Cyber Deterrence or: How We Learned to Stop Worrying and Love the Signal

Jonathan William Welburn, Justin Grana, Karen Schwindt

RAND National Security Research Division

WR-1294-OSD
July 2019

RAND working papers are intended to share researchers’ latest findings. Although this working paper has been peer reviewed and approved for circulation by RAND National Security Research Division, the research should be treated as a work in progress. Unless otherwise indicated, working papers can be quoted and cited without permission of the author, provided the source is clearly referred to as a working paper. RAND’s publications do not necessarily reflect the opinions of its research clients and sponsors. RAND® is a registered trademark.
Cyber Deterrence or: How We Learned to Stop Worrying and Love the Signal

Jonathan W. Welburn, Justin Grana, and Karen Schwindt

Draft: 7/23/2019

Abstract

Traditional deterrence theory relies on numerous assumptions that in new domains of attack — especially computer networks — may no longer all be valid. One central assumption of traditional deterrence theory is having common knowledge of each actor’s ability to retaliate effectively and accurately (i.e. with perfect attribution). Such an assumption is implausible in the domain of computer security. Instead, attribution is imperfect and it is difficult for each actor to know the retaliation capability of other actors with certainty. Motivated by these features of cyberattacks and retaliation, we examine a game of deterrence between an attacker and a defender. In the game, the attacker does not know the defender’s ability to retaliate but only receives a (possibly noisy) signal from the defender. Similarly, the defender is not able to perfectly attribute an attack but only receives a noisy signal that provides information about the potential attacker. We show that it is never in the best interest of the defender to perfectly signal its retaliation capability. There are (possibly several) equilibria where the defender does not signal any information about its retaliatory capability. However, we show that there are equilibria in which the defender can strategically release noisy information that is imperfectly correlated with its retaliation capability to increase its expected payoffs. While we reveal cases where the defender can use signaling to deter an attacker, we also uncover a counter-intuitive “anti-deterrent” result that illustrates how the defender can increase its expected utility though signaling by inducing the attacker to attack more. The new contributions of this approach have important implications for cyber policy. We find that it is never in the best interest of the defender to signal truthfully, that an effective cyber policy must be flexible and amenable to change, that enhancing the strength of attribution may be the most powerful deterrence tool, and elucidate the curious value of anti-deterrence – a finding which suggests that policy makers should consider whether deterrence is the only option.

Acknowledgements

We would like to thank Sina Beaghley and C. King Mallory for their considerable advice through the duration of this research. We would also like to thank Krista Langeland and Quentin Hodgson in addition to many other RAND colleagues for their useful feedback. This research was sponsored by the Office of the Secretary of Defense and conducted within the Cyber and Intelligence Policy Center of the RAND National Defense Research Institute, a federally funded research and development center (FFRDC). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.
# Table of Contents

Abstract ................................................................................................................. 2  
Acknowledgements .............................................................................................. 2  
1 Introduction ...................................................................................................... 4  
2 Model Outline .................................................................................................. 7  
3 Model Specification ......................................................................................... 9  
3.1 The Attribution Game .................................................................................. 9  
3.2 The Signaling Game ..................................................................................... 11  
4 Results and Analysis ....................................................................................... 14  
4.1 The Attribution Game .................................................................................. 15  
4.2 The Signaling Game ..................................................................................... 19  
4.2.1 Separating Equilibria .......................................................................... 20  
4.2.2 Pooling Equilibria .............................................................................. 22  
4.2.3 Semi-Separating Equilibria ................................................................. 25  
4.2.4 Gains from Signaling ........................................................................... 28  
5 Findings ........................................................................................................... 33  
6 Conclusion — Towards a Cyber Deterrence Policy ....................................... 34  
References .......................................................................................................... 37  
Appendix ............................................................................................................. 38
1 Introduction

Current cyber security best practices are not sufficient. Countless cyberattacks after cyberattack have breached the network defenses of organizations of all sizes and levels of sophistication across the world. Amongst the current of cyber incidents causing significant disruption and economic damage, a growing swell of nation state cyber aggression threatens key institutions. High profile cyberattacks against significant private and public institutions provide numerous examples of nation state led aggression in cyberspace with varying motives and consequences. Cyberattacks against the Democratic National Committee (DNC), Ukrainian power grid, and intrusion into the U.S. power grid over the past few years provide a worsening picture of nation state aggression that can present severe threats to national security. Unfortunately, the threat of cyberattacks against key U.S. institutions and critical infrastructure has outpaced defensive efforts to reduce vulnerabilities (Defense Science Board, 2017).

The result has led to increased calls for countering cyber aggression through deterrence. The recent National Defense Authorization Act (NDAA) for the fiscal year 2019 specifically calls for such a U.S. cyber deterrence policy:

"It shall be the policy of the United States, with respect to matters pertaining to cyberspace, cybersecurity and cyber warfare, that the United States should employ all instruments of national power, including the use of offensive cyber capabilities, to deter if possible, and respond to when necessary, all cyberattacks or other malicious cyber activities of foreign powers that target United States interests" Section 1636 of the 2019 NDAA (115th Congress, 2018).

The language in the latest NDAA tracks with the last two versions; the 2018 and 2017 NDAAAs made similar statements with Congress urging for a deterrence strategy to counter the increasingly obvious threat of cyber aggression. Notably, while there have been high profile data breaches and cyberattacks against the United States. (e.g., Sony Pictures, the U.S. Office of Personnel Management, the U.S. Department of State) and its allies (e.g., TV5Monde, the German Parliament, Saudi Aramco) the 2016 hacking of the DNC by Russian state actors amplified the threat. In the years since, aggression has continued with high profile nation state cyberaggression ranging from global malware attacks (e.g., WannaCry, NotPetya) to critical infrastructure (e.g., Russian cyber intrusions into Ukrainian and U.S. power grids). In each case, a defense-only solution of implementing or improving cyber best practices has been insufficient.
Deterrence should be in place to be leveraged, and potentially acted upon, when defense alone is not sufficient.

It is not surprising then that the strategy of deterrence, which came to prominence in an era of nuclear conflict when the potential reality of any one attack was too significant to bear, has returned to the forefront of the policy debate. Discussions of dissuading adversaries from launching an attack through a deterrence strategy follow from the seminal work of Thomas C Schelling (1960) and the significant body of work on nuclear deterrence (e.g., Robert Powell (1990), James D Morrow (1994)). It is tempting, as conflict shifts into the cyber domain, to draw parallels to previous domains of conflict and the desirability of deterrence. In fact, there are some notable similarities between the deterrence problem in the nuclear domain and in the cyber domain. The significance of the need for a credible threat stressed by Thomas C Schelling (1960) is a problem in both domains. Furthermore, as with nuclear deterrence, cyber deterrence is strengthened by attribution, the use of thresholds, and signaling. Additionally, for deterrence by punishment (as opposed to denial or entanglement), both benefit from the strategic and tactical use of weapons, the superiority of offense over defense, and a preserved second-strike capability. However, that may be where the similarities end.

Developing a strategy of cyber deterrence is not as simple as carrying forward the lessons of nuclear deterrence into a new domain. At even the most basic level, the clarity of the need for cyber deterrence is met with a sharp lack of clarity in what it entails. A report by the Defense Science Board (2017) defines cyber deterrence as:

"The use of both deterrence by denial and deterrence by cost imposition to convince adversaries not to conduct cyberattacks or costly cyber intrusions against the United States, and in at least some instances, to extend this deterrence to protect allies and partners."

Adding further to the unique nature of cyber deterrence, Martin C Libicki (2009), Joseph S Nye (2011), Eric Talbot Jensen (2012), Scott Jasper (2015), Edward Geist (2015), and others identify key distinctions between it and nuclear deterrence. To start, the cyber domain brings lower barriers to entry with respect to technical and financial capabilities, opening the door for numerous actors, state and non-state actors. Aggression in cyberspace has not required existing military actions; while some attacks have occurred during ongoing conflict (e.g., the cyberattack against the Ukrainian power grid) the vast majority (e.g., Titan Rain, Stuxnet, the DNC) have occurred during peace time. Yet, the most significant distinctions are imperfect and the potential for vertical escalation.
Imperfect information in cyberspace cuts in several directions including attack detection and attribution. While imperfect attribution makes identifying the original attacker a challenge, it also makes identifying any retaliation by cyber means and methods, including in a deterrence strategy, a challenge (Martin C Libicki, 2009) \(^1\). Thus, the certainty of cost imposition is not guaranteed. While the nuclear domain holds the potential for unintended consequences, the imperfect information of cyberspace makes unintended consequences a real possibility with each action. Furthermore, the challenge of attribution also inhibits the potential for signaling. Unlike in the nuclear domain where signals can be sent through nuclear weapons testing, signals in cyberspace are hard to send. A test of cyber tools would provide credibility and almost certainly reveal the vulnerabilities they exploit allowing others to quickly offer patches and render the cyberweapon useless. On the other hand, blanket statements of cyber capability may be dismissed as cheap talk. Credible signaling should, therefore, be a significant concern of effective cyber deterrence. Furthermore, the asymmetric nature of cyberwar (where highly capable countries may also have the largest vulnerabilities) and the potential for escalation into the kinetic domain where destruction would be permanent, leads many to believe there is no such thing as mutually assured destruction – the crux of nuclear deterrence – in cyberspace (Michael Sulmeyer, 2018).

For this reason, recent papers have forwarded game-theoretic models with imperfect attribution. Benjamin Edwards et al. (2017) and Sandeep Baliga et al. (2018) both present notable examples of game-theoretic models of cyberattacks with uncertain attribution. Benjamin Edwards, Alexander Furnas, Stephanie Forrest and Robert Axelrod (2017) present a blame game where an attacker chooses to attack based on imperfect knowledge of an attacker's vulnerabilities and a defender chooses to assign blame or not based on uncertain attribution. Our work is more similar to that of Sandeep Baliga, Ethan Bueno de Mesquita and Alexander Wolitzky (2018) which produces a model with a single defender and \( n \) attackers. The attackers choose whether to attack the defender and the defender receives an uncertain attribution signal and chooses whether to retaliate against one or more attackers. Interestingly, the Sandeep Baliga, Ethan Bueno de Mesquita and Alexander Wolitzky (2018) setting finds endogenous complementarity amongst attackers where increasing aggression from the most aggressive attacker incentivizes increasing aggression from all others. Furthermore, Sandeep Baliga, Ethan Bueno de Mesquita and Alexander Wolitzky

\(^1\) While this paper focuses on retaliation within the cyber domain, retaliation does not need to be contained to cyberspace. As described by Mallory (2018), retaliation can be cross-domain fitting within a broad strategy of deterrence.
(2018) find that enhancing attack detection and attacker identification simultaneously (which they jointly refer to as attribution) strengthens deterrence. However, a key component lacking in previous models is the ability for the defender to communicate before the start of the game and possibly convey its attack capabilities.

We advance the discussion on cyber deterrence by introducing a game-theoretic model in the spirit of Sandeep Baliga, Ethan Bueno de Mesquita and Alexander Wolitzky (2018) by considering the role of communication and signaling in cyber deterrence. Fundamentally, our findings support the belief expressed by others that the unique challenges of imperfect attribution ultimately undermine deterrence. With imperfect attribution, we find complete deterrence is unlikely, if not impossible. This challenge has lead others, notably argued by Martin C Libicki (2009), to conclude that a successful cyber deterrence policy is out of reach. While simply improving attribution capabilities, as demonstrated by Sandeep Baliga, Ethan Bueno de Mesquita and Alexander Wolitzky (2018) may be the single best deterrence mechanism, our findings take us further.

