Deterrence Games and the Disruption of Information

Siyu Ma[1] * Yair Tauman[2] † Richard Zeckhauser[3] ‡

July 23, 2023

Abstract

Deterrence is a generic situation where a "Retaliator" (Player R) threatens to bash an "Underminer" (Player U) should he take a stealth threatening move. A typical Underminer is a potential bomb builder, market invader or computer hacker. The Retaliator's decision whether to bash will depend on a noisy signal her intelligence receives about U's action. U may or may not have the ability to disrupt R's signal (type U^+ and U^-, respectively). U's type is his private information. If U can and does disrupt, the signal to R's intelligence is random, in effect noise. The equilibrium of the game is basically unique. U is better off with the disruption capability than without. More accurate intelligence makes R less likely to bash U. Accordingly, all expected payoffs increase. As R's belief about U's ability to disrupt increases, R is more aggressive and U (whether he is able to disrupt or not) is less aggressive. Yet, greater disruption potentially lowers the payoffs of the all players R, U^+ and U^-. Hence a more transparent information system with no potential disruption helps both sides.

Keywords Deterrence, Intelligence, Information Disruption, Noise

[*1] Beijing Jiaotong University

[†2] Adelson School of Entrepreneurship, Reichman University and Economics Department, Stony Brook University

[‡3] Harvard Kennedy School, Harvard University.

1 Introduction

Powerful players, be they nations, companies, or individuals, must always guard against a lesser player who could take secret moves that in expectation would bring them harm. For example, an enemy nation may be pursuing a cloaked effort to develop nuclear weapons to gain some parity with an existing nuclear power. If a potential Underminer, who is the secret mover, can complete the secret move, thereby creating a capability to harm, the powerful player (the Retaliator) will be in a much worse position. A nuclear-armed enemy will be both emboldened and more dangerous.

In some situations, the Retaliator will be able to impair or eliminate the Underminer's newly created capability once it becomes known. Thus, the United States could conceivably knock out much of North Korea's nuclear weapon and missile delivery capabilities. But doing so too late, once the capability is created, or mostly created, would be extremely expensive in terms of dollar cost, military risk, and world opinion. And if the knock-out blow was delivered in error, say because the capability did not exist, as was the case with Saddam Hussein's fictitious weapons of mass destruction, reputational costs could be massive.

A far preferable strategy for the Retaliator would be to deter the Underminer from taking a secret move that will enable him to create a capability of harming her. (To facilitate exposition, Underminers (U) are treated as male and Retaliators (R) as female.) Thus, R will threaten to make a move that imposes significant costs on the Underminer should he take that move. We refer to R's cost-imposing move as *bashing*, and assume that bashing eliminates U's capability of harming R in case he created such a capability.

Thus, this is a two-stage game: U moves first and chooses in secret whether to harm or not harm, then R based on a noisy intelligence system chooses whether to bash or not bash. The two-move sequence concludes the game and determines the outcome, hence the players' payoffs.

Given that the Underminer's move will be taken in secret, i.e., will not be public absent further efforts, the Retaliator will not know what move U took before bashing becomes impossible or extremely costly. Once U's move becomes public information, i.e., once the capability of harming R is created, bashing will no longer be desirable, and perhaps not even feasible. Given R's cost of either unjustified bashing or not bashing when it would avoid harm, she developed an intelligence system before the game began. That system gives her information, albeit imperfect (or noisy) information, on the move taken by U.

The Underminer, as is common in the real world, may have the ability to disrupt the Retaliator's intelligence. Thus, R's intelligence will give him a signal on the move U took, but U may have taken a disrupting action that impairs that signal. For example, Saddam Hussein disrupted U.S. intelligence, which concluded that he did possess weapons of mass destruction.

In this paper we analyze the case where U may or may not have a disruptive technology (whether he does is his private information) that renders the intelligence of R uninformative. The two types are denoted as U^+ and U^-, respectively.

This analysis has two features that distinguish it from classic conceptual models of deterrence: Its focus is on deterring a secret not a public move by the Underminer. It attends to intelligence gathering by the Retaliator. The classic theory, by contrast, focuses on deterring public moves by the Underminer. Hence, intelligence is not a major component. Not a small bit of confusion has arisen, we believe, because players (as well as analysts) have been guided by the classic theory of deterrence in contexts where R's objective in fact was to deter a secret move by U.

To avoid reacting too late, R needs an intelligence system. That system, usually imperfect, will help her to bash on a timely and appropriate basis. Misguided retaliation is costly to both parties. Of course, the system's imperfection blunts R 's ability to retaliate appropriately, and the Underminer likes best the outcome where he takes the harmful move but escapes retaliation.

This outcome of harm without retaliation is made more possible if the Underminer has a technology to disrupt R's intelligence system. The prime focus and main contribution of this paper is on the consequences of this disruption capability in a framework where the Underminer has a secret harmful move. It is shown (not surprisingly) that if there is the potential for U to have a disruptive technology, he is better off having it. Namely, the equilibrium payoff of U^+ is greater than that of U^-. Furthermore, U of both types and R are better off the lower is the belief of R about U possessing a disruptive capability. This result is quite intuitive with regards to U but less so with regards to R. Namely, R benefits as well if she is led to believe that U is less likely to have disruptive capabilities. The Underminer will rarely have a credible ability to forswear his use of his disruptive technology. As R's prior belief on U's ability to disrupt gets stronger, she acts more aggressively and increases the probability of punishing U; having anticipated this punishment, U will more likely be deterred. Overall, a stronger prior belief on U's disruptive capability increases the chance of punishing an innocent U by mistake, thus lowering both players' equilibrium payoffs. In line with Jelnov, Tauman and Zeckhauser (2017), a more accurate intelligence makes both players better off. Therefore, a transparent system with accurate and intact signals benefits both sides. If it can be made credibly, the Retaliator should be convinced that the Underminer has never developed any disruptive technology. However, this ideal result cannot be sustained, since an Underminer is better off with the disruption capability than without.

2 Related Literature

Most[1] analytic discussions of deterrence related to military affairs. It took many brilliant mind years to understand the underpinnings of the deterrent effects of massive retaliation. As Schelling (1960, p. 7) observes: "What is impressive is not how complicated the idea of deterrence has become and how carefully it has been refined and developed, but how slow the progress has been, how vague the concepts still are, and how inelegant the current theory of deterrence is." See Powell (1990) for an early extensive treatment of nuclear deterrence using game theory. These pioneering analyses, as opposed to our analysis, were concerned with deterring public moves. The prime illustration, of course, was deterring a nuclear attack. No move could be more public.

Recently, a new generation has shown how game theory, combined with deep thinking and Bayesian methods, can yield important insights into effective deterrence of moves taken in secret. However, when moves are secret there is the potential for mistaken harmful moves blossoms. Most of these recent analyses continue to be embedded in national security contexts. Debs and Monteiro (2014, p. 1) address secret build ups, preventive wars, including mistaken preventive wars: "states may be tempted to introduce power shifts as a fait accompli. (Concerns about the Underminer's fait accompli potential also feature in entry deterrence models.) They employ the Iraq War to illustrate. It stemmed, they observe, from Iraq's inability to commit not to develop nuclear weapons combined with the United States' inability not to launch a preventive war. Bas and Coe (2012), use historical examples that stretch from precolonial New Zealand to the 1967 Six-Days War to support their model of a potential power shift as a trigger to war.

Information lies at the heart of many analyses of deterrence related to military subjects, such as arms control. Coe and Vaynman (2020, p. 342) highlight the importance of information in any attempt to limit armaments: "The main impediment to arms control is the need for monitoring that renders a state's arming transparent enough to assure its compliance but not so much as to threaten its security." Effective monitoring is thus critical, lest cheating be worthwhile.

Signal disruption, a key feature of this analysis, impedes an opponent's ability to detect one's type. Refusing arms inspections is another way to deny information to a potential Retaliator. Baliga and Sjostrom (2008) show that the strategic ambiguity such refusal introduces can deter attacking. They also show, unfortunately, that ambiguity can lead to mistaken attacks.

Jelnov, Tauman and Zeckhauser (2017) address deterrence, also in the bomb-building context, where intelligence plays a salient role. This paper broadens their analysis by considering a potential signal disruption by the Underminer. Disruptive technologies are a common weapon in real world contexts where the Retaliator has intelligence capabilities. As expected, her

[1]The discussion paper version of this paper contains a much more extensive literature review. See Ma, Tauman and Zeckhauser (August 2020), Harvard Kennedy School Discussion Paper.

choice whether to act cautiously or aggressively depends not only on the intact precision of intelligence, but also on the Underminer's ability to disrupt her intelligence.

To illustrate, cyberattacks for ransom have become a major problem recently. Thus in 2021 DarkSide, a criminal firm probably located in Russia, extorted $4.4 million from each of Colonial Pipeline and Brenntag. The United States threatened retaliation. DarkSide then announced that it was shutting down operations, but experts think that the shutdown may be a ruse to disrupt the intelligence of potential Retaliators, and that Dark Side may be operating again under a changed identity. In context where cyberattacks cannot be perfectly attributed to hackers while a defender wants to retaliate against the guilty attacker only, Baliga, Mesquita and Wolitzky (2020) show that some improvements in attribution can backfire, weakening deterrence.

Patents represent a second area where deterrence and bashing are brought into play. A patent holder must worry about a potential entrant or weak-patent competitor innovating around its patent. The theoretical and empirical literature on using patents to deter entry, dating back to the 1980s, examines an array of sophisticated strategies. For example, Ellison and Ellison (2011) discuss the deterrence behavior of monopolist pharmaceutical firms in the period just prior to patent expiration. They find examples where incumbent firms decreased advertising slightly prior to patent expiration to shrink the market. In our terminology, they gain by pre-bashing the profits of a potential entrant. Of course, the incumbent firm has a countervailing consideration to possibly boost its advertising expenditures in anticipation of patent expiration as a means to reinforce brand loyalty[2].

Entry deterrence by patent holders is a subset of such deterrence by monopolists. Klemperer (1987) shows how an incumbent firm can protect itself against market entry by building brand loyalty, particularly if consumers perceive "switching costs" to be high. Ellison and Ellison (2011) found evidence of such loyalty-building efforts shortly prior to patent expiration. Sophisticated pricing strategies can also be used as a bashing weapon when firms do enter. If consumers value authenticity, firms can bash counterfeiters by raising their prices. Qian (2008) measured a 45% average price increase within two years of infringement in Chinese markets.

