects the largest technology level that does not induce war, or*

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

- *Fix a selected level of rising technology 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*

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

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 deterrence failure following a shift from* α *to* α'*.*

The Observation above can be derived by taking first order conditions on U_D.[11] In this equilibrium, A is selecting a level of rising technology 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 A will select a greater level of rising technology. 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 A selects a t that makes a type $\underline{\theta}$ indifferent between war and hassling (as it is optimal for A to do under the conditions above), this selected t is closer to the technology investment that would make a type $\theta \in (\underline{\theta}, \bar{\theta}]$ indifferent between war and a level of hassling, thus granting A 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 deterrence 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

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

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

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

8.4.1 Outcome 1: D always hassles

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

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

Note that this is strictly increasing in t; therefore, within this range, A 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 A's utility here as

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

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 A 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 investment level, 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, A 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

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

or

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

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

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

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 A's utility from this case as

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

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

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

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

Taking first order conditions yields

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

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 continuum of degrees of hassling, 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 A wants to invest in the rising technology to improve their bargaining position, but, because θ is private, A does not know how much they can invest before State D starts hassling to degrade A's investment, or before D goes to war to prevent A's investment 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 A 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 A selects some level of investment in rising technology $t \in \mathcal{T} \subset \mathbb{R}$. 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 levels of rising technology and hassling are allowed to be realized, and then states enter some crisis bargaining that is influenced by the hassling and rising technology levels. For ease, I decompose outcome function g. When states go to war, war occurs before the hassling or rising technology come to fruition, and states receive their wartime payoffs W_A and W_D.[12] When states do not go to war, they receive payoffs $V_A^g(h, t, a_A, a_D) - C_A^g(h, t)$ and $V_D^g(h, t, a_A, a_D) - C_D^g(h, \alpha, \theta)$. The V^g functions define, for a specific outcome function g, the value each state attains through future bargaining that is influenced by the rising technology and hassling. To capture the dynamics of t and h described earlier, V_A^g is increasing in t and decreasing in h, while V_D^g is increasing in h and decreasing in t (in all cases weakly).[13] The C^g functions denote the costs each state attains from hassling and the rising technology. State A incurs costs from both hassling and investing in the rising technology, making C_A weakly increasing in h and t. State D incurs costs from hassling, making C_D weakly increasing in h, and C_D is strictly decreasing in the parameter and type (α and θ, respectively).[14] 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

$$u_A^g(a_A, t, a_D, h) = (1 - \pi^g(a_A, a_D))W_A + \pi^g(a_A, a_D)\left(V_A^g(h, t, a_A, a_D) - C_A^g(t, h)\right) \quad (8)$$

[12] 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).

[13] One way to interpret the V function is that, after the rising technology and hassling comes to fruition, enter a new round of bargaining, where t and h affect A's and D's future bargaining positions. This is discussed in Example 2 below.

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

and

$$u_D^g(a_D, h, a_A, t \,|\, \alpha, \theta) = (1 - \pi^g(a_A, a_D))W_D + \pi^g(a_A, a_D)\left(V_D^g(h, t, a_A, a_D) - C_D^g(h, \alpha, \theta)\right). \tag{9}$$

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 $\sigma_D : (\mathcal{A}, \Theta) \rightarrow (A_D, \mathcal{H})$. Because State A does not know the value of θ, A's strategy is as mapping only from the known parameter α, or $\sigma_A : (\mathcal{A}) \rightarrow (A_A, t)$. 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 $(V_D^g(\sigma_A(\alpha), \sigma_D(\alpha, \theta)) - C_D^g(\sigma_D(\alpha, \theta), \alpha, \theta)$ and $V_A^g(\sigma_A(\alpha), \sigma_D(\alpha, \theta)) - C_A^g(\sigma_A(\alpha), \sigma_D(\alpha, \theta)))$. 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 $U_A(\theta, \alpha)$ are the expected utilities, or

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

and

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

With this defined, I can introduce the condition of interest: when a deterrence 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, C_A^g and C_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 A 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 $C_D^g(\sigma^*(\alpha,\theta),\alpha,\theta)$, the new direct mechanism considers outcome functions $\pi(\tilde{\theta},\alpha)$, $V_D(\tilde{\theta},\alpha)$ and $C_D(\tilde{\theta},\theta,\alpha)$, 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. $\pi(\tilde{\theta},\alpha)=\pi^g(\sigma^*(\alpha,\theta))$, $V_D(\tilde{\theta},\alpha)=V_D^g(\sigma^*(\alpha,\theta))$, $C_D(\tilde{\theta}|\theta,\alpha)=C_D^g(\sigma^*(\alpha,\theta),\alpha,\theta)$, 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 $\theta,\tilde{\theta}\in\Theta$,*

$$(1-\pi(\theta|\alpha))\,W_D+\pi(\theta|\alpha)\,(V_D(\theta|\alpha)-C_D(\theta|\theta,\alpha))\geq\left(1-\pi(\tilde{\theta}|\alpha)\right)W_D+\pi(\tilde{\theta}|\alpha)\left(V_D(\tilde{\theta}|\alpha)-C_D(\tilde{\theta}|\theta,\alpha)\right).$$

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 (Meyerson 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 rising technology 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 $C_D^g(h,\theta,\alpha)$ is differentiable in θ for all $h \in H$ and $\alpha \in \mathcal{A}$, and that the derivative of C_D^g is uniformly bounded, or that, for all α and h, $|\frac{\partial}{\partial\theta}C_D^g(h,\theta,\alpha)| \leq M < \infty$.*

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 $V_D(h,t,a_A,a'_D) - C_D(h,\alpha,\theta) \geq W_D$ 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 deterrence failure occurs following improvements in hassling capabilities.