Through the addition of signaling, we find that defenders can improve deterrence by through a noisy signal. Thus, even if perfect attribution is not possible (as it likely is not), we find an opportunity to enhance deterrence through the use of signaling.

However, our findings also raise a new question; should deterrence be the goal? While we find evidence that all cyber defenders can improve their welfare by deterring attacks, we also find that the strong defenders may improve their welfare by inducing attacks, which we call "anti-deterrence". The curious finding on the value of anti-deterrence underscores the complexities of cyberspace where a one-size fits all approach is not guaranteed.

2 Model Outline

We consider a two-player sequential-move game with imperfect information between an attacker and a defender. At the start of the game, the defender is given either a “high” or “low” capability to retaliate against attacks. That is, if the defender has a high capability then its retaliation against an attacker does more harm than if the defender was of low capability. We also allow the defender’s capability to determine how much the defender benefits from (correctly) retaliating against the attacker. Since a successful retaliation can be interpreted as stopping an ongoing
attack, it is plausible that a more capable defender receives more benefit from stopping an attack than a less capable defender. In the model, the defender knows its capability with certainty and the attacker does not. Instead, the attacker only knows the probability assigned to the defender’s retaliation capability.

After the defender observes its capability, it chooses how to signal its capability to the attacker. That is, in the sequence of the game, after nature (the personification of randomness) chooses the capability of the defender, the defender moves first. It can either signal that it has a high capability or a low capability. We do not place any restrictions on these signals and there is no cost to signaling. That is, regardless of what the defender’s true capability is, it can costlessly signal any capability.

The attacker moves next. It perfectly observes the defender’s signal and then chooses whether or not to attack. The attacker’s decision to attack is binary and it can only condition its decision on the signal it received from the defender and not the defender’s true capability.

Following the attacker’s decision to attack the defender’s cyber system or not, the defender receives a signal indicating whether or not the attacker has attacked the system or not. Intuitively, this signal would be the joint result of both identification and attribution where both positive identification and positive attribution are required for the defender to receive the signal that the system is under attack by the attacker. However, for clarity we focus on the challenge of attribution. As such, we assume that the defender is able to identify all attacks but is uncertain about attack origin. Thus, the signal simply indicates whether the system is “not under attack by the attacker” or is “under attack by the attacker.” These signals are correlated but not perfectly correlated with the attacker’s action. For example, it is possible for the attacker to choose to attack but the defender to receive a signal that its cyber system has not been attacked by the attacker. This represents an unattributed attack. On the other hand, it is possible for the attacker to choose not to attack but the defender receives a signal that it is under attack. This represents a false alarm or possibly an attack by an exogenous and unmodeled attacker. This signal generating process captures the imperfect attribution aspect of the model. That is to say, even when the attacker chooses to attack, the defender does not know with certainty whether it was actually attacked.

After observing the signal, the defender moves next by choosing whether to retaliate against the attacker. Since there is no restriction on the defender’s actions, it can choose to retaliate even when it receives a signal that its cyber systems are normal. Similarly, it can forego retaliation even if it receives a signal that its systems are under attack. For example, if the defender knows that its detection capabilities are
poor, it might still choose to retaliate against an attacker even if it received a signal that its systems were normal, just because the defender knows that it is likely that the attacker subverted detection methods.

The payoffs depend on the defender’s capability, the attacker’s decision to attack or not attack and the defender’s decision to retaliate. The attacker incurs a reward for attacking but also incurs a cost if the defender retaliates. If the attacker does not attack but the defender retaliates anyway, the attacker still incurs a cost. The attacker incurs a higher cost of retaliation from a defender that has a high capability. The defender incurs a cost when it is attacked but receives a small benefit for correctly retaliating. The defender incurs a cost if it incorrectly retaliates against the attacker. That is, if the attacker chooses not to attack but the defender retaliates, the defender incurs a cost. If the attacker does not attack and the defender does not retaliate, neither players incur rewards or costs.

3 Model Specification

We consider a game theoretic model with two players; attacker \( A \) and defender \( d \), as well as a nature player to capture stochastic elements. We introduce the game in two parts. First, we introduce the attribution game as a game where the defender only has one capability to retaliate and that capability is common knowledge. Second, we expand on the attribution game by introducing the signaling game which includes uncertain information on the defender’s capability, and the ability to signal. We adopt this presentation method because the signaling game, the game we are interested in analyzing, is the attribution with two additional steps: 1) nature chooses the defenders type and 2) the defender chooses how to signal. Therefore, a clear presentation of the attribution game allows for an easier exposition of the entire signaling game.

3.1 The Attribution Game

In the first stage of the attribution game, the attacker chooses to attack \( A \) or don’t attack \( DA \). The defender moves next, by observing whether or not it has been attacked and choosing whether to retaliate. The defender’s observation is drawn by nature. Specifically, the defender either observes a signal that it has been attacked by the attacker, \( o_1 \), or that it has not been attacked by the attacker, \( o_2 \).
For clarity in the definition of attribution, we assume that when the defender is attacked it is able to identify the attack but is uncertain about the attack origin. Thus, it receives signals $o_1$ and $o_2$ based on the actions of the attacker and the strength of attribution. Specifically, if the attacker chooses attack $A$ then the defender receives signal $o_1$ (i.e., attributes the attack to the attacker) with probability $\pi_1$ and receives signal $o_2$ (i.e., is unable to attribute the attack to the attacker) with probability $1-\pi_1$. If, however, the attacker chooses $DA$ then the defender receives observes $o_1$ with probability $\pi_2$ and $o_2$ with probability $1-\pi_2$. Without loss of generality, we assume that $\pi_1 \geq \pi_2$.

Finally, the defender must choose whether to retaliate $R$ or don’t retaliate $DR$ given its observation. The defender’s pure strategy maps its capability observation and the signal it sent to an action $\{R, DR\}$. That means that the defender’s mixed strategy is a function $F: \{o_1, o_2\} \rightarrow [0,1]$.

This strategy represents the probability that the attacker retaliates— chooses action $R$—given its signal. Let $q(o_1)$ be the probability the defender observes $o_1$ and chooses to retaliate and $q(o_2)$ be the probability that defender chooses to retaliate after observing $o_2$.

The payoffs depend on the attacker’s action and the defender’s choice to retaliate. If the attacker attacks, it accrues a payoff of 1. However, if the defender retaliates, it incurs a cost of $c$. If the attacker does not attack but the defender retaliates, we assume the attacker incurs a cost of $v$. We assume that $c > 1 + v$ which establishes two facts about the model. First, it says that an attacker would rather not attack and not be retaliated against than attack and incur the retaliation. Second, this assumption says that the cost incurred by attackers from retaliation is higher when they have actually carried out an attack then when they haven’t. In other words, the attacker incurs more of a cost when it is correctly retaliated against versus when it is incorrectly retaliated against. For example, if a real attacker is caught and retaliated against, a correct retaliation might mean public shaming or the loss of an ongoing attack, which would be costly. However, if the defender incorrectly retaliates, public shaming may be limited if the attacker can show that it was indeed innocent, and since there was no ongoing attack, the attacker does not incur a cost for the cessation of an ongoing attack. In reality, the ability for an attacker to prove themselves innocent may be more challenging than it seems.

For the defender, if it is attacked it incurs a cost of $-1$. If it correctly retaliates, it earns $r$. We assume that $r < 1$ which means that the defender would rather not be
attacked than be attacked and correctly retaliate. If the defender retaliates when the attacker didn’t actually attack, it incurs a cost of $-w$. If there is no attack and no retaliation, both players earn 0. Finally, we make a technical assumption that $1 - \pi_1 c - \pi_2 v \neq 0$ and $1 - \pi_1 c - \pi_2 v \neq 0$. This is an innocuous assumption that allows us to ignore sets of parameters that have measure 0.

3.2 The Signaling Game

The signaling game has the same two players; attacker $a$ and defender $d$, as well as a nature player to capture stochastic elements. However, the first stage of the game is different than the attribution game due to the inclusion of asymmetric information on the defender’s capability to retaliate (i.e., the strength of the defender to retaliate with force). Thus, in the first stage of the signaling game, nature chooses the type of the defender. The defender is of type $H$ (representing a high capability to retaliate) with probability $\gamma$ and is type $L$ (representing a low capability to retaliate) with probability $(1 - \gamma)$.

After the defender is assigned its type, it signals either $s_H$ or $s_L$. Specifically, the defender’s pure strategy is a mapping from $\{H, L\}$ to $\{s_H, s_L\}$. Therefore, the defender’s mixed strategy is a mapping $F: \{H, L\} \rightarrow [0,1]$. That is, the defender chooses the probability for which it signals a high capability for each of its possible types. This function can be represented by two real numbers $\alpha_H$ and $\alpha_L$. Specifically, $\alpha_H$ is the probability that the defender signals $s_H$ — a high capability signal — given it was assigned a high capability and $\alpha_L$ is the probability the defender signals $s_H$ given that nature assigned it a low capability. Analogously, $(1 - \alpha_H)$ and $(1 - \alpha_L)$ is the probability that the defender signals $s_L$ — a low capability signal — given that nature assigned it a high and low capability, respectively.

---

2 While this assumption is generally valid and consistent with requirements from the Law of Armed Conflict requiring proportionality in retaliation, one could imagine cases where retaliation strengthens the defender’s security posture.

3 A more technical justification for this assumption is to assume that at least one of the parameters are drawn from a continuous probability distribution at the start of the game and all players observe the parameter value. Then the probability that $1 - \pi_1 c - \pi_2 v = 0$ is exactly zero and such a case can be ignored without changing the fundamental nature of the game.
After the defender’s signal, the attacker and defender play the attribution game with the added complexity of signaling. As such, the attacker moves next; it observes the signal and chooses whether or not to attack. The attacker’s pure strategy is then a mapping from \( \{ s_H, s_L \} \) to \( \{ A, DA \} \), and thus the attacker’s mixed strategy is a mapping \( G; \{ s_H, s_L \} \rightarrow [0,1] \). Intuitively, the attacker’s mixed strategy assigns the probability of attack, \( A \), conditional on the signal that it received. This strategy can be represented by two real numbers \( \beta_H \) and \( \beta_L \) where \( \beta_H \) is the probability the attacker chooses \( A \) given that it received the signal \( s_H \) and \( \beta_L \) is the probability that attacker chooses \( A \) given it received the signal \( s_L \). Of course, \( (1-\beta_H) \) and \( (1-\beta_L) \) is the probability the attacker doesn’t attack (chooses action \( DA \)), given it received signal \( s_H \) and \( s_L \) respectively.