The general lesson from this literature review is that deterrence is a broadly observed phenomena. Intelligence gathering and intelligence disruption are potent accompaniments. Our analysis focuses on one-shot games, with two players, one powerful the other much weaker, facing off against one another. Its Bayesian framework with signals, however, is readily adapted to repeated games. Signals can also be player's purposefully revealed information, i.e., communications between players. Those communications can be public, as is common with threats, or conveyed through back channels, or a mediator.

Effective deterrence almost always involves some form of monitoring, such as inspection or intelligence. The player being deterred, because he would like to get away playing his harmful

[2]We are indebted to a referee for making this point.

action, will often seek to disrupt that monitoring. However, as the analysis below will show, his interests may suffer if he has such a disruptive capability.

3 Model

There are two players: an Underminer (U) and a Retaliator (R). U moves first and wants to take an action H that will *harm* R. The Retaliator seeks to deter U by threatening to bash him if he has played H. "Successful" deterrence lowers the probability that the Underminer plays H (taking the harmful action). U can choose H or NH (not taking the harmful action). The Retaliator can either bash, B, or not bash, NB. Table 2 describes the two players' payoffs for the four possible outcomes:

Table 1: Payoff Table

U \ R	NB	B
NH	$w_1, 1$	r_1, r_2
H	$1, 0$	$0, w_2$

It is assumed that $0 < r_i < w_i < 1,\ i = 1, 2$. That is, the Underminer ranks the outcomes (from best to worst) as follows: (H, NB), (NH, NB), (NH, B) and (H, B). The Retaliator ranks the four possible outcomes (from best to worst) as follows: (NH, NB), (H, B), (NH, B) and (H, NB). To simplify the analysis, it is assumed that the number of pure strategies by U are discrete, and not continuous.

The major challenge to the Retaliator is that the Underminer's action is taken in secret. To determine whether or not to bash, R employs a noisy intelligence system to spy on U and thereby detect whether he has taken a harmful action. The intelligence sends one of the two possible signals: h or nh, indicating imperfectly whether U takes a harmful action. The precision of intelligence is α, $\frac{1}{2} < \alpha < 1$. Namely, if U chooses either H or NH, then with probability α, the intelligence sends the signal h or nh, respectively. If $\alpha = \frac{1}{2}$, the signal is completely random.

Whether U has a disruptive capability is his private information. We refer to U^+ as the player U who has a disruptive capability and U^- who does not. It is commonly known that R believes U possesses a disruptive technology with probability $\beta \in (0, 1)$.

If U^+ chooses to operate the technology and Disrupt (D) the signal, then the signal is intercepted and the precision drops to $\frac{1}{2}$ (a random signal is sent to R). If U^+ chooses Not to Disrupt (ND) or if the player is U^-, the precision of the signal, α, remains unchanged.

In summary, the U^+ type of Underminer has four strategies, $A_U^+ = \{H, NH\} \times \{D, ND\}$, the U^- type of Underminer has two strategies, $A_U^- = \{H, NH\}$.

Asymmetric information arises here: U knows his type, but R only knows the probability β of her opponent's type, namely whether or not he possesses the disruptive capability. Based on the binary signal $s \in \{h, nh\}$, R chooses whether (or with what probability) to Bash. The set of pure strategies for R is $A_R = \{B_hB_{nh}, B_hN_{nh}, N_hB_{nh}, N_hN_{nh}\}$.

It is assumed that the game Γ_1 described above is commonly known. Define the set of the six parameters by $\mathcal{W} \times \mathcal{L}$, where

$$\mathcal{W} \equiv \left\{(\alpha, \beta)|\frac{1}{2} < \alpha < 1, 0 < \beta < 1\right\}$$
$$\mathcal{L} \equiv \{(r_1, w_1, r_2, w_2)|0 < r_i < w_i < 1, i = 1, 2\}$$

The sets $\mathcal{W}$ and $\mathcal{L}$ are common knowledge. The case $\beta \in (0, 1)$ indicates the asymmetric information about the disruption capability. Disruption plays no role in Jelnov, Tauman and Zeckhauser (2017) ($\beta = 0$). As a benchmark case, we will compare their result with ours, thereby highlighting the role of information disruption. The case of $\beta = 1$, where it is commonly known that U has the disruptive capability, will be discussed separately (see Proposition 2(iii)).

3.1 Equilibrium Analysis

In this section, we first simplify the original game Γ_1 to its reduced form Γ_0. Then we describe the unique equilibrium of Γ_0 (Proposition 1). Based on the equilibrium outcome, we show how the intelligence quality α and R's belief on U's disruption capability, β, impact the strategies (Proposition 2) and payoffs (Proposition 3) in the equilibrium of Γ_0. All the lemmas and propositions are proven in the Appendix.

The next two lemmas are essential for the equilibrium analysis.

Lemma 1. *The strategy N_hB_{nh} of R is strictly dominated by B_hN_{nh}. In equilibrium, R plays B_hN_{nh} with positive probability, and she does not mix B_hB_{nh} and N_hN_{nh}.*

Corollary 1. *Every equilibrium of Γ_1 is one of the following three types: (i) R plays purely B_hN_{nh}, (ii) R mixes B_hN_{nh} with N_hN_{nh}, or (iii) R mixes B_hN_{nh} with B_hB_{nh}.*

By Lemma 1, following the intelligence's recommendation (B_hN_{nh}) is strictly better for R than acting opposite to it (N_hB_{nh}). It further narrows down R's equilibrium strategy to three possibilities as shown in the corollary, simplifying the analysis.

Lemma 2. *The strategies (H, ND) and (NH, D) of U^+ do not survive iterative elimination of weakly dominated strategies.*

By Lemma 2, Player U^+ is (weakly) better off disrupting the signal when choosing H (a disrupted intelligence is more likely to send the nh signal, thus helping to camouflage his H action). He is (weakly) better off not disrupting when choosing NH (an accurate intelligence is more likely to send the nh signal).
Lemma 2 allows us to consider a simplified game in which the U^+ type Underminer has only two strategies, (H, D) and (NH, ND), and ignore his other two pure strategies (H, ND) and (NH, D) that do not survive iterative elimination of weakly dominated strategies. To minimize notation, we denote the two active strategies, (H, D) and (NH, ND) of U^+, as H and NH, respectively. No confusion should arise, since U^+ prefers combining D with H and ND with NH. We use the same notations H and NH for U^-, who has no capability to disrupt.

Denote the reduced game of Γ_1 by Γ_0. The extensive form of Γ_0 is presented below.

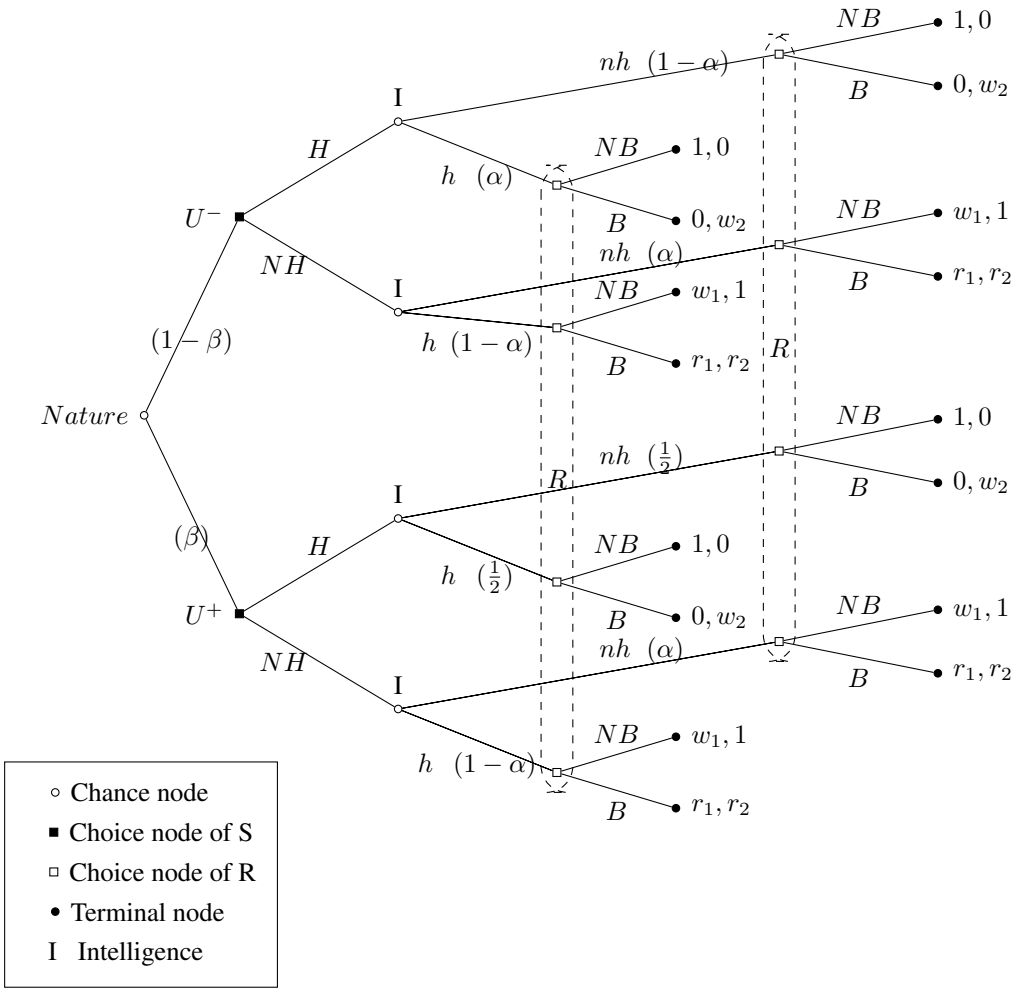

Figure 1: The Reduced Game Γ_0

By Lemma 1, the equilibrium can be of three types, one of which involves a pure strategy of R. In the next lemma, we solve this case.

Lemma 3. *Consider an equilibrium where R plays purely B_hN_{nh}. Then U^+ plays pure H and U^- plays pure NH.*

We have the following lemma for the case where R plays a mixed strategy.

Lemma 4. *Consider an equilibrium where R plays a mixed strategy.*

(i) If $\alpha = \frac{(1-r_2)(1-\beta)-\frac{1}{2}\beta w_2}{(1-r_2)(1-\beta)}$ or $\alpha = \frac{\frac{1}{2}\beta w_2}{(1-r_2)(1-\beta)}$, then U^+ plays pure H and U^- plays pure NH.