Definition: *In crisis bargaining game G with equilibrium σ^*, D experiences a* ***deterrence 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* ***deterrence 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 deterrence 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

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

and

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

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

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

and

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

Combining and simplifying the inequalities above yields

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

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

□

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

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

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

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

which violates the Voluntary Agreements assumption.

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

In this case, the "if" clause implies

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

Using incentive compatibility, I can say

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

Simplifying the above gives

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

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

□.

9.4 General Results on α

In this section, I can define the necessary conditions for a deterrence failure following a shifts in α. This is somewhat more difficult to accomplish because changes in α can produce shifts in how A 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 Segal (2003) envelope theorem result.

At this point, I introduce two ways that A's choices in the hassling game G can influence the final outcomes of D's utility. These two ways characterize how A plays the game and offers some simplification to the conditions for a deterrence 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^**, A is "**unconstrained**" in parameter* $\alpha \in \mathcal{A}$ *when* $U_D(\underline{\theta}, \alpha) = W_D$.

A is unconstrained when A is not sufficiently deterred from investing in the rising technology. 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 A is "unconstrained," it implies that A either selects a level rising technology 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 level of rising technology 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).[15] When would A be "constrained"? In the example where A faces costs to selecting t, sometimes A was constrained because their costs from developing the rising technologies were too high to justify selecting a t that gave type $\underline{\theta}$ D their wartime payoff.

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

When A "avoids war," A is not willing to select a level of rising technology 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 A 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 $U_D(\theta, \alpha) = U_D(\underline{\theta}, \alpha) + \int_{\underline{\theta}}^{\theta} \frac{\partial}{\partial \theta} \left(-\pi(\tau|\alpha) C_D(\tau|\tau, \alpha) d\tau\right)$.[16] Additionally, the following conditions hold:*

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

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

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

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

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

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

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

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

(iii.) If

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

in the direct mechanism, then improvements in public hassling capabilities produce a deterrence 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 deterrence failure is determined by three factors: whether improved hassling capabilities results in D becoming more predictable, A becoming emboldened, or A becoming constrained.

When A 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

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

The W_D term represents the costs from going to war. The expression $\int_{\underline{\theta}}^{\theta} \frac{\partial C_D}{\partial \theta} (\tau|\tau, \alpha) \, d\tau$ 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 $\alpha \in \mathcal{A}$, A 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 $\int_{\underline{\theta}}^{\theta} \frac{\partial C_D}{\partial \theta}(\tau|\tau,\alpha)\,d\tau$ expression can be simplified to $C_D(h'|\theta,\alpha) - C_D(h'|\underline{\theta},\alpha)$, making the full for a deterrence failure $C_D(h'|\theta,\alpha') - C_D(h'|\underline{\theta},\alpha') - (C_D(h'|\theta,\alpha) - C_D(h'|\underline{\theta},\alpha)) \geq 0$; 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 deterrence failure.[17] 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 A's selected t; both of which intuitively could be thought of as "predictability."

When A is unconstrained but does not always avoid war (Case ii. in Proposition 9), now whether A 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

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

Outside of D's wartime utility, D's expected utility now consists of the $\int_{\underline{\theta}}^{\theta} \pi(\tau,\alpha)\frac{\partial C_D}{\partial \theta}(\tau|\tau,\alpha)\,d\tau$ 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 A 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 A's greater selected level of rising technology is also eating into D's bargaining surplus.

Finally, when A is constrained (Case iii.), the utility that type $\underline{\theta}$ D attains becomes a salient factor. In the model in the text, A selected a level of rising technology where type $\underline{\theta}$ either went to war (Proposition 1, Case 2) or hassled and was indifferent between war and hassling (Proposition 1, Case 1). However, sometimes the costs that A incurs from hassling or from investing in the rising technology (or both) makes A hold back on their level of

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

investing, 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 deterrence 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 deterrence failures can still occur, but not necessarily (it will be contingent upon the remaining terms).

9.5 Proof of Proposition 10

What follows here is a derivation of Milgrom Segal's 2003 envelope theorem results.[18] 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 $C_D(\theta|\theta, \alpha) = C_D^g(\sigma^*(\theta, \alpha)|\theta, \alpha)$. I define $\Phi(\theta'|\theta, \alpha)$ as the utility a type (θ, α) D receives from announcing type θ',or

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

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

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

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 $C_D(\theta'|\theta, \alpha)$ (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

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

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

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

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, $\frac{\partial}{\partial\theta}C_D \leq 0$.

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

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

And similarly

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

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

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

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

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

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

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

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

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

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

I can then derive the conditions above through algebra.

□