After the attacker’s action is drawn according to the attacker’s mixed strategy, the defender’s observation is drawn by nature. As in the attribution game, the defender either observes \( o_1 \) or \( o_2 \) where the probability of each signal depends on the attacker’s action. Specifically, if the attacker chooses to attack, the defender observes \( o_1 \) with probability \( \pi_1 \) and \( o_2 \) with probability \( (1-\pi_1) \). If the attacker does not attack, then the defender observes \( o_1 \) with probability \( \pi_2 \) and \( o_2 \) with probability \( (1-\pi_2) \). Intuitively, \( \pi_1 \) and \( \pi_2 \) represent the defender’s ability to attribute an attack. For example, if \( \pi_1 = \pi_2 \), then the signal does not depend on the attacker’s action and the defender does not learn anything from the signal. If \( \pi_1 = 1 \) and \( \pi_2 = 0 \), then the defender can perfectly attribute attacks. Without loss of generality, we assume that \( \pi_1 \geq \pi_2 \).

Finally, the defender must choose whether or not to retaliate given its observation. The defender’s pure strategy maps its capability, observation and the signal it sent to an action \( \{ R, DR \} \) (\( R \) for retaliate and \( DR \) for don’t retaliate). That means that the defender’s mixed strategy is a function \( F; \{ o_1, o_2 \} \times \{ s_H, s_L \} \times \{ H, L \} \rightarrow [0,1] \). This strategy represents the probability that the attacker retaliates—chooses action \( R \)—given its signal. Let \( q(x,y,z) \) be the probability the defender retaliates after observing observation \( x \), signaling \( y \) and having type \( z \). For example, \( q(o_1,s_H,H) \) is the probability that a defender of high capability that signaled \( s_H \) and observed \( o_1 \) chooses to retaliate. Let \( q \) (without subscripts) be shorthand for the set of \( q(x,y,z) \) in the defender’s strategy that give the retaliation probabilities.

As in the attribution game, the payoffs depend on the attacker’s action, the defender’s capability and the defender’s choice to retaliate. If the attacker attacks, it
accrues a payoff of 1. However, if the defender retaliates, the attacker incurs a cost of $c_H$ if the defender has high capability and $c_L$ if the defender has low capability. If the attacker does not attack but the defender retaliates, we assume the attacker incurs a cost of $v$, regardless of the defender’s type. Since a defender of high capability is more able to punish, we assume $c_H > c_L$. We also assume that $c_H > 1 + v$. This assumption implies that when the defender has high capability, the attacker would prefer to not attack and not be retaliated over attacking and incurring a retaliation. Secondly, we assume that $c_L > v$. This means that the cost to being correctly retaliated against is always worse than being incorrectly retaliated against. For technical convenience, we assume that $1 - \pi_1 c - \pi_2 v \neq 0$ and $1 - \pi_1 c - \pi_2 v \neq 0$. This is an innocuous assumption that allows us to ignore sets of parameters that have measure 0.4

For the defender, if it is attacked it incurs a cost of $-1$. If it correctly retaliates, it earns $r_H$ if it is has a high capability to retaliate and $r_L$ if it has a low capability. We assume $r_H > r_L$. We also assume that $r_H r_L < 1$ which means that the defender would rather not be attacked than be attacked and correctly retaliate. If the defender retaliates when the attacker didn’t actually attack, it incurs a cost of $-w$. If there is no attack and no retaliation, both players earn 0. The extensive form version of the game is given in Figure 1 which illustrates the sequence of events, the information sets and the payoffs.

---

4 A more technical justification for this assumption is to assume that at least one of the parameters are drawn from a continuous probability distribution at the start of the game and all players observe the parameter value. Then the probability that $1 - \pi_1 c - \pi_2 v = 0$ is exactly zero and such a case can be ignored without changing the fundamental nature of the game.
Figure 1: Extensive form representation of the signaling game. “Circle” nodes are attacker nodes and “square” nodes are defender nodes. Nodes of the same color are in the same information set. Probabilities for the nature player are given in Greek letters and actions for the attacker and defender are given in Latin letters.

The solution concept we will use to analyze this game is the Perfect Bayesian Equilibrium (henceforth equilibrium). Under such a solution concept, player’s strategies are optimal given their beliefs and beliefs are derived using Bayes’ rule wherever possible. Since all beliefs can be defined using Bayes’ rule at all equilibria, no further equilibrium refinement is necessary.

4 Results and Analysis

To derive equilibria in the full signaling game described above, we will first analyze a subgame that we call the attribution game. In a fashion similar to that of the games forwarded by Benjamin Edwards, Alexander Furnas, Stephanie Forrest and Robert Axelrod (2017) and Sandeep Baliga, Ethan Bueno de Mesquita and Alexander Wolitzky (2018), the attribution game is a game where the defender only has one capability and that capability is common knowledge, thus making signals irrelevant. In reference to Figure 1, the attribution game is a game that starts at any of the attacker’s action nodes as if it were a singleton information set. The sequence of events in the attribution game are then: 1) the attacker chooses whether or not to attack, 2) the
defender receives a signal that is correlated with the attack and then 3) the defender chooses whether or not to retaliate. Since we assume that the defender’s capability is common knowledge in the attribution (but not in the signaling game) game, let \( r \) be the reward the defender receives from correctly retaliating and \( c \) be the cost to the attacker from being retaliated against after attacking.

4.1 The Attribution Game

We begin by solving for the unique Nash equilibrium of the attribution game described above. In the proofs, we assume that \( 1-c<-v \) in the attribution game. Otherwise, there is an equilibrium where the attacker always attacks and the defender always retaliates. However, when considering the full signaling game later, a key equilibrium arises when \( 1-c_H<-v \) but \( 1-c_L>0 \), which is obscured in the attribution game. All formal proofs are given in the appendix.

**Proposition 1 (Equilibrium in the Attribution Game)** Suppose \( 1-c<-v \). Let \( \beta \) be the probability the attacker attacks in the attribution game. Then:

1. If \( 1-\pi_1 c + \pi_2 v > 0 \), there exists a Nash equilibrium of the attribution game where
   - the attacker randomizes with probability \( \beta_1^* = \frac{\pi_1 w}{\pi_1 r + \pi_2 w} \) and the defender never retaliates after observing \( o_2 \) and randomizes between retaliating and not retaliating after \( o_1 \) with probability \( \varrho_1^* = \frac{1}{\pi_1 c - \pi_2 v} \).

2. If \( 1-\pi_1 c + \pi_2 v < 0 \), there exists a Nash equilibrium of the attribution game where
   - the attacker randomizes with probability \( \beta_2^* = \frac{(1-\pi_2)w}{(1-\pi_1)r + (1-\pi_2)w} \) and the defender always retaliates after observing \( o_1 \) and randomizes between retaliating and not retaliating after \( o_2 \) with probability \( \varrho_2^* = \frac{1-\pi_1 c + \pi_2 v}{(1-\pi_1)c - (1-\pi_2)v} \).

**Corollary 1 (Uniqueness of Equilibrium in Attribution Game)** The Nash equilibria in proposition 4.1 are unique.

Proposition 4.1 and corollary 1 establish that there is a unique equilibrium in the attribution game, but the nature of the equilibrium depends on the value of the parameters. To better understand the equilibrium, first consider why it is impossible for there to be a Nash equilibrium in pure strategies. If the attacker always attacks,
then the defender has a best response to retaliate, regardless of its signal. However, if the defender always retaliates, the attacker is better off not attacking. Similarly, from the perspective of the defender, if it chooses the pure strategy of never retaliating, the attacker’s best response to that is to always attack, in which case the defender would have a profitable deviation to always retaliate. Therefore, there cannot be a pure strategy equilibrium.

For there to be a mixed strategy equilibrium, both the defender and the attacker randomize to make the other player indifferent among at least two of their strategies. The defender only has to consider three out of its four pure strategies:

1. Always retaliate
2. Never retaliate
3. Retaliate after \( o_1 \) and don’t retaliate after \( o_2 \).

The strategy ”Retaliate after \( o_2 \) and don’t retaliate after \( o_1 \)” is dominated because for any fixed value of attack probability, \( \beta \), Bayesian beliefs necessitate that an attack was more likely if the defender observes \( o_1 \) then if it observed \( o_2 \). Therefore, retaliating after \( o_2 \)—when the defender is less certain there was an attack—and not retaliating after \( o_1 \)—when the defender is more certain there was an attack—is a dominated strategy.

Figure 2 illustrates the defender’s expected utility \( U_d \) for each of its three strategies as a function of the attack probability, \( \beta \). The purple line extending from the origin is the defender’s expected utility from never retaliating. The blue line with an intercept at \(-\pi_2 w\) is the defender’s expected utility from retaliating after observing \( o_1 \) and not retaliating after \( o_2 \). The red line is the defender’s expected utility from always retaliating. Finally, the gray shaded outlines the defender’s best response for each value of \( \beta \). Specifically, for \( \beta < \frac{\pi_2 w}{\pi_1 r + \pi_2 w} \) the defender’s best response is to never retaliate. For \( \frac{\pi_2 w}{\pi_1 r + \pi_2 w} < \beta < \frac{(1-\pi_2) w}{(1-\pi_1) r + (1-\pi_2) w} \) the defender’s best response is to retaliate only after observing \( o_1 \). Finally, for \( \beta > \frac{(1-\pi_2) w}{(1-\pi_1) r + (1-\pi_2) w} \) the defender’s best response is the always retaliate. The legend in the Figure lists the equations of each of the lines.
The slope of the blue line and the red line in Figure 2 are determined by $\pi_1 r + \pi_2 w - 1$ and $r + w - 1$, respectively. By assumption, these are never 0. However, there is no restriction on their sign (except that $\pi_1 r + \pi_2 w - 1 < r + w - 1$). Therefore, the defender’s best response curve may appear qualitatively different, as shown in Figure 3.
Independent of the slope of the curves, there are two points where the defender is indifferent. These occur at
\[ \beta_1^* = \frac{\pi_2 w}{\pi_1 r + \pi_2 w} \quad \text{and} \quad \beta_2^* = \frac{(1-\pi_2)w}{(1-\pi_1)r + (1-\pi_2)w}. \]
Since the defender cannot have a pure strategy in equilibrium, it must be willing to randomize between at least two strategies. Therefore, the equilibrium attacker randomization probability must be at either one of these two values of \( \beta \).

From the attacker’s perspective, it is willing to randomize if it is indifferent between attacking and not attacking. That means the defender must randomize either after observing \( o_1 \) or after \( o_2 \) to make the attacker indifferent. Suppose the defender randomizes after \( o_1 \) and never retaliates after \( o_2 \). The randomization probability that would make the attacker indifferent between attacking and not attacking is
\[ \frac{1}{\pi_1 c - \pi_2 v}. \]
Of course, this is only a proper probability if \( \pi_1 c - \pi_2 v > 1 \). This means that if the cost of an attacker getting caught is too low (low value of \( c \)) or if its penalty of being incorrectly retaliated against is too high (high value of \( v \)), the defender cannot retaliate with a high enough probability after \( o_1 \) to make the attacker indifferent between attacking and not attacking. In other words, if the cost of being correctly retaliated against is sufficiently close to the cost of being incorrectly retaliated against, the attacker should always just attack since the cost it incurs due to a retaliation does not significantly depend on whether or not it attacked.
The previous analysis can also be phrased in terms of the defender’s attribution probabilities. If the defender’s attribution ability is low ($\pi_1$ is only slightly greater than $\pi_2$), then the attacker knows that its attack is not correlated with the defender’s signal and thus attacking has no effect on whether or not the defender will retaliate. If this is the case, it is always in the attacker’s best interest to attack. Conversely, if the defender’s attribution ability is high ($\pi_1$ significantly greater than $\pi_2$) the attacker knows that if it attacks, it is likely to generate a signal leading to detection and therefore the defender is capable of randomizing to make the attacker indifferent between attacking and not attacking.