(ii) If $\alpha \neq \frac{(1-r_2)(1-\beta)-\frac{1}{2}\beta w_2}{(1-r_2)(1-\beta)}$ and $\alpha \neq \frac{\frac{1}{2}\beta w_2}{(1-r_2)(1-\beta)}$, one type of U plays a pure strategy and the other type randomizes H and NH. U^+ plays more aggressively than U^-.

By Corollary 1, Lemma 3 and Lemma 4, it follows that every equilibrium of Γ_0 (except for $\alpha = \frac{(1-r_2)(1-\beta)-\frac{1}{2}\beta w_2}{(1-r_2)(1-\beta)}$ or $\alpha = \frac{\frac{1}{2}\beta w_2}{(1-r_2)(1-\beta)}$) falls into one of the five categories described in Table 2. Moreover, Proposition 1 below asserts that for every $(r_1, w_1, r_2, w_2) \in \mathcal{L}$, the set $\mathcal{W}$ is partitioned into five regions and each is characterized by one of the categories of Table 2.

Table 2: Equilibrium Categories

Category	R's strategy	U^+'s strategy	U^-'s strategy
1	pure B_hN_{nh}	pure H	pure NH
2	mixes B_hN_{nh} with N_hN_{nh}	pure H	mixes H and NH
3	mixes B_hN_{nh} with N_hN_{nh}	mixes H and NH	pure NH
4	mixes B_hN_{nh} with B_hB_{nh}	pure H	mixes H and NH
5	mixes B_hN_{nh} with B_hB_{nh}	mixes H and NH	pure NH

Proposition 1. *For any $(w_1, r_1, w_2, r_2) \in \mathcal{L}$ and for any $(\alpha, \beta) \in \mathcal{W}$, there is a partition of $\mathcal{W}$ into disjoint regions shown in Figures 2 and 3 s.t. in the interior of each region there exists a unique equilibrium in Γ_0 described in Table 4 and Table 5 in the Appendix.*

(i) Suppose $1 - w_1 \geq r_1$. Then $\mathcal{W}$ is partitioned into five regions (see Figure 2).

- *Γ_0 has a pure strategy equilibrium iff $\frac{1-r_1}{1+w_1-r_1} < \alpha < \frac{\frac{1}{2}-r_1}{w_1-r_1}$ and $\frac{2(1-\alpha)(1-r_2)}{2(1-\alpha)(1-r_2)+w_2} < \beta < \frac{2\alpha(1-r_2)}{2\alpha(1-r_2)+w_2}$. In equilibrium, R plays B_hN_{nh}, U^+ plays H and U^- plays NH.*
- *Γ_0 has a mixed strategy equilibrium in Regions 2 and 3, where α is high. The Retaliator in equilibrium acts mildly. She does not bash if the signal is nh and with positive probability even if the signal is h. The Underminer is more aggressive in Region 2 (where β is low) than in Region 3 (where β is high).*

- Γ_0 *has a mixed strategy equilibrium in Regions 4 and 5, where* α *is low. The Retaliator in equilibrium acts aggressively. She for sure bashes if the signal is* h *and with positive probability even if the signal is* nh*. The Underminer is more aggressive in Region 4 (where* β *is low) than in Region 5 (where* β *is high).*

(ii) Suppose $1 - w_1 < r_1$*. Then* $\mathcal{W}$ *is partitioned into just two regions and the equilibrium strategies coincide with those of Region 2 and Region 3 in part (i) (see Figure 3 below). Only the mild mixed strategy equilibrium exists, and the other three regions of (i) are empty.*

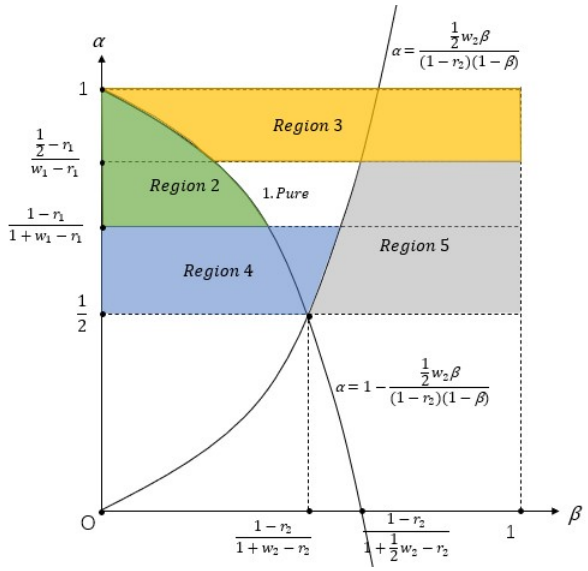

Figure 2: Partition of $\mathcal{W}$ in case $1 - w_1 \geq r_1$

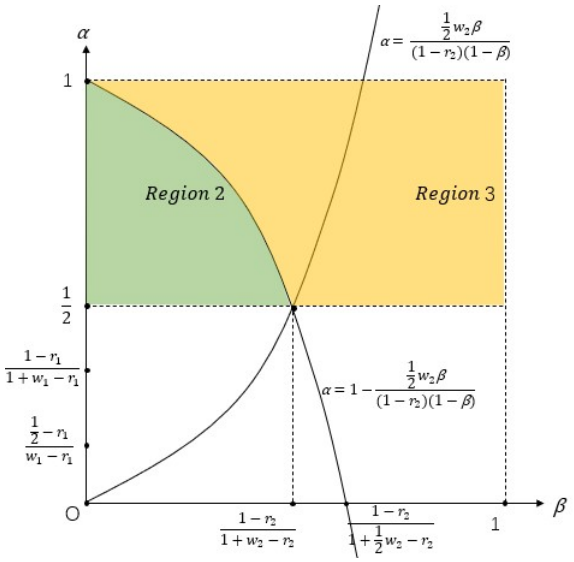

Figure 3: Partition of $\mathcal{W}$ in case $1 - w_1 < r_1$

Before we elaborate on the equilibrium in each region, we define the notion of aggressive and mild strategies and actions of the players.

Definition: (i) We say that U^i, $i \in \{-, +\}$, plays *more aggressively* if he plays H with greater probability. He plays *more mildly* if he plays NH with greater probability.

(ii) For R, the strategy B_hB_{nh} is more aggressive than B_hN_{nh}, and B_hN_{nh} is more aggressive than N_hN_{nh}. We say that R plays *more aggressively* if she plays a more aggressive strategy with greater probability. She plays *more mildly* if she plays a less aggressive strategy with greater probability.

Suppose first $1 - w_1 \geq r_1$, namely, the Underminer is more concerned with missing the opportunity of harming R than about getting caught. See Figure 2.

Region 1. $\frac{1-r_1}{1+w_1-r_1} < \alpha < \frac{\frac{1}{2}-r_1}{w_1-r_1}$ *and* $\frac{2(1-\alpha)(1-r_2)}{2(1-\alpha)(1-r_2)+w_2} < \beta < \frac{2\alpha(1-r_2)}{2\alpha(1-r_2)+w_2}$

In this region, both α and β take intermediate values and R's belief about U's type is ambiguous. Her best choice is simply to follow her intelligence which is moderately accurate but still valuable. U^+ harms her with probability 1 and generates a completely random signal. Hence with probability $\frac{1}{2}$, he will not be bashed (his best outcome). Given R's full trust on her signal, U^- is best off not harming. The outcome is a pure strategy equilibrium.

Region 2. $\frac{1-r_1}{1+w_1-r_1} < \alpha < 1$ *and* $0 < \beta < \frac{2(1-\alpha)(1-r_2)}{2(1-\alpha)(1-r_2)+w_2}$

In this region, α is high and β is low. The signal is relatively accurate and R believes that U is unlikely to disrupt her signal. Since the signal is quite reliable, U^- knows that his action has a good chance to be detected, and he chooses a non-harmful action (NH) with high probability. R then expects to receive the signal nh, and if indeed this is the signal, she for sure does not bash (NB). If, however, unexpectedly she obtains the signal h, she strongly believes this signal is an error of her intelligence since U^- is deterred and the chance of U^+ is quite low. She reduces the probability of mistakenly bashing U by playing NB with positive probability even if the signal is h. Given the milder behavior of R, Player U^+ masks R's signal and with probability 1 harms her (H).

Region 3. $\frac{\frac{1}{2}-r_1}{w_1-r_1} < \alpha < 1$ *and* $\frac{2(1-\alpha)(1-r_2)}{2(1-\alpha)(1-r_2)+w_2} < \beta < 1$ *(it is empty if* $w_1 \leq \frac{1}{2}$*)*

In this region, α is high and β is high. Since the intelligence is more reliable, U^- is again deterred. He lowers the probability to play H in fear of the harmful move being detected. Since R highly believes she faces U^+ who can disrupt her signal, she plays B with a slightly higher probability than in Region 2. This also deters U^+ and result in mild behaviors of all players.

Region 4. $\frac{1}{2} < \alpha < \frac{1-r_1}{1+w_1-r_1}$ *and* $0 < \beta < \frac{2\alpha(1-r_2)}{2\alpha(1-r_2)+w_2}$

In this region, α is low and β is low. R punishes excessively since she cannot trust the signal. R bashes U for sure if the signal is h and with positive probability even if the signal is nh.

However, R believes that her signal is likely intact and this encourages U^+ to knock down R's intelligence. He harms R with probability 1. The result is aggressive behaviors by all players.

Region 5. $\frac{1}{2} < \alpha < \frac{\frac{1}{2}-r_1}{w_1-r_1}$ *and* $\frac{2\alpha(1-r_2)}{2\alpha(1-r_2)+w_2} < \beta < 1$

In this region, α is low and β is high. Knowing that R cannot trust much her signal, U^+ has good reasons to harm R. To deter him, R plays even more aggressively than she plays in Region 4. She bashes U for sure if the signal is h, and bashes U even if the signal is nh with higher probability than in Region 4. R's fierce retaliation deters U, so U acts less aggressively than in Region 4.

Suppose next $1 - w_1 < r_1$. Namely, U is more concerned with getting caught than with missing the opportunity of harming R. The shift of his concern eliminates the aggressive equilibrium regions, leaving the game with only mild equilibrium outcomes. In this case, R never bashes U if the signal is nh and with positive probability even if the signal is h. If she strongly believes U has no ability to disrupt the signal ($0 < \beta < \frac{2\alpha(1-r_2)}{2\alpha(1-r_2)+w_2}$), she bashes him if the signal is h but with probability smaller than in case $\frac{2\alpha(1-r_2)}{2\alpha(1-r_2)+w_2} < \beta < 1$. In the former case, U^+ uses this opportunity to mask her signal and he harms her with probability 1. In the latter case he acts more cautiously and with positive probability he does not disrupt her signal. The pattern of U^-'s behavior is similar, with positive probability he does not harm U if β is small and for sure he does not harm her if β is large.