The converse argument holds for the case when the defender always retaliates after $o_1$ and randomizes its retaliation after $o_2$. The attacker can only be made indifferent between attacking and not attacking if $\pi_1 c - \pi_2 v < 1$. In this case, if $c$ is relatively high and $v$ is relatively low and the defender will always retaliate after $o_1$, the attacker will incur a high cost when it does attack and is retaliated against. Since the defender’s strategy says to always retaliate after $o_1$, the defender cannot randomize with a low enough probability after $o_2$ to ever induce the attacker to attack because the potential cost to attacking is so high.

Again, the analysis can be phrased in terms of the defender’s attribution parameters. If the defender has high attribution ability ($\pi_1$ much greater than $\pi_2$) then the attacker knows that if it attacks, it is likely that the defender retaliates. This is because the defender retaliates after observing $o_1$ and when $\pi_1$ is high, the defender is more likely to observe $o_1$. Therefore, the defender cannot retaliate with a low enough probability after observing $o_2$ to compensate for the loss the attacker incurs when it attacks and is retaliated against because the attack generated signal $o_1$.

<table>
<thead>
<tr>
<th></th>
<th>$1 - \pi_1 c + \pi_2 v \leq 0$</th>
<th>$1 - \pi_1 c + \pi_2 v \geq 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_d$</td>
<td>$-\beta_1^*$</td>
<td>$\beta_2^<em>(r - 1) - (1 - \beta_2^</em>)w$</td>
</tr>
<tr>
<td>$U_a$</td>
<td>$-\pi_2 \rho_1 v$</td>
<td>$-(\pi_2 + (1 - \pi_2) \rho_2^*) v$</td>
</tr>
</tbody>
</table>

Table 1: Equilibrium expected utilities in the attribution game

Since $\beta_1^* < \beta_2^*$, the attacker attacks with a lower probability when $\pi_1 c - \pi_2 v > 1$, which occurs when either the defender’s ability to attribute with a high degree of certainty or the punishment for correctly retaliating is significantly higher than the punishment for incorrect retaliation. However, this does not mean that the defender is better off in the equilibrium where the attacker attacks less. Table 1 shows the defender’s expected utility under each equilibrium. If $\pi_1 r_H + \pi_2 w > 1$, then the
defender’s expected utility is higher when the attacker attacks *more*. We will return to this point in section 4.2.4 when we discuss the question “is deterrence the right goal?”

### 4.2 The Signaling Game

We now turn our attention to the signaling game. Specifically, we examine whether there are equilibria in which the defender attempts to signal its true capability to the attacker. An important parametric assumption is whether $1-c_L < -v$ (by assumption $1-c_H < -v$ always holds). If $1-c_L < -v$, then the attacker would be willing to attack *if it knew with certainty* that the defender is type $L$ with a strategy to retaliate regardless of the signal. This is because the attacker’s payoff for attacking and getting punished $(1-c_L)$ is greater than its payoff from not attacking and getting punished $(-v)$. Therefore, if the attacker knew the defender were type $L$ and would always retaliate, its best response would be to always attack.

#### 4.2.1 Separating Equilibria

First, we establish that there is no separating equilibrium in which the defender’s signal truthfully reveals its type, regardless of the sign of $1-c_L + v$.

**Proposition 2 (No Separating Equilibrium)** Assume $1-c_L < -v$. Then, there is no equilibrium where the defender truthfully signals its type in the signaling game. Formally, there is no PBE where $\alpha_1 = 1$ and $\alpha_2 = 0$.

**Proposition 3 (No Separating Equilibrium II)** Assume $1-c_L > -v$. Then, there is no equilibrium where the defender truthfully signals its type in the signaling game. Formally, there is no PBE where $\alpha_1 = 1$ and $\alpha_2 = 0$.

Although the formal proofs of propositions 2 and 3 proceed differently, the logic is similar. If the defender truthfully signals its type to the attacker, then after the signal, the attacker and defender just play the attribution game analyzed in the previous section. However, the defender of type $L$ or type $H$ would be better off if it could deceive the attacker in playing a different attribution game. In other words, in some cases a defender of type $H$ would be better off if it could convince the attacker it is type $L$ and have the attacker choose its strategy *as if* the defender is type $L$. In other cases, the defender of type $L$ would be better off if it could convince the attacker that it was type $H$ and have the attacker play the attribution game *as if* the defender were type $H$.
Figure 4: The defender’s best responses in the signaling game. The solid purple line extending from the origin is the defender’s payoffs from never retaliating. The solid lines represent the defender’s payoff when it is type $H$ and the dashed lines represent its payoffs when it is type $L$. The four labeled values of $\beta$ denote the points where the defender is indifferent between two of its strategies.

Figure 4 illustrates the payoff for the defender of type $L$ and type $H$ in the signaling game and illustrates why there can’t be a separating equilibrium. If the defender did truthfully signal its type, then after the signal, the attacker and defender play the attribution game described above. Since there is a unique equilibrium in the attribution game, the attacker’s randomization probabilities after receiving signal $s_H$ are either $\frac{\pi_2 w}{\pi_1 r_H + \pi_2 w}$ or $\frac{(1-\pi_2) w}{(1-\pi_1) r_H + (1-\pi_2) w}$ and after receiving signal $s_L$, randomizes with probability $\frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$ or $\frac{(1-\pi_2) w}{(1-\pi_1) r_L + (1-\pi_2) w}$. All four of these randomization probabilities are annotated in Figure 4.

For any two of the equilibrium probabilities annotated in Figure 4, it is clear that either a defender of type $H$ or a defender of type $L$ would have an incentive to switch its signal. For example, suppose the parameters were such that in the attribution game, when the defender is type $H$ the attacker randomizes with probability...
\[ \beta_H = \frac{\pi_2 W}{\pi_1 r_H + \pi_2 W} \] and when the defender is type \( L \) randomizes with probability

\[ \beta_L = \frac{\pi_2 W}{\pi_1 r_L + \pi_2 W} \]

These points are where the solid blue line and the dashed blue line intersect the purple line extending from the origin. In this case, the defender of type \( L \) would have an incentive to signal \( s_H \) because its expected utility along its best response curve is higher at \( \beta_H \). Of course, there are other possible equilibrium randomization probabilities in the attribution game and other versions of the graph (versions with the blue lines sloping upward) but in all cases, either a defender of type \( H \) or a defender of type \( L \) would have an incentive to not truthfully signal.

### 4.2.2 Pooling Equilibria

Before analyzing semi-separating equilibria, we present the possible pooling equilibria. In a pooling equilibrium, the defender’s signal conveys no information regarding its true type and thus the attacker ignores the signal and only chooses one value of \( \beta \) for which to randomize its attack.

There exists a pure strategy equilibrium if \( \gamma(1-c_H) + (1-\gamma)(1-c_L) > -v \). In such an equilibrium, the attacker always attacks and the defender always retaliates. This is because the (net) cost to the attacker of being correctly retaliated against when the defender is type \( L \) is relatively small compared to its cost of being incorrectly retaliated against. Therefore, if the defender is highly likely to be of type \( L \) (low value of \( \gamma \)), then an equilibrium attacker strategy is to always attack because the frequency in which the defender is type \( H \) and retaliates is not enough to deter the attacker from attacking when the defender is type \( L \) and has a relatively weak ability to punish.

If \( \gamma(1-c_H) + (1-\gamma)(1-c_L) < -v \), there is no pure strategy equilibrium in which the attacker always attacks or never attacks. This implies that the defender must also play a mixed strategy in order to make the attacker indifferent between attacking and not attacking\(^5\). For the defender to be willing to randomize at one of its information sets, it must be indifferent between two actions at that information set. This implies that a pooling equilibrium must have the attacker randomize with one of the four probabilities given in Figure 4. Table 2 gives the possible pooling equilibria.

---

5 Again, there is a measure zero set of parameters where the defender would not have to randomize to make the attacker indifferent, but we ignore such a case for the reasons given above.
<table>
<thead>
<tr>
<th></th>
<th>(\beta)</th>
<th>((L, o_1))</th>
<th>((L, o_2))</th>
<th>((H, o_1))</th>
<th>((H, o_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(\frac{\pi_2 w}{\pi_1 r_L + \pi_2 w})</td>
<td>DR</td>
<td>DR</td>
<td>(P_1)</td>
<td>DR</td>
</tr>
<tr>
<td>2</td>
<td>(\frac{(1 - \pi_2) w}{(1 - \pi_1) r_L + (1 - \pi_2) w})</td>
<td>R</td>
<td>(P_2)</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>3</td>
<td>(\frac{\pi_2 w}{\pi_1 r_L + \pi_2 w})</td>
<td>(P_3)</td>
<td>DR</td>
<td>R</td>
<td>DR</td>
</tr>
<tr>
<td>4</td>
<td>(\frac{\pi_2 w}{\pi_1 r_L + \pi_2 w})</td>
<td>(P_4)</td>
<td>DR</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>5</td>
<td>(\frac{(1 - \pi_2) w}{(1 - \pi_1) r_H + (1 - \pi_2) w})</td>
<td>DR</td>
<td>DR</td>
<td>R</td>
<td>(P_5)</td>
</tr>
<tr>
<td>6</td>
<td>(\frac{(1 - \pi_2) w}{(1 - \pi_1) r_H + (1 - \pi_2) w})</td>
<td>R</td>
<td>DR</td>
<td>R</td>
<td>(P_6)</td>
</tr>
</tbody>
</table>

Table 2: Possible Pooling Equilibria. Column’s 2-5 give the defender’s actions give its type and observation. For example, column \((L, o_1)\) gives the defender’s action when the defender is type \(L\) and observes \(o_1\). The defender’s action can be either pure—\(\text{R or DR}\)—or mixed. When the defender’s action is mixed, the probability is the probability that the defender retaliates. \(P_1 = \frac{1}{y(\pi_1 c_H - \pi_2 y)}\).

Not all of the equilibria in Table 2 are possible simultaneously. First, the parameters must be such that the defender’s randomization probabilities are between 0 and 1. In addition, the equilibria in lines 3 and 4 cannot exist simultaneously and lines 5 and 6 cannot exist simultaneously. To see why the equilibria in lines 5 and 6 cannot exist simultaneously, consider Figure 5. As with previous figures, the gray highlighted line traces the defender’s best response for each of its types. If the attacker randomizes with probability \(\beta = \frac{(1 - \pi_2) w}{(1 - \pi_1) r_H + (1 - \pi_2) w}\) then a defender of type \(H\) is indifferent between always retaliating and retaliating after \(o_1\) only. However, a defender of type \(L\) may either have a best response of either retaliating after \(o_1\) and not retaliating after \(o_2\) or never retaliate, as shown in Figure 5. With the exception of a measure zero set of parameters, the defender cannot be indifferent between never retaliating and retaliating after \(o_1\) only when the attacker randomizes with probability \(\beta = \frac{(1 - \pi_2) w}{(1 - \pi_1) r_H + (1 - \pi_2) w}\) and thus only one of the two equilibria can exist for a given value of the parameters. The same type of argument
can be used to show that only one of the equilibria in rows 5 and 6 can exist simultaneously.\footnote{Formally, when }\begin{equation} r_H > \frac{\pi_1}{\pi_2} \frac{1 - \pi_2}{1 - \pi_1}, \end{equation}\end{equation} row 4 and row 5 are possible equilibria. Otherwise, row 3 and 6 are possible equilibria. The measure zero parameter set we ignore occurs when }\begin{equation} r_H = \frac{\pi_1}{\pi_2} \frac{1 - \pi_2}{1 - \pi_1}\end{equation}
Since there always exists a babbling equilibrium in cheap talk games, at least one equilibrium in Table 2 exists. On the other hand, for some values of the parameters, there are multiple babbling equilibria. For example, when \( \pi_1 = .9, \pi_2 = .1, c_H = 4, c_L = 2, \) and \( \gamma = .4 \), the randomization probabilities in rows 1, 2, 4, and 5 are all proper probabilities and thus there are multiple equilibria. Additionally, at those parameter values, the equilibrium in which the attacker always attacks and the defender always retaliates also exists. We will continue the analysis of pooling equilibria when we discuss each equilibrium relative to the semi-separating equilibrium derived in the following section.

### 4.2.3 Semi-Separating Equilibria

Thus far, we have established that there are never separating equilibria, and depending on the parameter regime, many possible pooling equilibria and one possible pure strategy equilibrium. In this section, we establish the conditions in which there are equilibria where the defender’s signal contains some — but not perfect information — regarding its true type.

**Proposition 4 (No Semi-Separating Equilibria when \( 1 - c_L < -\nu \))**

Assume \( 1 - c_L < -\nu \). Then:

1. There is no equilibrium where the defender of type L always signals \( s_L \) and a defender of type H randomizes between signaling \( s_L \) and \( s_H \) and the attacker randomizes with probability \( \beta_L \) after receiving \( s_L \) and \( \beta_H \) after receiving \( s_H \) and \( \beta_L \neq \beta_H \).

2. There is no equilibrium where the defender of type H always signals \( s_H \) and a defender of type L randomizes between signaling \( s_L \) and \( s_H \) and the attacker randomized with probability \( \beta_L \) after receiving \( s_L \) and \( \beta_H \) after receiving \( s_H \) and \( \beta_L \neq \beta_H \).

Proposition 4.2.34 says that if an attacker of type L has relatively high ability to punish (relatively high value of \( c_L \)), then there is no equilibrium in which the defender truthfully signals when it is one type and randomizes its signal when it is another type. While this proposition, in isolation is a “negative result,” it is useful to understand the following result, established in proposition 5.

---

Proposition 5 (Semi-Separating Equilibrium with Low Punishment Power) Assume the following conditions

1. $1-c_L>−\nu$, 
2. $\pi_1c_L−\pi_2\nu>1$, 
3. $\pi_2<\pi_1r_L+\pi_2w<1$. 
4. $r_L+w>1$ 
5. $\gamma<\frac{c_L−\nu−1}{1−\gamma}<\frac{c_H+v}{c_H+v}$ 
6. $w(\pi_1r_L+\pi_2w−\pi_2)<(r_L+w−1)(\pi_1r_L+\pi_2w)$

Then there exists an equilibrium where:

- A defender of type $H$ always signals $s_H$ and always retaliates.
- A defender of type $L$ signals $s_H$ with probability $\alpha_L=\frac{\gamma}{1−\gamma}<\frac{c_H+v}{c_H+v}$. After signaling $s_H$, the defender always retaliates. After signaling $s_L$, and retaliates with probability $\varrho(o_1,s_L,L)=\frac{1}{\pi_1c_L−\pi_2\nu}$ after observing $o_1$ and defender never retaliates after observing $o_2$.
- An attacker that receives signal $s_H$ attacks with probability $\beta_H=\frac{w(\pi_1r_L+\pi_2w−\pi_2)}{(r_L+w−1)(\pi_1r_L+\pi_2w)}$.
- An attacker that receives signal $L$ attacks with probability $\beta_L=\frac{\pi_2w}{\pi_1r_L+\pi_2w}$.

To understand proposition 5, it is beneficial to understand the parameter regime in which the equilibrium exists. The first condition of proposition 5 ($1-c_L>−\nu$) says that an attacker’s best response to a defender of type $L$ that always retaliates is to always attack. Intuitively, this means that a defender of type $L$ has a relatively low capability to deliver an impactful correct retaliation. Condition 2 says that if the attacker knew with certainty that the defender were type $L$, there is an equilibrium in the induced attribution game where the attacker randomizes with probability $\frac{\pi_2w}{\pi_1r_L+\pi_2w}$. This condition together with condition 3 establishes that the defender has a relatively high
attrition capability (high values of $\pi_1$ and low values of $\pi_2$ expand the parameter region). Condition 4 says that a defender of type $L$ has enough ability to punish that as the attacker attacks more, the utility of the defender from always punishing increases. Condition 5 says that the defender of type $H$ too often. Finally, condition 6 says that there is a wide enough range of attacker randomization probabilities where the defender's best response is to always retaliate. In summary, for the signaling equilibrium to exist the defender must have relatively high attribution capability, must not possess a high retaliatory capability too often, and that the defender of type $L$ does not deliver a relatively strong punishment when it successfully retaliates.

![Figure 6: Semi-Separating Equilibrium](image)

To see how such an equilibrium exists, consider Figure 6. The two values of $\beta$ that are labeled are the attacker randomization probabilities in the signaling equilibrium. When the attacker attacks with probability $\beta_L$, the defender is indifferent between never retaliating and retaliating after $o_1$. This is where the purple line intersects the dotted blue line. When the attacker attacks with probability $\beta_H$, the defender's best response is to always retaliate. The horizontal green line illustrates that a defender of type $L$ is indifferent between these two outcomes and thus is willing to randomize its signal when it is type $L$. A defender of type $H$ receives a higher utility when the attacker attacks with probability $\beta_H$ and always retaliates (solid red line) than when the attacker attacks with probability $\beta_L$ and the attacker retaliates after $o_1$ only (dotted blue line). Therefore, the defender of type $H$ would always signal $\beta_H$ as indicated in the semi-separating equilibrium.
4.2.4 Gains from Signaling

Finally, we investigate whether it is possible for the defender to gain from signaling. To carry out such an analysis, we say that there is a gain from signaling if for a fixed set of parameters, the signaling equilibrium exists and the defender’s expected utility in the signaling equilibrium is higher than it’s expected utility in a pooling equilibrium that exists simultaneously. Of course, this ignores any notion of equilibrium selection and how the players might arise at such an equilibrium, which is an interesting future endeavor. We also say that it is possible for the defender to gain with signaling through a *deterrence effect* if the defender’s expected utility under the signaling equilibrium is higher than in the pooling equilibrium and the attacker’s attack probability is lower in the signaling equilibrium than in the pooling equilibrium.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_1$ (probability of signal $o_1$, attribution to attacker)</td>
<td>.8</td>
</tr>
<tr>
<td>$\pi_2$ (probability of signal $o_2$, no attribution to attacker)</td>
<td>.45</td>
</tr>
<tr>
<td>$\epsilon_h$ (attacker cost from correct retaliation with high capability)</td>
<td>4</td>
</tr>
<tr>
<td>$\epsilon_L$ (attacker cost from correct retaliation with low capability)</td>
<td>3</td>
</tr>
<tr>
<td>$\nu$ (attacker cost from incorrect retaliation)</td>
<td>2.6</td>
</tr>
<tr>
<td>$\gamma$ (probability defender has high capability)</td>
<td>.4</td>
</tr>
<tr>
<td>$\rho_h$ (defender reward from correct retaliation with high capability)</td>
<td>.9</td>
</tr>
<tr>
<td>$\rho_L$ (defender reward from correct retaliation with low capability)</td>
<td>.65</td>
</tr>
<tr>
<td>$w$ (defender cost from incorrect retaliation)</td>
<td>.8</td>
</tr>
</tbody>
</table>

Table 3: Parameter values where there are gains from signaling through deterrence effect.
First, we illustrate that there exist parameter regimes where the defender can gain by deterring the attacker. Consider the parameter values in Table 3. In this parameter regime, the conditions in proposition 5 are satisfied and the signaling equilibrium exists. At this equilibrium, the attacker attacks with probability 0.715 and the defender’s expected utility before realizing its type is −0.298. At these parameter values, the equilibrium in row 2 of Table 2 also exists. At this equilibrium, the attacker attacks with probability 0.772 and the defender earns an expected utility of −0.36.

Figure 7 illustrates the deterrent effect of the signaling equilibrium. In the pooling equilibrium, the attacker randomizes with probability $\beta_p = \frac{(1-\pi_2)w}{(1-\pi_1)r_L + (1-\pi_2)w'}$. In the signaling equilibrium, the attacker will randomize either at probability $\beta_H$ or $\beta_L$. While $\beta_H$ is slightly higher than $\beta_p$, $\beta_L$ is sufficiently low such that on average, the attacker

---

8 The intuition that we demonstrate in this section holds for a set of parameters of positive measure. We only select specific parameter values to highlight the main insights.
attacks less and the defender’s expected utility increases by reducing the probability in which the attacker attacks. Since the defender of type $L$ has a higher expected utility at $\beta_H$ than at $\beta_P$, the defender gains with signaling through a deterrence effect.

After establishing that the defender can benefit through signaling by deterring an attacker, we now present our final result that shows that a defender can benefit through signaling by inducing the attacker to attack more. Consider the parameter set in Table 4. In this parameter regime, the signaling equilibrium exists and the attacker attacks with probability 0.729 and the defender earns an expected payoff of −0.249. At those parameter values, the equilibrium in row 1 of Table 2 also exists. This equilibrium is the pooling equilibrium in which the attacker randomizes its attack with the lowest probability. At that pooling equilibrium, the attacker attacks with probability 0.260 and the defender’s expected payoff is −0.260. This means that the defender can increase its expected utility from signaling by inducing the attacker to attack more.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_1$ (probability of signal $o_1$, attribution to attacker)</td>
<td>.95</td>
</tr>
<tr>
<td>$\pi_2$ (probability of signal $o_2$, no attribution to attacker)</td>
<td>.5</td>
</tr>
<tr>
<td>$c_H$ (attacker cost from correct retaliation with high capability)</td>
<td>5</td>
</tr>
<tr>
<td>$c_L$ (attacker cost from correct retaliation with low capability)</td>
<td>3</td>
</tr>
<tr>
<td>$v$ (attacker cost from incorrect retaliation)</td>
<td>3</td>
</tr>
<tr>
<td>$\gamma$ (probability defender has high capability)</td>
<td>.32</td>
</tr>
<tr>
<td>$r_H$ (defender reward from correct retaliation with high capability)</td>
<td>.9</td>
</tr>
<tr>
<td>$r_L$ (defender reward from correct retaliation with low capability)</td>
<td>.7</td>
</tr>
<tr>
<td>$w$ (defender cost from incorrect retaliation)</td>
<td>.6</td>
</tr>
</tbody>
</table>

Table 4: Parameter values where there are gains from signaling through deterrence

There are two reasons for this counter-intuitive result. The first reason has to do with the trade-off the defender faces between an incorrect retaliation and an undetected attack. If the cost to an incorrect retaliation is relatively high, then the defender has little incentive to retaliate because it risks the possibility of being incorrect. However, if the attacker were to attack with a higher probability, then the defender would incorrectly retaliate less often. If the cost of an incorrect retaliation is high enough, then the defender would benefit from being attacked slightly more but incorrectly retaliating less often.
The second effect is due to the defender inducing the attacker to attack against a defender of type $H$ where the defender gains the most from a correct retaliation. Consider Figure 8. The Figure illustrates the pooling equilibrium at $\beta_p$ and the semi-separating equilibrium where the attacker randomizes either at $\beta_L$ and $\beta_H$ depending on the signal it gets from the defender. The defender of type $L$’s expected utility when the attacker attack with probability $\beta_p$ is higher than if the attacker were to attack with probability $\beta_L$ or $\beta_H$ indicating that a defender of type $L$ is worse off when the attacker attacks more. However, the expected utility of an attacker of type $H$ is higher at $\beta_H$ than at $\beta_p$. Therefore, if the defender can randomize its signal such that the benefit it receives from retaliating when it is type $H$ outweighs the loss it would incur from a higher attack probability when the defender is type $L$, then the defender stands to gain from a higher attack probability. In other words, the defender can gain when the attacker attacks more as long as the increased attack probability mostly occurs when the defender is type $H$ and can more effectively retaliate.