Proposition 2. *The unique equilibrium of* Γ_0 *satisfies the following.*

(i) As α *increases,* R*'s equilibrium strategy is (weakly) less aggressive.*

(ii) As β *increases,* R*'s equilibrium strategy is (weakly) more aggressive, and the probability of* U *(either type) harming* R *is weakly decreasing.*

(iii) For each $\alpha \in (\frac{1}{2}, 1)$*, the equilibrium strategy profile and payoffs of* Γ_0 *are upper-semi-continuous in* β*,* $\beta \in (0, 1)$*, and are continuous in* $\beta = 0$ *and* $\beta = 1$*.*

A similar result to Proposition 2 part (i) appears in Jelnov, Tauman, Zeckhauser (2017), dealing with the case $\beta = 0$, where it is common knowledge that U has no ability to disrupt[3]. A more accurate intelligence induces a milder behavior of R.

Our new insight is the impact of β on the equilibrium strategies. Proposition 2 part (ii) shows that if R strongly suspects U capable of disrupting the signal, she would act more aggressively and as a result both U^+ and U^- are deterred. Since the equilibrium outcome is continuous in $\beta \in [0, 1]$, $\beta = 1$ (namely, it is common knowledge that U is of type U^+) induces the most

[3]The model in JTZ (2017) is slightly different than the one we have. There U has the option to open his facility fo public inspection. Yet, the conclusion in Proposition 2 (i) and Proposition 3 (i) remain true if this option is eliminated. In line with JTZ (2017), we show that the probability of either type of U harming R is decreasing in α in Regions 2 and 3, independent of α in Region 1, and increasing in α in Regions 4 and 5.

aggressive strategy of R and the mildest strategy of U^+; $\beta = 0$ (it is common knowledge that U is of type U^-) induces the mildest strategy of R and the most aggressive strategy of U^-.

Proposition 3. *(i) As α increases, U^+'s, U^-'s and R's payoffs are strictly increasing in α with one exception: if $1 - w_1 > r_1$, then in Region 1 (the pure strategy equilibrium region), U^+'s payoff is $\frac{1}{2}$, regardless of α. However, better disruption capabilities reduce R's extra benefit gained from a more precise intelligence*

(ii) As β increases, U^+'s, U^-'s and R's payoffs are non-increasing in β.

(iii) U^+ plays more aggressively than U^- (that is, $p^*_+ > p^*_-$), and the payoff of U^+ is at least as high as the payoff of U^-.

A special case of Proposition 3 part (i) was noticed in Jelnov, Tauman, Zeckhauser (2017) (see footnote 3). Greater precision α increases the payoff of U as well as the payoff of R.[4] Moreover, the cross partial of R's equilibrium payoff with respect to α and β is either negative or zero (see Appendix, Table 7).

Part (iii) is intuitive: fixing R's belief β of U's ability to disrupt, U is better off with a disruptive capability than without.

As for Part (ii), the result can be explained by Proposition 2(ii). As β increases, R retaliates more excessively, which deters both U^+ and U^- and lowers U's payoff. Since U of both types acts more mildly, then R's excessive retaliation results in a higher probability of mistakenly bashing an innocent U. This in turn lowers R's payoff as well. By Proposition 2(iii) (the continuity of the equilibrium in $\beta = 0$), the best outcome happens when $\beta = 0$ where it is commonly known that U cannot disrupt.

We just saw that higher assessment of R about U's ability to deploy a disruptive technology hurts him as well as R. In light of Proposition 3(ii) and Proposition 2 (iii), a natural question arises: why doesn't U^+ simply turn off his disruptive technology and make R believe that she faces only Type U^-? Both U^+ and R would gain from a convincing turn-off. The unfortunate reality is that U^+ can not make a "turn off" action credible, since he would be better to keep it turned on, if R for sure believed that it was turned off.

In practice, R can't be sure if U possesses or not a disruptive capability and she assigns some positive probability that U has it. U^+ is better off having the disruptive technology, and at the same time convincing R of his innocence.

[4]This result is based on α being common knowledge. For the analysis of the case where α is private information of R, see Biran and Ma (2022).

4 Discussion

4.1 Endogenous Intelligence and Disruption Capability

In this section[5], we briefly discuss the establishment of the intelligence / counterintelligence abilities, assuming an added first stage where U and R create these abilities at a cost.

Γ_1 without an intelligence is a simultaneous game whose strategic form is shown in Table 1. It can be verified that a unique mixed strategy equilibrium exists there, and R's equilibrium payoff is $f_R^0 \equiv \frac{w_2}{1+w_2-r_2}$ and U's equilibrium payoff is $f_U^0 = \frac{r_1}{1+r_1-w_1}$.

Suppose R creates an intelligence system with precision α at a cost $C_R(\alpha)$. Such an investment is worthwhile as long as R's net payoff is greater than f_R^0. The equilibrium intelligence precision α^* is the maximizer of R's net payoff

$$f_R(\alpha) = \begin{cases} f_R^0 & \text{if } R \text{ does not build IS} \\ \Pi_R(\alpha) - C_R(\alpha) & \text{if } R \text{ builds IS with precision } \alpha \end{cases}$$

where $\Pi_R(\alpha)$ is R's payoff in Table 5 as shown in the Appendix.

As for the disruptive capability of U, the paper assumes that if U possesses such a technology, it functions perfectly. Namely, it destroys R's signal and turns it completely random with probability 1. β is not the quality of the technology; it is the probability R assigns to U possessing such a technology.

A slightly different model can be analyzed where β is the quality of the disruptive technology. Consider a complete information game where it is common knowledge that U has a disruptive technology of quality β. Namely, if U disrupts, then with probability β, the signal turns random and with probability $1-\beta$, it has no effect on R's intelligence. U finds out if his technology is effective only after using it. Suppose we add a preceding stage where simultaneously U and R choose β and α, respectively, and choose whether and to what level to build their technologies. Let $C_t(\alpha, \beta)$, $t \in \{U, R\}$, be t's cost of building an intelligence/ counterintelligence ability, which may depend on both α and β. Then t's net payoff is

$$f_t(\alpha, \beta) = \begin{cases} f_t^0 & \text{if } R \text{ does not build IS} \\ \Pi_t(\alpha, \beta) - C_t(\alpha, \beta) & \text{if } R \text{ builds IS with precision } \alpha \\ & \text{and } U \text{ builds Disruption with quality } \beta \end{cases}$$

where $\Pi_t(\alpha, \beta)$ is t's equilibrium payoff in the subgame starting after the intelligence and counterintelligence capabilities are established. The equilibrium of the two-person game with the payoffs $f_U(\alpha, \beta)$ and $f_R(\alpha, \beta)$ determines α and β endogenously.

[5]We thank the associate editor and a referee for suggesting the two extensions of the model.

4.2 Higher Order of Information Asymmetry

This analysis assumes that all parameters are commonly known. With respect to α and β the paper shows that both players' ex ante expected payoffs are weakly increasing in α and weakly decreasing in β. A recent paper (Biran and Ma (2023)) deals with the case where α is R's private information (disruption of information is not considered there). That paper shows (different from ours) that for some $(r_1, w_1, r_2, w_2) \in \mathcal{L}$, U's ex-post expected payoff is decreasing in α.

We next show that similar outcome happens if β is not common knowledge. Let us make one change in our game Γ_1 and assume the belief of R on U's capability to disrupt is not common knowledge, rather it is R's private information with a commonly known distribution $F(\beta)$. Assume that $F(\beta)$ is binomial: with probability δ, $\beta = \beta_{high}$, and with probability $1-\delta$, $\beta = \beta_{low}$. Denote the two types of R as R^{high} and R^{low}, respectively.

Claim 1. *Suppose $F(\beta)$ follows a binomial distribution $F(\beta) \sim B(1, \delta)$. There exists an equilibrium where R^{high} plays purely B_hB_{nh}, R^{low} plays a mixed strategy $(q^*, 1-q^*)$ over (B_hB_{nh}, B_hN_{nh}), $0 < q^* < 1$, U^+ plays pure H and U^- mixes $(p^*_-, 1-p^*_-)$ over (H, NH).*

Proof see the Appendix.

As an example, for the parameters
$\alpha = 0.55, \beta_{high} = 0.9, \beta_{low} = 0.1, \delta = 0.1, r_1 = 0.125, w_1 = 0.625, r_2 = 0.5, w_2 = 0.75,$

such equilibrium exists, and U^+ plays pure H, U^- plays $(0.38, 0.62)$ over (H, NH), R^{high} plays pure B_hB_{nh}, and R^{low} plays $(0.21, 0.79)$ over (B_hB_{nh}, B_hN_{nh}). Finally, R^{high} obtains 0.735, and R^{low} obtains 0.611. Opposite to the case where β is common knowledge, the expected payoff of R^{high} is greater than that of R^{low}.

Therefore, the common knowledge assumption is essential to drive our main result in the original model.

5 Concluding Thoughts

Deterrence models have played a significant role in the academic literature, and ultimately strategic thinking, since the late 1950s. These models were initially inspired by the need to deter nuclear attacks and military aggression more generally. As would be expected, they received renewed attention in 2022 when the threat of sanctions failed and Russia invaded the Ukraine. Deterrence is a strategy that is employed broadly. When it is, potential Underminers will seek to choose in secret. Potential Retaliators will therefore need to employ an intelligence system lest they bash without justification, and thus lose value. Potential

Underminers may counter that system with a disruptive technology. Surprisingly, employing that technology may hurt rather than help them. Yet, potential Underminers may be unable to foreswear the use of a technology that actually hurts them.

6 Bibliography

1. Avenhaus, Rudolf, Bernhard Von Stengel, and Shmuel Zamir. "Inspection games." Handbook of game theory with economic applications 3 (2002): 1947-1987.

2. Baliga, S. and Sjostrom, T.. "Strategic ambiguity and arms proliferation." Journal of political Economy (2008) 116(6) , pp.1023-1057.

3. Baliga, S., de Mesquita, E.B. and Wolitzky, A.. "Deterrence with imperfect attribution." American Political Science Review 114.4 (2020): 1155-1178.