As a final illustration of the equilibrium and its properties, Figure 9. shows the parameter region where there are gains to be made from signaling by inducing the
attacker to attack more. This region is defined by the conditions in proposition 5 and the region where \( \frac{1}{\gamma (\pi_1 c_H - \pi_2 v)} \) is a proper probability. Generally speaking, the lower right corner of Figure 9 is where attribution capabilities are the highest (high value of \( \pi_1 \) and low value of \( \pi_2 \)). The figure shows that as the cost of an incorrect retaliation increases, the defender’s attribution ability must increase in order to support an equilibrium where there are gains from signaling though an anti-deterrence effect.

Figure 9: Region in \( \pi_1, \pi_2 \) space where the signaling equilibrium exists and the defender can gain from signaling by inducing the attacker to attack more. The parameters other than \( \pi_1, \pi_2 \) and \( w \) are as in Table 4.
5 Findings

By expanding upon the attribution game through the inclusion of signaling, our results contribute new insights on the open question of cyber deterrence. Fundamentally, the signaling game yields four important findings. We briefly summarize these findings here. First, it is never in the best interest of the defender to signal truthfully. This is because if the defender could convince the attacker that it is truthfully signaling, the defender would no longer have an incentive to truthfully signal. This implies that it is virtually impossible to deter an attack by signaling a strong capability. This means that policymakers should prepare for a scenario in which they must deliver a powerful retaliation, even though they warned an adversary that they have the ability to retaliate effectively.

Second, since there are different qualitative equilibria for different parameter values (i.e., the results of the game theoretic model can qualitatively change as inputs to the model change), policy makers must maintain a flexible cyber policy to adapt for a changing world. This means that a small change in the parameters (the defender’s attribution capabilities, for example) may trigger a large change in the attacker behavior and thus necessitate a significant revision in the defender’s strategy. To account for this abrupt shift in equilibrium strategies, an effective cyber policy must be flexible and amenable to change.

Third, enhancing the strength of attribution may be the powerful deterrence tool. Our results underscore the importance of this finding which has been argued many others before us. Confident attribution leads to gains for the defender in both deterring the attacker and in enabling retaliation. The first point is clear; in a world of perfect attribution the attacker will never attack. For the second, our findings convey that enhancing attribution would reduce the likelihood that a defender retaliates incorrectly thereby encouraging the use of retaliation. Our final result follows this point into new directions.

Fourth, our findings point to the interesting and thought-provoking value of anti-deterrence and suggest that policy makers should consider whether deterrence is the only option. In a world without complete information, complete attribution, and complete deterrence, our results suggest that anti-deterrence may be a means of increasing the defender’s well-being while also welcoming a higher probability of attack. We show that in some cases, a defender would be willing to take on a higher probability of being attacked if 1) the higher probability of attack reduces the
probability (and therefore costs) of an incorrect retaliation and 2) the increase in the probability of attack is more heavily weighted to when the defender is most able to respond to attack and not when it is cannot effectively respond. Our results therefore elucidate a potential strategy for policy makers to consider when crafting a cyber deterrence policy; when attribution is uncertain and the cost of incorrect retaliation is high but the defender is able to respond to an attack by a given attacker with strength, use signaling to induce more attacks.

6 Conclusion — Towards a Cyber Deterrence Policy

The challenges emanating from the cyber domain on conflict are a marked departure from the challenges of the nuclear domain. Thus, while the fundamental building blocks of nuclear deterrence outlined by Thomas C Schelling (1960) may apply in the current cyber domain, other aspects of nuclear deterrence theory may not. We offer a game-theoretic model with two players, imperfect information, and signaling to discuss the strategy of deterrence in cyberspace. In addition to highlighting the significant limitations to cyber deterrence, our findings reveal unique opportunities and important considerations for the development of a cyber deterrence policy. Finally, this work identifies numerous areas that warrant additional research.

As eluded to in the previous section, deterrence likely will not come easy for the foreseeable future. In a world with imperfect attribution, we find no equilibria in which the attacker will never attack, and therefore find deterrence of all attacks from all attackers unlikely. However, this does not lead us to the conclusion that deterrence is infeasible or impossible. On the contrary, improvements in attribution and signaling capabilities may deter some attacks from some attackers. When we introduce signaling into the attribution game, there are equilibria in which the attacker attacks with a lower probability than in a game without signaling. The result is increased benefits received by the defender from the reduced attacks. While signaling does not bring the attack probability to zero (i.e. does not deter the attacker from all attacks), signaling helps to reduce the chances of an attack.

Our results also convey several key points for consideration of policy makers in the development of cyber deterrence policy. The reality that it is better to retaliate with strength than force underpins many of these findings. For one, increasing the cost of a correct retaliation relative to the cost of an incorrect retaliation improves
deterrence. This may mean choosing mechanism of retaliation (i.e., cover cyber, overt cyber, kinetic force) based on the strength of attribution. Enhancing the strength of attribution is potentially the strongest tool in achieving cyber deterrence; perfect attribution would be necessary (although maybe not sufficient) to deter all attacks by all attackers. However, in the world we currently occupy where perfect attribution does not exist and incorrect retaliation can come with a cost to the defender, we find a new result which we call anti-deterrence. In this world, the anti-deterrence result finds that it may be in the best interest to use signaling to actually induce some attackers to attack thereby decreasing the likelihood of incorrect retaliation and allowing for stronger retaliation. The curious value of anti-deterrence raises the question is deterrence the only option?

Capturing these findings in policy in a way that establishes tactics and procedures for the effective execution of cyber deterrence policy poses several challenges. Our two-player game elucidates the importance of signaling but does not capture the interpretations or responses of multiple actors operating in a cyber environment to signaling attempts. In order to incorporate signaling into a cyber deterrence policy, the U.S. will need to consider more specific questions; how does the U.S. signal (through what mechanism)? When does the U.S. signal? How does the U.S. signal credibly? Do different cyber actors require different signals (is the signal the same for North Korea as it is for Iran?) Additional work on this topic would answer key questions about signaling in cyberspace and help policymakers integrate signaling into existing cyber policy. Furthermore, while there is some existing literature on the role of signaling in foreign policy, little is focused on signaling in cyberspace and even less that quantifies the benefits the defender receives when it signals properly. Additional research could construct a cyber policy framework for signaling that would outline when and in what ways to signal a capability or willingness to retaliate that afford the defender the most benefit and greatest probability of deterring a potential attacker.

Future exploration of this subject could elucidate answers to important questions: would the cost of retaliating – at times correctly and at times incorrectly – outweigh the cost of being attacked and not responding? Would the use of both correctly and incorrectly attributed retaliatory capabilities lead to more attacks or less? How does the value of anti-deterrence change when multiple actors are introduced into the game?

Expansions of this game accompanied by additional research could identify a coherent strategy for deterring adversary actions and reducing costs in cyberspace as well as a framework for its implementation. Another factor not considered in this
game but that certainly warrants additional research is the role of multiple actors in cyberspace and potentially the need for unique deterrence strategies designed to deter specific cyber actors. For example, would a cyber deterrence strategy have the same impact on China and North Korea? The results of our two-player game suggest not, confirming the suspicions of others that there is no "one size fits all" cyber deterrence strategy. This also echoes debates in the nuclear realm as to whether rational and irrational (e.g., some non-state actors such as extremist groups) would be evenly deterred. The motivations, capabilities, and appetites for risk vary among actors operating in cyberspace and should be considered when developing a cyber deterrence strategy. Future extensions of the signaling game including multiple attackers with unique characteristics could reveal valuable insights regarding the costs and benefits that the defender receives.

Outstanding questions include: what can be deterred? Additional research could examine the types of adversaries, types of attacks and types of capabilities that are actually deterrable. From a policy and legal perspective, how much attribution proof is needed to justify a response? What is the threshold for a cyberattack that warrants a response (either a signaled response or actual retaliation?) Additional work could also be conducted to examine past cases where signaling has been used and past cases where retaliation has been used and not used.

The conversation on cyber deterrence is ripe for further debate. Additional work could enhance our findings by endogenizing the defender's capability in which the defender undertakes cost investment to improve its detection and attribution abilities. Furthermore, while we introduce signaling to the discussion on cyber deterrence, additional work should be done to discuss the single use nature of cyber weapons. In the end, our findings on cyber deterrence do not lead us far astray from Thomas C Schelling (1960). Increasing the credible threat of punishment either for deterrence or anti-deterrence is in the best interest of the defender. In a world of cyber weapons, however, the strongest weapon is likely attribution.
References


A Appendix

- **Proposition 1** To prove proposition 1, we begin with two lemmas that establish there are no pure strategy equilibria and then prove the proposition by looking for mixed strategy equilibria.

**Lemma 1** (No Pure Strategy Equilibrium for the Attacker in the Attribution Game). If $1 - c < -v$ there is no equilibrium in the attribution game where the attacker plays a pure strategy.

**Proof.** Suppose the attacker’s pure strategy is to never attack. The defender’s best response against such a strategy is to never retaliate. However, the attacker’s best response to the defender never retaliating is to always attack. Therefore, there is no pure strategy equilibrium where the attacker never attacks. Now, suppose the attacker’s pure strategy is to always attack. In this case, the defender’s best response is to always retaliate (since by assumption, the total payoff to attacking and being correctly retaliated against, $1 - c$, is less than the total payoff of being incorrectly retaliated against $-v$). Therefore, there is no pure strategy equilibrium where the attacker always attacks. \[\square\]

**Lemma 2** (No Pure Strategy Equilibrium for the Defender in the Attribution Game). If $1 - c < -v$ there is no equilibrium in the attribution game where the defender plays a pure strategy.