4. Bas, Muhammet A., and Andrew J. Coe. "Arms diffusion and war." Journal of Conflict Resolution 56 (2012): 651-674.

5. Biran Dov and Siyu Ma, "Attacking a Nuclear Facility: the Impact of a Noisy Intelligence with Unknown Quality" (2023) Working paper.

6. Coe, Andrew J., and Jane Vaynman. "Why arms control is so rare." American Political Science Review 114 (2020): 342-355.

7. Debs, Alexandre, and Nuno P. Monteiro. "Known unknowns: Power shifts, uncertainty, and war." International Organization 68 (2014): 1-31.

8. Ellison, Glenn, and Sara Fisher Ellison. "Strategic Entry Deterrence and the Behavior of Pharmaceutical Incumbents Prior to Patent Expiration." American Economic Journal: Microeconomics 3 (1) (2011): 1–36.

9. Jelnov Artyom, Yair Tauman, and Richard Zeckhauser. "Attacking the unknown weapons of a potential bomb builder: The impact of intelligence on the strategic interaction." Games and Economic Behavior 104 (2017): 177-189.

10. Jelnov Artyom, Yair Tauman, and Richard Zeckhauser. "Confronting an enemy with unknown preferences: Deterrer or provocateur?." European Journal of Political Economy 54 (2018): 124-143.

11. Klemperer, Paul. "Entry Deterrence in Markets with Consumer Switching Costs." The Economic Journal (1987) 97: 99–117.

12. Powell, Robert. “Nuclear Deterrence Theory: The Search for Credibility.” Cambridge, UK: Cambridge University Press (1990).

13. Qian, Yi. “Impacts of Entry by Counterfeiters.” The Quarterly Journal of Economics (2008) 123(4): 1577–1609.

14. Schelling, Thomas C. “The Strategy of Conflict.” Cambridge, MA: Harvard University Press (1980 edition) (initial edition 1960).

7 Appendix

Proof of Lemma 1. Since $\alpha > \frac{1}{2}$, it is easy to verify from Table 3 that N_hB_{nh} is strictly dominated by B_hN_{nh} for the Retaliator, so N_hB_{nh} will never be played in equilibrium. Clearly, no equilibrium exists in which R plays pure N_hN_{nh} or pure B_hB_{nh}.[6]

Based on the signal, R assigns conditional probabilities over U's actions, denoted by $P(H|h), P(NH|h) = 1 - P(H|h), P(H|nh)$ and $P(NH|nh) = 1 - P(H|nh)$. By Baye's Rule,

$$P(H|h) = \frac{\frac{1}{2}\beta\, p_+ + (1-\beta)\alpha p_-}{\beta(1-p_+)(1-\alpha) + \frac{1}{2}\beta p_+ + (1-\beta)(1-p_-)(1-\alpha) + (1-\beta)\alpha p_-} \tag{1}$$

$$P(H|nh) = \frac{\frac{1}{2}\beta\, p_+ + (1-\beta)(1-\alpha) p_-}{\beta(1-p_+)\alpha + \frac{1}{2}\beta p_+ + (1-\beta)(1-p_-)\alpha + (1-\beta)p_-(1-\alpha)} \tag{2}$$

R's conditional expected payoffs are

$$\begin{aligned}
\Pi_2(B|h) &= P(H|h)w_2 + P(NH|h)r_2, \\
\Pi_2(NB|h) &= P(NH|h)\cdot 1 + (1 - P(H|h))\cdot 0 = P(NH|h) \\
\Pi_2(B|nh) &= P(H|nh)w_2 + P(NH|nh)r_2, \\
\Pi_2(NB|nh) &= P(NH|nh)\cdot 1 + (1 - P(NH|nh))\cdot 0 = P(NH|nh)
\end{aligned}$$

R weakly prefers B to NB given h iff

$$\Pi_2(B|h) \geq \Pi_2(NB|h) \longleftrightarrow P(H|h) \geq \frac{1-r_2}{1-r_2+w_2} \tag{3}$$

Similarly, R weakly prefers B to NB given nh iff

$$\Pi_2(B|nh) \geq \Pi_2(NB|nh) \longleftrightarrow P(H|nh) \geq \frac{1-r_2}{1-r_2+w_2} \tag{4}$$

Suppose R in equilibrium mixes N_hN_{nh} and B_hB_{nh}. Then $P(H|h) = P(H|nh) = \frac{1-r_2}{1-r_2+w_2}$. This implies that for $\beta < 1$ the signal is uninformative, contradicting $\alpha > \frac{1}{2}$. □

Proof of Lemma 2. The game Γ_1 in strategic form is given by Table 3 below.

[6] If R plays in equilibrium N_hN_{nh}, both U^+ and U^- are better off playing pure H in which case R is better off deviating to B_hB_{nh}. Similarly, if R plays in equilibrium B_hB_{nh}, both U^+ and U^- are better off playing pure NH, and R is better off deviating to N_hN_{nh}.

Table 3: Γ_1 in Strategic Form

U^+, U^- \ R	N_hN_{nh}	B_hN_{nh}	N_hB_{nh}	B_hB_{nh}
$(NH,ND),NH$	$w_1, w_1, 1$	$\alpha w_1+(1-\alpha)r_1,$ $\alpha w_1+(1-\alpha)r_1,$ $\alpha+(1-\alpha)r_2$	$(1-\alpha)w_1+\alpha r_1,$ $(1-\alpha)w_1+\alpha r_1,$ $1-\alpha+\alpha r_2$	r_1, r_1, r_2
$(NH,D),NH$	$w_1, w_1, 1$	$\frac{1}{2}w_1+\frac{1}{2}r_1,$ $\alpha w_1+(1-\alpha)r_1,$ $\beta(\frac{1}{2}+\frac{1}{2}r_2)+(1-\beta)[\alpha+(1-\alpha)r_2]$	$\frac{1}{2}w_1+\frac{1}{2}r_1,$ $(1-\alpha)w_1+\alpha r_1,$ $\beta(\frac{1}{2}+\frac{1}{2}r_2)+(1-\beta)(1-\alpha+\alpha r_2)$	r_1, r_1, r_2
$(NH,ND),H$	$w_1, 1, \beta$	$\alpha w_1+(1-\alpha)r_1,$ $1-\alpha,$ $\beta[\alpha+(1-\alpha)r_2]+(1-\beta)\alpha w_2$	$(1-\alpha)w_1+\alpha r_1,$ $\alpha,$ $\beta(1-\alpha+\alpha r_2)+(1-\beta)(1-\alpha)w_2$	$r_1, 0, \beta r_2+(1-\beta)w_2$
$(NH,D),H$	$w_1, 1, \beta$	$\frac{1}{2}w_1+\frac{1}{2}r_1,$ $1-\alpha,$ $\beta(\frac{1}{2}+\frac{1}{2}r_2)+(1-\beta)\alpha w_2$	$\frac{1}{2}w_1+\frac{1}{2}r_1,$ $\alpha,$ $\beta(\frac{1}{2}+\frac{1}{2}r_2)+(1-\beta)(1-\alpha)w_2$	$r_1, 0, \beta r_2+(1-\beta)w_2$
$(H,D),NH$	$1, w_1, 1-\beta$	$\frac{1}{2},$ $\alpha w_1+(1-\alpha)r_1,$ $\beta\frac{1}{2}w_2+(1-\beta)[\alpha+(1-\alpha)r_2]$	$\frac{1}{2},$ $(1-\alpha)w_1+\alpha r_1,$ $\beta\frac{1}{2}w_2+(1-\beta)(1-\alpha+\alpha r_2)$	$0, r_1, \beta w_2+(1-\beta)r_2$
$(H,ND),NH$	$1, w_1, 1-\beta$	$1-\alpha,$ $\alpha w_1+(1-\alpha)r_1,$ $\beta\alpha w_2+(1-\beta)[\alpha+(1-\alpha)r_2]$	$\alpha,$ $(1-\alpha)w_1+\alpha r_1,$ $\beta(1-\alpha)w_2+(1-\beta)(1-\alpha+\alpha r_2)$	$0, r_1, \beta w_2+(1-\beta)r_2$
$(H,D),H$	$1, 1, 0$	$\frac{1}{2},$ $1-\alpha,$ $\beta\frac{1}{2}w_2+(1-\beta)\alpha w_2$	$\frac{1}{2},$ $\alpha,$ $\beta\frac{1}{2}w_2+(1-\beta)(1-\alpha)w_2$	$0, 0, w_2$
$(H,ND),H$	$1, 1, 0$	$1-\alpha,$ $1-\alpha,$ αw_2	$\alpha,$ $\alpha,$ $(1-\alpha)w_2$	$0, 0, w_2$

By Lemma 1, we can eliminate N_hB_{nh} of R. Since $\alpha > \frac{1}{2}$, the expected payoff to U^+ when playing (NH, ND) is greater than when playing (NH, D) (See Table 3). Similarly, U^+'s expected payoff by playing (H, D) is greater than that by playing (H, ND). Hence in equilibrium, U^+ must assign zero probability to both (NH, D) and (H, ND).

□

Proof of Lemma 3. Suppose R plays pure B_hN_{nh}. U^+ obtains $\alpha w_1 + (1-\alpha)r_1$ by playing NH, and obtains $\frac{1}{2}$ by playing H. He strictly prefers NH to H iff $\alpha > \frac{\frac{1}{2}-r_1}{w_1-r_1}$.

U^- obtains $\alpha w_1 + (1-\alpha)r_1$ by playing NH, and obtains $1-\alpha$ by playing H. He strictly prefers NH to H iff $\alpha > \frac{1-r_1}{1+w_1-r_1}$.

Hence R is better off deviating to N_hN_{nh} if $\alpha > \max\{\frac{\frac{1}{2}-r_1}{w_1-r_1}, \frac{1-r_1}{1+w_1-r_1}\}$ and she is better off deviating to B_hB_{nh} if $\alpha < \min\{\frac{\frac{1}{2}-r_1}{w_1-r_1}, \frac{1-r_1}{1+w_1-r_1}\}$.

If a pure strategy equilibrium exists, either $\frac{1-r_1}{1+w_1-r_1} \leq \alpha \leq \frac{\frac{1}{2}-r_1}{w_1-r_1}$ or $\frac{\frac{1}{2}-r_1}{w_1-r_1} \leq \alpha \leq \frac{1-r_1}{1+w_1-r_1}$. Note that the latter is true iff $1-w_1 \leq r_1$, but in this case $\frac{\frac{1}{2}-r_1}{w_1-r_1} \leq \frac{1-r_1}{1+w_1-r_1} \leq \frac{1}{2} < \alpha$, a

contradiction. Hence if a pure strategy equilibrium exists, then

$$\frac{1-r_1}{1+w_1-r_1} \leq \alpha \leq \frac{\frac{1}{2}-r_1}{w_1-r_1} \tag{5}$$

and U^+ chooses H and U^- chooses NH. Note that $\alpha \in (\frac{1}{2}, 1)$ satisfies (5) iff

$$w_1 < 1 - r_1 \tag{6}$$

Using Table 3, it is easy to verify that R has no incentive to deviate from $B_h N_{nh}$ given that U^+ chooses H and U^- chooses NH if

$$(1-\beta)[\alpha + (1-\alpha)r_2] + \frac{1}{2}\beta w_2 \geq \max\{\beta w_2 + (1-\beta)r_2, 1-\beta\}$$

Equivalently if

$$\alpha \geq \max\left\{\frac{\frac{1}{2}\beta w_2}{(1-\beta)(1-r_2)}, 1 - \frac{\frac{1}{2}\beta w_2}{(1-\beta)(1-r_2)}\right\} \tag{7}$$

Combining (5) and (7),

$$\max\left\{\frac{\frac{1}{2}\beta w_2}{(1-\beta)(1-r_2)}, 1 - \frac{\frac{1}{2}\beta w_2}{(1-\beta)(1-r_2)}, \frac{1-r_1}{1+w_1-r_1}\right\} \leq \alpha \leq \frac{\frac{1}{2}-r_1}{w_1-r_1} \tag{8}$$

Given $1 - w_1 > r_1$, the interval in (8) is not empty iff $\frac{\frac{1}{2}\beta w_2}{(1-\beta)(1-r_2)} \leq \frac{\frac{1}{2}-r_1}{w_1-r_1}$ and $1 - \frac{\frac{1}{2}\beta w_2}{(1-\beta)(1-r_2)} \leq \frac{\frac{1}{2}-r_1}{w_1-r_1}$. Hence (8) is not void iff

$$\frac{(1-r_2)(w_1-\frac{1}{2})}{(1-r_2)(w_1-\frac{1}{2}) + \frac{1}{2}w_2(w_1-r_1)} \leq \beta \leq \frac{(1-r_2)(\frac{1}{2}-r_1)}{(1-r_2)(\frac{1}{2}-r_1) + \frac{1}{2}w_2(w_1-r_1)} \tag{9}$$

Note that the LHS of (9) imposes no restriction on β if $w_1 < \frac{1}{2}$. If the denominator is positive then the LHS of (9) is negative and if the denominator is negative then LHS> 1 and the sign of the inequality changes to $\beta \leq$ LHS. Also since $w_1 - \frac{1}{2} < \frac{1}{2} - r_1$, (9) defines a non-empty interval.

Hence a pure strategy equilibrium exists iff (8) and (9) hold and in this case R plays pure $B_h N_{nh}$ and obtains $(1-\beta)[\alpha + (1-\alpha)r_2] + \frac{1}{2}\beta w_2$, U^+ plays pure H and obtains $\frac{1}{2}$, and U^- plays pure NH and obtains $\alpha w_1 + (1-\alpha)r_1$. □

Proof of Lemma 4. Consider an equilibrium where R mixes $B_h N_{nh}$ and $N_h N_{nh}$ with probability q and $1-q$, respectively, $0 < q < 1$.

Given q, U^+'s and U^-'s payoffs when playing pure strategies are

$$\Pi_1^+(NH|q) = \Pi_1^-(NH|q) = q[\alpha w_1 + (1-\alpha)r_1] + (1-q)w_1 \tag{10}$$

$$\Pi_1^+(H|q) = \frac{1}{2}\,q + (1-q) > \Pi_1^-(H|q) = q(1-\alpha) + (1-q) \tag{11}$$

Suppose the equilibrium is separating. Then it has to be $p_+ = 1$ and $p_- = 0$ s.t. $\Pi_1^+(H|q) > \Pi_1^+(NH|q) = \Pi_1^-(NH|q) > \Pi_1^-(H|q)$. [7] Since R is indifferent with B_hN_{nh} and N_hN_{nh}, by (3) and (4),

$$P(H|h) = \frac{1-r_2}{1-r_2+w_2} \text{ and } P(H|nh) \leq \frac{1-r_2}{1-r_2+w_2} \tag{12}$$

Plugging in $p_+ = 1$ and $p_- = 0$ into (1) and setting it equal $\frac{1-r_2}{1-r_2+w_2}$, we have

$$\alpha = \frac{(1-r_2)(1-\beta) - \frac{1}{2}\beta w_2}{(1-r_2)(1-\beta)}$$

For other parameters, by (10) and (11), if U^+ randomizes H and NH, U^- must play pure NH; if U^- randomizes H and NH, U^+ must play pure H.

Next consider an equilibrium where R mixes B_hN_{nh} and B_hB_{nh} with probability q and $1-q$, respectively, $0 < q < 1$.

Given q, Player U^+'s and U^-'s payoffs if playing pure strategies are

$$\Pi_1^+(NH|q) = \Pi_1^-(NH|q) = q[\alpha w_1 + (1-\alpha)r_1] + (1-q)r_1 \tag{13}$$

$$\Pi_1^+(H|q) = q \cdot \frac{1}{2} + (1-q)\cdot 0 > \Pi_1^-(H|q) = q(1-\alpha) \tag{14}$$

Suppose the equilibrium is separating. Then it has to be $p_+ = 1$ and $p_- = 0$. Since R is indifferent with B_hN_{nh} and B_hB_{nh}, by (3) and (4),

$$P(H|h) \geq \frac{1-r_2}{1-r_2+w_2} \text{ and } P(H|nh) = \frac{1-r_2}{1-r_2+w_2} \tag{15}$$

Plugging in $p_+ = 1$ and $p_- = 0$ into (2) and setting it equal $\frac{1-r_2}{1-r_2+w_2}$, we have

$$\alpha = \frac{\frac{1}{2}\beta w_2}{(1-r_2)(1-\beta)}$$

For other parameters, by (13) and (14), if U^+ randomizes H and NH, U^- must play pure NH; if U^- randomizes H and NH, U^+ must play pure H.

□

[7] If it is to the contrary that $p_+ = 0$ and $p_- = 1$, then it suggests $\Pi_1^-(H|q) > \Pi_1^-(NH|q) = \Pi_1^+(NH|q) > \Pi_1^+(H|q)$, contradicting with $\Pi_1^+(H|q) > \Pi_1^-(H|q)$.

Proof of Proposition 1.

Case 1. *Pure Strategy Equilibrium.*

This case is shown in Lemma 3.

Next we deal with Case 2 and 3. Consider an equilibrium where R mixes B_hN_{nh} and N_hN_{nh} with probability q and $1-q$, respectively, $0<q<1$. Suppose U^+ chooses H with probability p_+; U^- chooses H with probability p_-.

By (12) and (1), $P(H|h)=\frac{1-r_2}{1-r_2+w_2}$, that is

$$\alpha=\frac{(1-r_2)[\beta(1-p_+)+(1-\beta)(1-p_-)]-\frac{1}{2}\beta w_2 p_+}{(1-r_2)[\beta(1-p_+)+(1-\beta)(1-p_-)]+w_2(1-\beta)p_-} \tag{16}$$

R 's equilibrium payoff is the same whether she plays B_hN_{nh} or N_hN_{nh}, that is

$$\Pi_2(p_+,p_-)=\beta(1-p_+)+(1-\beta)(1-p_-) \tag{17}$$

Case 2. *(Region 2) Suppose* $p_+^*=1$ *and* $0<p_-^*<1$.

Since $\Pi_1^-(H|q)=\Pi_1^-(NH|q)$, by (10) and (11), we have

$$q^*=\frac{1-w_1}{\alpha-(1-\alpha)(w_1-r_1)}$$

and $q^*\in(0,1)$ iff $\alpha>\frac{1-r_1}{1+w_1-r_1}$. This imposes no restriction if $1-w_1<r_1$ since then $\frac{1-r_1}{1+w_1-r_1}<\frac{1}{2}$.

$$\Pi_1^{+*}=\frac{(1+w_1)(\alpha-\frac{1}{2})+r_1(1-\alpha)}{\alpha-(1-\alpha)(w_1-r_1)},\ \Pi_1^{-*}=\frac{(2w_1-r_1)\alpha-(w_1-r_1)}{\alpha-(1-\alpha)(w_1-r_1)} \tag{18}$$

Since $p_+=1$, by (16) we have

$$p_+^*=1,\ \ p_-^*=\frac{(1-\alpha)(1-r_2)(1-\beta)-w_2\frac{1}{2}\beta}{(1-\beta)[\alpha w_2+(1-\alpha)(1-r_2)]}$$

and $0<p_-^*<1$ iff $\beta<\frac{(1-\alpha)(1-r_2)}{(1-\alpha)(1-r_2)+\frac{1}{2}w_2}$ By (17),

$$\Pi_2^*=\frac{(1-\beta)\alpha w_2+w_2\frac{1}{2}\beta}{\alpha w_2+(1-\alpha)(1-r_2)} \tag{19}$$

Case 3. *(Region 3) Suppose* $0 < p_+^* < 1$ *and* $p_-^* = 0$.

Since $\Pi_1^+(H|q) = \Pi_1^+(NH|q)$, by (10) and (11), we have

$$q^* = \frac{1 - w_1}{\frac{1}{2} - (1-\alpha)(w_1 - r_1)}$$

and $q^* \in (0,1)$ iff $\alpha > \frac{\frac{1}{2} - r_1}{w_1 - r_1}$. Note that this region is not empty iff $w_1 > \frac{1}{2}$. The equilibrium payoffs of U^+ and U^- satisfy

$$\Pi_1^{+*} = \Pi_1^{-*} = \frac{w_1 \frac{1}{2} - (1-\alpha)(w_1 - r_1)}{\frac{1}{2} - (1-\alpha)(w_1 - r_1)} \tag{20}$$

Since $p_- = 0$, by (16) we have

$$p_+^* = \frac{(1-\alpha)(1-r_2)}{w_2 \frac{1}{2}\beta + (1-r_2)(1-\alpha)\beta}, \;\; p_-^* = 0$$

and $0 < p_+^* < 1$ iff $\beta > \frac{(1-\alpha)(1-r_2)}{(1-\alpha)(1-r_2) + \frac{1}{2}w_2}$. By (17),

$$\Pi_2^* = \frac{w_2 \frac{1}{2}}{w_2 \frac{1}{2} + (1-\alpha)(1-r_2)} \tag{21}$$

Finally we analyze Case 4 and 5. Consider an equilibrium where R mixes $B_h N_{nh}$ and $B_h B_{nh}$ with probability q and $1-q$, respectively, $0 < q < 1$. In this case, R is indifferent between playing $B_h N_{nh}$ and $B_h B_{nh}$. By (1), (2) and (15), setting $P(H|nh) = \frac{1-r_2}{1-r_2+w_2}$ we have

$$\alpha = \frac{\frac{1}{2}\beta w_2 p_+ + w_2(1-\beta)p_-}{(1-r_2)[\beta(1-p_+) + (1-\beta)(1-p_-)] + w_2(1-\beta)p_-} \tag{22}$$

R obtains the same expected payoff whether she plays $B_h N_{nh}$ or $B_h B_{nh}$,

$$\Pi_2(p_+, p_-) = r_2[\beta(1-p_+) + (1-\beta)(1-p_-)] + w_2[\beta p_+ + (1-\beta)p_-] \tag{23}$$

Case 4. *(Region 4) Suppose* $0 < p_-^* < 0$ *and* $p_+^* = 1$.