**Proof.** Suppose the defender’s pure strategy is to always retaliate. Then the attacker’s best response would be to never attack and thus the defender’s best response would be to never retaliate. Therefore, always retaliating cannot be part of an equilibrium. A parallel argument shows that never retaliating cannot be part of equilibrium. The only other pure strategy for the defender in the attribution game is to always retaliate after receiving signal $o_1$ and never retaliate after signal $o_2$ (since $\pi_1 > \pi_2$, the strategy always retaliate after $o_2$ and never retaliate after $o_1$ is strictly dominated). If the defender adopts this strategy, the attacker’s expected utility from attacking and not attacking is given by:

$$U_a(A, (R, DR)) = Pr(o_1|A)(1 - c) + Pr(o_2|A) = \pi_1(1 - c) + (1 - \pi_1)$$
$$U_a(NA, (R, DR)) = Pr(o_1|NA)(-v) + Pr(o_2|NA) \times 0 = \pi_2 v$$

where $U_a(X, (Y, Z))$ is the expected utility of the attacker for choosing action $X$ where the defender chooses (the probability of) action $Y$ after observing $o_1$ and (the probability of action) $Z$ after observing $o_2$. These expected utilities implies that if $1 - \pi_1 c - \pi_2 v > 0$, then the attacker would always attack and if $1 - \pi_1 c - \pi_2 v < 0$ then attacker would never attack both of which by lemma 1 cannot be part of an equilibrium (Recall that by assumption $1 - \pi_1 c - \pi_2 v \neq 0$). \[\square\]

Lemmas 1 and 2 establish that there cannot be an equilibrium in which either the attacker or the defender play pure strategies. Therefore, we prove proposition 1 by searching for mixed strategy equilibria only.

**Proof of Proposition 1.** Suppose the attacker randomizes with probability $\beta$. Then equilibrium defender beliefs are determined as follows:

$$Pr(A|o_1) = \frac{Pr(o_1|A)Pr(A)}{Pr(o_1)} = \frac{\pi_1 \beta}{\pi_1 \beta + \pi_2 (1 - \beta)}$$
$$Pr(DA|o_1) = \frac{Pr(o_1|DA)Pr(DA)}{Pr(o_1)} = \frac{\pi_2 (1 - \beta)}{\pi_1 \beta + \pi_2 (1 - \beta)}$$
$$Pr(A|o_2) = \frac{Pr(o_2|A)Pr(A)}{Pr(o_2)} = \frac{(1 - \pi_1) \beta}{(1 - \pi_1) \beta + (1 - \pi_2)(1 - \beta)}$$
$$Pr(DA|o_2) = \frac{Pr(o_2|DA)Pr(DA)}{Pr(o_2)} = \frac{(1 - \pi_2) (1 - \beta)}{(1 - \pi_1) \beta + (1 - \pi_2)(1 - \beta)}$$

38
With these probabilities, it is possible to write the defender’s expected utility from retaliating and not retaliating for each of its observations. They are given by:

\[ U_d(R, \beta; o_1) = Pr(A|o_1)(r - 1) - Pr(DA|o_1)w \]
\[ = \frac{\pi_1\beta}{\pi_1\beta + \pi_2(1 - \beta)} (r - 1) - \frac{\pi_2(1 - \beta)}{\pi_1\beta + \pi_2(1 - \beta)}w \]

\[ U_d(DR, \beta; o_1) = -Pr(A|o_1) - \pi \times Pr(DA|o_1) \]
\[ = -\frac{\pi_1\beta}{\pi_1\beta + \pi_2(1 - \beta)} \]
\[ U_d(R, \beta; o_2) = Pr(A|o_2)(r - 1) - Pr(DA|o_2)w \]
\[ = \frac{(1 - \pi_1)\beta}{(1 - \pi_1)\beta + (1 - \pi_2)(1 - \beta)} (r - 1) - \frac{(1 - \pi_2)(1 - \beta)}{(1 - \pi_1)\beta + (1 - \pi_2)(1 - \beta)}w \]

\[ U_d(DR, \beta; o_2) = -Pr(A|o_2) - \pi \times Pr(DA|o_2) \]
\[ = -\frac{(1 - \pi_1)\beta}{(1 - \pi_1)\beta + (1 - \pi_2)(1 - \beta)} \]

where \( U_d(X, \beta; o) \) is the expected utility of the defender by choosing action \( X \) given the attacker randomizes with probability \( \beta \) and the defender observed observation \( o \). Since there is no equilibrium where the defender plays a pure strategy and must randomize, it must be indifferent at (at least) one of its information sets.

After observing \( o_1 \) the defender is indifferent between retaliating and not retaliating when:

\[ \frac{\pi_1\beta}{\pi_2(1 - \beta)} = \frac{w}{r} \]
\[ \implies \beta = \frac{\pi_2w}{\pi_1r + \pi_2w} \]  \hspace{1cm} (1)

where it is optimal for the defender to retaliate if \( \beta \) is greater than the right hand side (RHS) of equation 1 and not retaliate if \( \beta \) is less than the RHS.

After observing \( o_2 \), the defender is indifferent between retaliating and not retaliating when

\[ \frac{(1 - \pi_1)\beta}{(1 - \pi_2)(1 - \beta)} = \frac{w}{r} \]
\[ \implies \beta = \frac{(1 - \pi_2)w}{\pi_1r + (1 - \pi_2)w} \]  \hspace{1cm} (2)

where it is optimal for the defender to retaliate if \( \beta \) is greater than the right hand side (RHS) of equation 2 and not retaliate if \( \beta \) is less than the RHS. Since \( \pi_1 > \pi_2 \) by assumption, the RHS of 1 is less than the RHS of 2.

Since the defender must be indifferent at at least one of its information sets, equations 1 and 2 give the only two possible values of equilibrium attacker randomization probability. We will now establish the sufficient conditions for the values in equations 1 and 2 to be part of a PBE.

**Case 1:** \( \beta = \frac{\pi_2w}{\pi_1r + \pi_2w} \). Suppose the attacker randomizes with probability \( \beta = \frac{\pi_2w}{\pi_1r + \pi_2w} \), then the defender will never retaliate after receiving \( o_2 \) and is indifferent after observing \( o_1 \), and therefore is willing to randomize with probability \( \rho \) after observing \( o_1 \). The necessary and sufficient condition for the attacker to be willing to randomize is if its expected utility from attacking is equal to its expected utility from not attacking. This is satisfied when:

\[ U_d(A, (\rho, DR)) = U_d(NA, (\rho, DR)) \]
\[ \implies Pr(o_1|A)Pr(R|o_1)(1 - c) + Pr(o_1|A)Pr(NR|o_1) + Pr(o_2|A) = Pr(o_1|A)Pr(R|o_1)(-v) \]
\[ \implies \pi_1\rho(1 - c) + \pi_1(1 - \rho) + 1 - \pi_1 = -\pi_2\rho \]
\[ \implies \rho = \frac{1}{\pi_1c - \pi_2v}. \]  \hspace{1cm} (3)
Since \( \rho \) is a probability, it only takes on the values between 0 and 1, which holds only when \( 1 - \pi_1 c + \pi_2 v \leq 0 \). Therefore, if \( 1 - \pi_1 c + \pi_2 v \leq 0 \), then there is a Nash equilibrium where the attacker randomizes with probability \( \beta = \frac{\pi_{2} u}{\pi_{1} c - \pi_{2} v} = \beta^{*}_{1} \) and the defender never retaliates after observing \( o_2 \) and randomizes with probability \( \rho = \frac{\pi_{1} c - \pi_{2} v}{\pi_{1} c - \pi_{2} v} = \rho^{*}_{1} \) after observing \( o_1 \). This proves item 1 of the proposition.

**Case 2**: \( \beta = \frac{(1 - \pi_2) w}{(1 - \pi_1) r + (1 - \pi_2) w} \) Suppose the attacker randomizes with probability \( \beta = \frac{(1 - \pi_2) w}{(1 - \pi_1) r + (1 - \pi_2) w} \), then the defender will always retaliate after receiving \( o_1 \) and is willing to randomize with probability \( \rho \) after observing \( o_2 \). The necessary and sufficient condition for the attacker to be willing to randomize is if its expected utility from attacking is equal to its expected utility from not attacking. This is satisfied when:

\[
U_{o}(A, (R, \rho)) = U_{o}(NA, (R, \rho)) \implies Pr(o_1|A)(1 - c) + Pr(o_2|A)\rho(1 - c) + Pr(o_2|A)(1 - \rho) = Pr(o_1|NA)(-v) + Pr(o_2|NA)\rho(-v)
\]

\[
\implies \pi_{1}(1 - c) + (1 - \pi_1)\rho(1 - c) + (1 - \pi_1)(1 - \rho) = -v(\pi_2 + (1 - \pi_2)\rho)
\]

\[
\implies \rho = \frac{1 - \pi_1 c + \pi_2 v}{(1 - \pi_1)c - (1 - \pi_2)v}.
\] (4)

Since by assumption \( c - v > 1 \) the numerator is greater than the denominator and thus the only way that \( \rho \) represents a proper probability is if \( 1 - \pi_1 c + \pi_2 v > 0 \). Therefore, if \( 1 - \pi_1 c + \pi_2 v > 0 \), then there is a Nash equilibrium where the attacker randomizes with probability \( \beta = \frac{(1 - \pi_2) w}{(1 - \pi_1) r + (1 - \pi_2) w} = \beta^{*}_{2} \) and the defender always retaliates after observing \( o_1 \) and randomizes with probability \( \rho = \frac{(1 - \pi_1) c + \pi_2 v}{(1 - \pi_1)c - (1 - \pi_2)v} = \rho^{*}_{2} \) after observing \( o_2 \). This proves item 2 of the proposition.

**Proof of Corollary 1.** As shown in the proof of proposition 1, there are only two possible values of \( \beta \) that can be part of a Nash equilibrium. For each value of \( \beta \), there is only one mixed strategy for the defender that would make the attacker indifferent and thus willing to randomize. This suggests that there may be two equilibria. However, under one value of \( \beta \) the existence of a defender’s equilibrium mixed strategy relies on \( 1 - \pi_1 c + \pi_2 v > 0 \) where for the other value of \( \beta \), the existence of the defender’s mixed strategy relies on \( 1 - \pi_1 c + \pi_2 v > 0 \). Since both conditions can not hold simultaneously, there is a unique Nash equilibrium determined by the sign of \( 1 - \pi_1 c + \pi_2 v \).

**Proof of Proposition 2.** If the defender truthfully signals its capability, then Bayes rule dictates that the attacker assigns probability 1 to the defender’s true capability and 0 otherwise. Therefore, after the truthful signal, a separating equilibrium would have the players play the equilibrium profile of the attribution game where the parameters are determined by the defender’s true type and the payoffs would be as in table 1 where the values of \( r \) and \( c \) are given according to the defender’s type. This proof shows that for for all possible values of \( 1 - \pi_1 c + \pi_2 v \) where \( k \in \{H, L\} \), the defender can improve its expected utility by lying to the attacker about its type.

Formally, let \( \beta^{*}_{L} \) be the attacker’s equilibrium probability of attacking when the players play the attribution game and the defender is type \( L \) and let \( \beta^{*}_{H} \) be the attacker’s equilibrium probability of attacking when they play the attribution game and the defender is type \( H \).