Since $\Pi_1^-(H|q) = \Pi_1^-(NH|q)$, by (13) and (14), we have

$$q^* = \frac{r_1}{1 - \alpha - \alpha(w_1 - r_1)}$$

Clearly, $q^* \in (0,1)$ iff $\alpha < \frac{1-r_1}{1+w_1-r_1}$. Note that $\frac{1-r_1}{1+w_1-r_1} > \frac{1}{2}$ iff $1 - w_1 > r_1$ and this region of α is not empty.

Since $p_+ = 1$, by (22), we have

$$p_+^* = 1, \;\; p_-^* = \frac{(1-r_2)(1-\beta)\alpha - \frac{1}{2}\beta w_2}{(1-\beta)[w_2 + (1-r_2-w_2)\alpha]}$$

and $0 < p_-^* < 1$ iff $\beta < \frac{\alpha(1-r_2)}{\alpha(1-r_2)+\frac{1}{2}w_2}$.

By (23),

$$\Pi_2^* = w_2 \cdot \frac{(w_2 - r_2)\beta(\frac{1}{2} - \alpha) + (1-2\alpha)r_2 + \alpha}{w_2 + (1-r_2-w_2)\alpha} \tag{24}$$

The equilibrium payoffs of U^+ and U^- are

$$\Pi_1^{+*} = \frac{r_1 \frac{1}{2}}{1-\alpha-\alpha(w_1-r_1)}, \;\; \Pi_1^{-*} = \frac{r_1(1-\alpha)}{1-\alpha-\alpha(w_1-r_1)} \tag{25}$$

Case 5. *(Region 5) Suppose* $0 < p_+^* < 1$ *and* $p_-^* = 0$.

Since $\Pi_1^+(H|q) = \Pi_1^+(NH|q)$, by (13) and (14), we have

$$q^* = \frac{r_1}{\frac{1}{2} - \alpha(w_1 - r_1)}$$

and $q^* \in (0,1)$ iff $\alpha < \frac{\frac{1}{2}-r_1}{w_1-r_1}$. Since $\alpha > \frac{1}{2}$, we must have $1 - w_1 > r_1$ in this case.

Since $p_- = 0$, by (22),

$$p_+^* = \frac{\alpha(1-r_2)}{[\frac{1}{2}w_2 + (1-r_2)\alpha]\beta}, \;\; p_-^* = 0 \tag{26}$$

and $0 < p_+^* < 1$ iff $\beta > \frac{\alpha(1-r_2)}{\alpha(1-r_2)+\frac{1}{2}w_2}$.

By (23) and (26),

$$\Pi_2^* = w_2 \cdot \frac{\frac{1}{2}r_2 + \alpha(1-r_2)}{\frac{1}{2}w_2 + \alpha(1-r_2)} \tag{27}$$

In this case, U of both types has expected payoff

$$\Pi_1^{+*} = \Pi_1^{-*} = \frac{\frac{1}{2}r_1}{\frac{1}{2} - \alpha(w_1 - r_1)} \tag{28}$$

□

Table 4: Equilibrium Strategies in the five regions

(p^*_+ is the probability of U^+ disrupting and harming. p^*_- is the probability of U^- harming.)

Region	U^+'s eq strategy	U^-'s eq strategy	R's eq strategy
1.Pure	$p^*_+ = 1$	$p^*_- = 0$	$P(B_hN_{nh}) = 1$
2. high α low β	$p^*_+ = 1$	$p^*_- = \frac{(1-\alpha)(1-r_2)(1-\beta)-\frac{1}{2}w_2\beta}{(1-\beta)[\alpha w_2+(1-\alpha)(1-r_2)]}$	$P(B_hN_{nh}) = \frac{1-w_1}{\alpha-(1-\alpha)(w_1-r_1)}$ $P(N_hN_{nh}) = \frac{\alpha w_1-(1-\alpha)(1-r_1)}{\alpha-(1-\alpha)(w_1-r_1)}$
3. high α high β	$p^*_+ = \frac{(1-\alpha)(1-r_2)}{[\frac{1}{2}w_2+(1-r_2)(1-\alpha)]\beta}$	$p^*_- = 0$	$P(B_hN_{nh}) = \frac{1-w_1}{\frac{1}{2}-(1-\alpha)(w_1-r_1)}$ $P(N_hN_{nh}) = \frac{\alpha w_1+(1-\alpha)r_1-\frac{1}{2}}{\frac{1}{2}-(1-\alpha)(w_1-r_1)}$
4. low α low β	$p^*_+ = 1$	$p^*_- = \frac{\alpha(1-r_2)(1-\beta)-\frac{1}{2}w_2\beta}{(1-\beta)[(1-\alpha)w_2+\alpha(1-r_2)]}$	$P(B_hN_{nh}) = \frac{r_1}{1-\alpha-\alpha(w_1-r_1)}$ $P(B_hB_{nh}) = \frac{(1-\alpha)(1-r_1)-\alpha w_1}{1-\alpha-\alpha(w_1-r_1)}$
5. low α high β	$p^*_+ = \frac{\alpha(1-r_2)}{[\frac{1}{2}w_2+(1-r_2)\alpha]\beta}$	$p^*_- = 0$	$P(B_hN_{nh}) = \frac{r_1}{\frac{1}{2}-\alpha(w_1-r_1)}$ $P(B_hB_{nh}) = \frac{\frac{1}{2}-\alpha w_1-(1-\alpha)r_1}{\frac{1}{2}-\alpha(w_1-r_1)}$

Table 5: Equilibrium (ex ante) Payoffs in the five regions

Region	U^+'s payoff	U^-'s payoff	R's payoff
1.Pure	$\frac{1}{2}$	$\alpha w_1 + (1-\alpha)r_1$	$(1-\beta)[\alpha+(1-\alpha)r_2] + \beta\frac{1}{2}w_2$
2	$\frac{(1+w_1)(\alpha-\frac{1}{2})+r_1(1-\alpha)}{\alpha-(1-\alpha)(w_1-r_1)}$	$\frac{(2w_1-r_1)\alpha-(w_1-r_1)}{\alpha-(1-\alpha)(w_1-r_1)}$	$\frac{(1-\beta)\alpha w_2+w_2\frac{1}{2}\beta}{\alpha w_2+(1-\alpha)(1-r_2)}$
3	$\frac{\frac{1}{2}w_1-(1-\alpha)(w_1-r_1)}{\frac{1}{2}-(1-\alpha)(w_1-r_1)}$		$\frac{\frac{1}{2}w_2}{\frac{1}{2}w_2+(1-\alpha)(1-r_2)}$
4	$\frac{\frac{1}{2}r_1}{1-\alpha-\alpha(w_1-r_1)}$	$\frac{r_1(1-\alpha)}{1-\alpha-\alpha(w_1-r_1)}$	$w_2 \cdot \frac{\alpha-(2\alpha-1)r_2-(w_2-r_2)\beta(\alpha-\frac{1}{2})}{w_2+(1-r_2-w_2)\alpha}$
5	$\frac{\frac{1}{2}r_1}{\frac{1}{2}-\alpha(w_1-r_1)}$		$w_2 \cdot \frac{\frac{1}{2}r_2+\alpha(1-r_2)}{\frac{1}{2}w_2+\alpha(1-r_2)}$

Proof of Proposition 2. (i) and (ii) can be verified by Table 4.

(iii) By Table 4, the strategies of U^+ and U^- are continuous in β within each region. By Table 5, the payoff of R is continuous in β within each region. If we plug in the value of β on the border line, it can be verified that they are also continuous on the border. By Table 4, R's strategy is constant in β within each region. By Table 5, the payoffs of U^+ and U^- are constant in β within each region. It can be verified that the value of β on the border line supports any equilibrium of the adjacent regions. Hence the equilibrium is upper-semi-continuous in β.

The continuity of the equilibrium in $\beta = 0$ follows immediately from Jelnov, Tauman and Zeckhauser(2017) and Table 4.

Suppose $\beta = 1$. That is, it is common knowledge that U is of type U^+. Consider an equilibrium where R mixes B_hN_{nh} and B_hB_{nh} with probability q and $1-q$, respectively, $0 < q < 1$. In this case, R is indifferent between playing B_hN_{nh} and B_hB_{nh}. By (1), (2) and

(15), setting $P(H|h) = \frac{1-r_2}{1-r_2+w_2}$ we have

$$\alpha = \frac{w_2 \frac{1}{2} p_+}{(1-r_2)(1-p_+)}$$

equivalently, $p_+^* = \frac{\alpha(1-r_2)}{\frac{1}{2}w_2+\alpha(1-r_2)}$, and

$$\Pi_2^* = r_2(1-p_+) + w_2 p_+ = w_2 \cdot \frac{\frac{1}{2}r_2 + \alpha(1-r_2)}{\frac{1}{2}w_2 + \alpha(1-r_2)}$$

Since U^+ is indifferent with NH and H.

$$\Pi_1^+(NH|q) = q[\alpha w_1 + (1-\alpha)r_1] + (1-q)r_1 = \Pi_1^+(H|q) = \frac{1}{2}q$$

Hence, $q^* = \frac{r_1}{\frac{1}{2}-\alpha(w_1-r_1)}$ and $Pr_2(B_h B_{nh}) = 1 - q^*$. Note that $q^* \in (0,1)$ iff $\alpha < \frac{\frac{1}{2}-r_1}{+w_1-r_1}$. U's expected payoff is $\Pi_1^* = \frac{\frac{1}{2}r_1}{\frac{1}{2}-\alpha(w_1-r_1)}$.
Consider next an equilibrium where R mixes $B_h N_{nh}$ and $N_h N_{nh}$ with probability q and $1-q$, respectively. By (1), (2) and (12), setting $P(H|nh) = \frac{1-r_2}{1-r_2+w_2}$ we have

$$\alpha = \frac{(1-r_2)(1-p_+) - \frac{1}{2}p_+ w_2}{(1-r_2)(1-p_+)}$$

and $p_+^* = \frac{(1-\alpha)(1-r_2)}{\frac{1}{2}w_2+(1-\alpha)(1-r_2)}$. R obtains $\Pi_2^* = 1 - p_+^* = \frac{\frac{1}{2}w_2}{\frac{1}{2}w_2+(1-\alpha)(1-r_2)}$.

Since U is indifferent between H and NH,
$\Pi_1^+(NH) = q[\alpha w_1 + (1-\alpha)r_1] + (1-q)w_1 = \Pi_1^+(H) = 1 - \frac{1}{2}q$. Hence $q^* = \frac{1-w_1}{\frac{1}{2}-(1-\alpha)(w_1-r_1)}$.
Finally, $q^* \in (0,1)$ iff $\alpha > \frac{\frac{1}{2}-r_1}{w_1-r_1}$, and U's expected payoff is $\Pi_1^* = 1 - \frac{\frac{1}{2}(1-w_1)}{\frac{1}{2}-(1-\alpha)(w_1-r_1)}$. This is the same outcome obtained if we substitute $\beta = 1$ in Table 4 and Table 5.

□

Proof of Proposition 3. (i) First note that by Table 5, $\frac{\partial^2 \Pi_1^{t*}}{\partial \alpha \, \partial \beta} = 0$ and $\frac{\partial \Pi_1^{t*}}{\partial \beta} = 0$, for $t \in \{+,-\}$.

By Table 5, we can calculate the derivatives of the equilibrium payoffs with respect to α and β, respectively, as well as the cross partial derivatives of the equilibrium payoffs with respect to α and β. We present these derivatives in Tables 6 and 7.

(ii) Since Player R's payoff within every region is weakly decreasing in β and it is continuous in β in all regions, we conclude that R's payoff is weakly decreasing in β for all $\beta \in [0,1]$.

Table 6: Impact of α on U's Equilibrium (ex ante) Payoff

Region	$\frac{\partial \Pi_1^{+*}}{\partial \alpha}(\geq 0)$	$\frac{\partial \Pi_1^{-*}}{\partial \alpha}(> 0)$
1	0	$w_1 - r_1$
2	$\frac{(1+w_1-r_1)(1-w_1)}{2[(r_1-1-w_1)\alpha-r_1+w_1]^2}$	$\frac{(2w_1-r_1)\alpha-(w_1-r_1)}{\alpha-(1-\alpha)(w_1-r_1)}$
3	$\frac{2(1-w_1)(w_1-r_1)}{(2\alpha r_1-2\alpha w_1-2r_1+2w_1-1)^2}$	
4	$\frac{r_1(1+w_1-r_1)}{2[1+(r_1-1-w_1)\alpha]^2}$	$\frac{r_1(w_1-r_1)}{[1+(r_1-1-w_1)\alpha]^2}$
5	$\frac{r_1(w_1-r_1)}{2\left[\frac{1}{2}+(r_1-w_1)\alpha\right]^2}$	

Table 7: Impact of α and β on R's Equilibrium (ex ante) Payoff

Region	$\frac{\partial \Pi_2^*}{\partial \alpha}(> 0)$	$\frac{\partial \Pi_2^*}{\partial \beta}(\leq 0)$	$\frac{\partial^2 \Pi_2^*}{\partial \alpha \partial \beta}(\leq 0)$
1	$(1-r_2)(1-\beta)$	$-\alpha(1-r_2) - r_2 + \frac{w_2}{2}$	$-(1-r_2)$
2	$\frac{[(2-\beta)(1-r_2)-\beta w_2]w_2}{2((\alpha-1)r_2+\alpha w_2-\alpha+1)^2}$	$-\frac{(\alpha-\frac{1}{2})w_2}{(r_2+w_2-1)\alpha-r_2+1}$	$-\frac{w_2(1+w_2-r_2)}{2[\alpha w_2+(1-\alpha)(1-r_2)]^2}$
3	$\frac{2w_2(1-r_2)}{(2\alpha r_2-2\alpha-2r_2+w_2+2)^2}$	0	0
4	$w_2(w_2-r_2)\cdot\frac{(1-r_2)(1-\frac{1}{2}\beta)-\frac{1}{2}\beta w_2}{[w_2+(1-r_2-w_2)\alpha]^2}$	$-\frac{w_2(w_2-r_2)(\alpha-\frac{1}{2})}{w_2+(1-r_2-w_2)\alpha}$	$-\frac{w_2(w_2-r_2)(1+w_2-r_2)}{2[\alpha(1-r_2-w_2)+w_2]^2}$
5	$\frac{2w_2(1-r_2)(w_2-r_2)}{(2\alpha r_2-2\alpha-w_2)^2}$	0	0

U^+'s and U^-'s payoffs are independent of β within each of the five regions. By Table 5 and using Maple, we can show that for a fixed α, U^+'s and U^-'s payoffs are both decreasing across regions as β increases.

(iii) It can be verified from Table 4 and Table 5. □

Proof of Claim 1. In such equilibrium, both U^+ and U^- assign a probability $\delta + (1-\delta)q$ on R playing B_hB_{nh} and a probability $(1-\delta)(1-q)$ on R playing B_hN_{nh}. The payoffs of U^+ and U^- if playing pure strategies are

$$\Pi_1^+(NH|q,\delta) = \Pi_1^-(NH|q,\delta) = (1-\delta)(1-q)[\alpha w_1 + (1-\alpha)r_1] + (\delta + (1-\delta)q)r_1$$
$$\Pi_1^+(H|q,\delta) = (1-\delta)(1-q)\cdot\frac{1}{2} > \Pi_1^-(H|q,\delta) = (1-\delta)(1-q)(1-\alpha)$$

U^- mixes H and NH, then $\Pi_1^-(H|q,\delta) = \Pi_1^-(NH|q,\delta)$. The solution in q is

$$q^* = 1 - \frac{r_1}{(1-\delta)[1-\alpha-\alpha(w_1-r_1)]}$$

R's best reply (as a function of the signal and β) is the same as in Γ_0. Since R^{low} mixes B_hB_{nh}

and $B_h N_{nh}$, we have (see (15))

$$P_{low}(H|nh) = \frac{1-r_2}{1-r_2+w_2} \tag{29}$$

where $P_{low}(H|nh)$ is R^{low}'s conditional probability on U playing H, given nh. By (2), replacing β with β_{low} and substituting $p_+ = 1$, we have

$$P_{low}(H|nh) = \frac{\frac{1}{2}\beta_{low} + (1-\beta_{low})(1-\alpha)p_-}{\frac{1}{2}\beta_{low} + (1-\beta_{low})(1-p_-)\alpha + (1-\beta_{low})p_-(1-\alpha)} \tag{30}$$

By (29) and (30), it can be verified that

$$p_-^* = \frac{(1-r_2)(1-\beta_{low})\alpha - \frac{1}{2}\beta_{low}w_2}{(1-\beta_{low})[w_2 + (1-r_2-w_2)\alpha]}$$

R^{low} obtains

$$\Pi_2^{low}(1, p_-^*) = r_2[(1-\beta_{low})(1-p_-^*)] + w_2[\beta_{low} + (1-\beta_{low})p_-^*] \tag{31}$$

$$= r_2(1-p_-^*) + w_2 p_-^* + (w_2 - r_2)(1-p_-^*)\beta_{low} \tag{32}$$

and since R^{high} plays purely $B_h B_{nh}$,

$$\Pi_2^{high}(1, p_-^*) = r_2(1-p_-^*) + w_2 p_-^* + (w_2 - r_2)(1-p_-^*)\beta_{high} \tag{33}$$

Next we need to ensure that R^{high} has no incentive to deviate from $B_h B_{nh}$. By (3) and (4), $P_{high}(H|h) \geq \frac{1-r_2}{1-r_2+w_2}$ and $P_{high}(H|nh) \geq \frac{1-r_2}{1-r_2+w_2}$ mus hold. By (1) and (2), for $\beta = \beta_{high}$ and $p_+ = 1$, we have

$$P_{high}(H|h) = \frac{\frac{1}{2}\beta_{high} + (1-\beta_{high})\alpha p_-}{\frac{1}{2}\beta_{high} + (1-\beta_{high})(1-p_-)(1-\alpha) + (1-\beta_{high})\alpha p_-} \geq \frac{1-r_2}{1-r_2+w_2} \tag{34}$$

$$P_{high}(H|nh) = \frac{\frac{1}{2}\beta_{high} + (1-\beta_{high})(1-\alpha)p_-}{\frac{1}{2}\beta_{high} + (1-\beta_{high})(1-p_-)\alpha + (1-\beta_{high})p_-(1-\alpha)} \geq \frac{1-r_2}{1-r_2+w_2} \tag{35}$$

To guarantee the existence of this equilibrium, we need to show that there exist parameters $(\alpha, \beta_{high}, \beta_{low}, \delta, r_1, w_1, r_2, w_2)$ s.t. $p_-^* \in (0,1)$, $q^* \in (0,1)$ and (34) and (35) are satisfied. Namely, $\alpha < \frac{1-\delta-r_1}{(1-\delta)(1+w_1-r_1)}$, $\beta_{low} < \frac{\alpha(1-r_2)}{\alpha(1-r_2)+\frac{1}{2}w_2}$ and β_{high} satisfies (34) and (35). It can be shown that such a parameter set is non-empty. For example, if $\alpha = 0.55$, $\beta_{high} = 0.9$, $\beta_{low} = 0.1$, $\delta = 0.1$, $r_1 = 0.125$, $w_1 = 0.625$, $r_2 = 0.5$, $w_2 = 0.75$, such equilibrium exists. U^+ plays pure H, U^- plays $(0.38, 0.62)$ over (H, NH), R^{high} plays pure $B_h B_{nh}$, and R^{low} plays $(0.21, 0.79)$ over $(B_h B_{nh}, B_h N_{nh})$.

Comparing R^{high}'s payoff (33) to that of R^{low} (31), we show that in the equilibrium of this game R is better off with a higher belief of U being able to disrupt. □