- **Case 1**: \( 1 - \pi_1 c_H + \pi_2 v < 0 \) and \( 1 - \pi_1 c_L + \pi_2 v < 0 \). Consider when the defender truthfully signals \( s_L \), indicating that it has a low capability. In this case, the attacker’s equilibrium probability in the attribution game is \( \beta^{*}_{L} = \frac{\pi_{2} w}{\pi_{1} r_{H} + \pi_{2} w} \) and the defender’s expected utility is \( -\beta^{*}_{L} \). If instead of signaling \( s_L \) the defender signaled \( s_H \), the attacker would randomize with probability \( \beta^{*}_{H} = \frac{\pi_{2} w}{\pi_{1} r_{L} + \pi_{2} w} \) and the defender’s payoff would be \( -\beta^{*}_{H} > -\beta^{*}_{L} \). Therefore the defender has an incentive to signal it is type \( H \) when it is truly type \( L \) and thus there is not a separating equilibrium when \( 0 < 1 - \pi_1 c_H + \pi_2 v < 0 \) and \( 0 < 1 - \pi_1 c_L + \pi_2 v < 0 \).

- **Case 2**: \( 1 - \pi_1 c_H + \pi_2 v > 0 \) and \( 1 - \pi_1 c_L + \pi_2 v > 0 \). In this regime, if the defender were to signal its true capability, the attacker would attack with probability \( \beta^{*}_{L} = \frac{\pi_{2} w}{(1 - \pi_1) r_{L} + (1 - \pi_2) w} \) where \( k \) is either \( H \) or \( L \) depending on the defenders signal. The defender of type \( k \) has an expected
utility of $\beta^*_L(r - 1) - (1 - \beta^*_H)w = \beta^*_L(r_L - 1 + w) - w$. If $(r_H - 1 + w) > 0$, then the defender’s utility is increasing in the attack probability. Therefore when the defender is type $H$ it would prefer that the attacker attack with a higher probability thus would signal that it is type $L$ and induce the attacker to attack with probability $\beta^*_L > \beta^*_H$. Therefore, there cannot be a separating equilibrium if $(r_H - 1 + w) > 0$. Now suppose $(r_H - 1 + w) \leq 0$. Then it must be the case that $(r_L - 1 + w) < 0$, which implies that when the defender is type $L$, its expected utility is decreasing in the attack probability. This means that when the defender is truly type $L$ it would prefer to signal that it was type $H$ so that the attacker randomizes with probability $\beta^*_H < \beta^*_L$. As a result, there cannot be a separating equilibrium when $(r_H - 1 + w) \leq 0$.

- **Case 3:** $1 - \pi_1 c_H + \pi_2 v < 0$ and $1 - \pi_1 c_L + \pi_2 v > 0$ In this regime, if the defender is type $H$ and signals such, the attacker randomizes with probability $\beta^*_H = \frac{\pi_2 w}{\pi_1 c_H + \pi_2 v}$ and the defender’s expected utility when it is type $H$ is $-\beta^*_H$. If the defender is type $L$ and it signals such, the attacker randomizes with probability $\beta^*_L = \frac{\pi_1 c_L + (1 - \pi_2)w}{\pi_1 c_L + \pi_2 w}$ and the defender earns an expected utility of $\beta^*_L \times (\pi_1 c_L + \pi_2 w - 1) - \pi_2 w$ (which by indifference is the same as $\beta^*_L (r_L - 1 + w) - w$). Since $\pi_1 > \pi_2$ and $r_H > r_L$, it can be shown that $\beta^*_L > \beta^*_H$. Consider the defender’s deviation of signaling that it is type $L$ when it is actually type $H$ and changing its strategy to retaliating after $o_1$ and not retaliating after $o_2$. In this case, the defender’s utility is given by:

$$U_d((R, DR), \beta^*_L) = \pi_1 c_L (r_H - 1) - (1 - \pi_1)\beta^*_L (r_L - 1 + w) - (1 - \beta^*_L)\pi_2 v$$

where $U_d((A, B), C)$ is the defender’s expected utility from playing $A$ after $o_1$ and $B$ after $o_2$, when the attacker randomizes with probability $C$. For this deviation not to be profitable for the defender it must be that:

$$
\begin{align*}
U_d((R, DR), \beta^*_H) &\geq U_d((R, DR), \beta^*_L) \\
-\beta^*_H &\geq \pi_1 c_L (r_H - 1) - (1 - \pi_1)\beta^*_L (r_L - 1 + w) - (1 - \beta^*_L)\pi_2 v \\
-\beta^*_H &\geq \beta^*_L \pi_1 c_H - \beta^*_L - \pi_2 w + \beta^*_L \pi_2 w \\
-\beta^*_H &\geq \beta^*_L \pi_1 c_H - \beta^*_L - \beta^*_H (\pi_1 c_H + \pi_2 w) + \beta^*_L \pi_2 w \\
\beta^*_H (\pi_2 w + \pi_1 c_H - 1) &\geq \beta^*_L (\pi_2 w + \pi_1 c_H - 1).
\end{align*}
$$

Since $\beta^*_H < \beta^*_L$ the inequality in equation 6 only holds if $\pi_2 w + \pi_1 c_H - 1 \leq 0$. So if $\pi_2 w + \pi_1 c_H - 1 > 0$, then the deviation is profitable and there is no incentive for the defender to truthful signal its type. What remains to be shown is that the defender does not have an incentive to truthfully signal its type when $\pi_2 w + \pi_1 c_H - 1 > 0$. To do this, consider the defender’s deviation of signaling that it is type $L$ when it is actually type $H$ and changing its strategy after $o_1$ and not retaliating after $o_2$. Under this deviation, the defender’s expected utility is

$$U_d((R, DR), \beta^*_H) = \beta^*_H (\pi_1 c_L + \pi_2 w - 1) - \pi_2 w$$

For this deviation to not be profitable for the defender it must be that:

$$U_d((R, DR), \beta^*_H) \geq U_d((R, DR), \beta^*_L)$$

$$\beta^*_H (\pi_1 c_L - 1 + \pi_2 w) - \pi_2 w \geq \beta^*_H (\pi_1 c_L - 1 + \pi_2 w) - \pi_2 w.$$  

Since $\beta^*_L > \beta^*_H$, the only way for the inequality in equation 8 to hold is if $\pi_1 c_L - 1 + \pi_2 w \leq 0$. However, if $\pi_1 c_L - 1 + \pi_2 w \geq 0$, then the defender would have an incentive to deviate when it is type $L$. Therefore, when $\pi_1 c_H - 1 + \pi_2 w \geq 0$, the defender has an incentive to deviate from truthful signaling and when $\pi_1 c_H - 1 + \pi_2 w \leq 0$, the defender has an incentive to deviate from truthful signaling. Ignoring the measure 0 case where $\pi_1 c_H - 1 + \pi_2 w = 0$, the defender always has an incentive to deviate from truthful signaling.

All three cases cover all possible parameter values and illustrate that for all values of the parameters, there is always a profitable deviation from truthful signaling for the defender and thus there is no separating equilibrium.
• Proof of Proposition 3. In this case, if the attacker knows the defender is type $L$, then it is always a best response for the attacker to attack. As a result, it is always the defender’s best response to retaliate when it truthfully signals it is type $L$. To show this cannot be an equilibrium, consider the following two cases:

- **Case 1**: Suppose $\pi_1 r_H + \pi_2 w - 1 > 0$. Under a separating equilibrium, when the defender signals it is type $L$, the attacker attacks with probability $1$. Also, when the defender is type $H$ and truthfully signals its type, its expected utility is $-\beta_H$ when $1 - \pi_1 c_H + \pi_2 v < 0$ and $\beta_H (r_H + w - 1) - w$ when $1 - \pi_1 c_H + \pi_2 v > 0$. Consider each of the two cases separately:
  * Suppose $1 - \pi_1 c_H + \pi_2 v > 0$. Since by assumption $\pi_1 r_H + \pi_2 w - 1 > 0$, then $r_H + w - 1 > 0$ and thus the attacker’s expected utility is increasing in $\beta_H$ when $1 - \pi_1 c_H + \pi_2 v > 0$. Therefore, if $1 - \pi_1 c_H + \pi_2 v > 0$ and the defender is type $H$, it would be better off signaling it is type $L$ and thus there cannot be a separating equilibrium in which it truthfully signals its type.
  * Suppose $1 - \pi_1 c_H + \pi_2 v < 0$. Then when the defender is type $H$ and truthfully signals its type, its expected utility is $-\beta_H = \frac{-\pi_2 w}{\pi_1 r_H - \pi_2 w}$. If it were to instead switch its strategy by signaling that it is type $L$ and always retaliating, its expected utility is $r_H - 1$. For the defender to not have an incentive to make this switch it must be that:

$$\frac{-\pi_2 w}{\pi_1 r_H - \pi_2 w} \geq r_H - 1$$

$$\rightarrow 0 \geq r_H (\pi_1 r_H + \pi_2 w - \pi_1)$$

(9)

However by assumption, $\pi_1 r_H + \pi_2 w - 1 > 0$ so it is impossible for the inequality in equation 9 to hold and this there cannot be a separating equilibrium.

- **Case 2**: Suppose $\pi_1 r_H + \pi_2 w - 1 < 0$. Again, there are two sub-cases:
  * Suppose $r_L + w - 1 < 0$. The defender’s expected utility by truthfully signaling when it is type $L$ is $r_L - 1$. If instead it signaled it was type $H$ and always retaliated, its expected utility would be $\beta_H^*(r_L + w - 1) - w$. For it to not gain anything from such a deviation it must be:

$$r_L - 1 \geq \beta_H^*(r_L + w - 1) - w$$

$$\rightarrow 1 \leq \beta_H^*$$

(10)

Since $\beta_H^*$ is a probability less than 1, the inequality in equation 10 cannot hold and therefore there cannot be a separating equilibrium.

  * Suppose $r_L + w - 1 > 0$. If the defender signals that it is type $L$ when it is type $H$ and always retaliates, it will earn $r_H - 1$. If it signals it is type $L$ when it is truly type $L$, it earns $r_L - 1$. If $1 - \pi_1 c_H + \pi_2 v < 0$, then when the defender signals it is type $H$, the attacker attacks with probability $-\beta_H^* = \frac{-\pi_2 w}{\pi_1 r_H - \pi_2 w}$. Two possible deviations the defender can make is 1) when $L$ signal that it is type $H$ and never retaliate and 2) when $H$ signal that it is type $L$ and always retaliate. For neither of these deviations to be profitable it must be:

$$r_H - 1 < \beta_H^*$$

$$r_L - 1 > \beta_H^*$$

Since $r_H > r_L$, it is impossible for both conditions to hold simultaneously. Finally, If $1 - \pi_1 c_H + \pi_2 v > 0$, then when the defender signals it is type $H$, the attacker attacks with probability $-\beta_H^* = \frac{-\pi_2 w}{\pi_1 r_H - \pi_2 w}$ and the defender earns an expected utility of $\beta_H^* (\pi_1 r_H + \pi_2 v - 1) - w$. If instead, the defender signaled it was type $L$ when it is type $H$, and always retaliate it would earn $r_H - 1$. For there to be no incentive for the defender to deviate it must be that:

$$r_H - 1 < \beta_H^* (\pi_1 r_H + \pi_2 v - 1) - w$$

$$\rightarrow \frac{r_H - 1 + w}{\pi_1 r_H + \pi_2 v - 1} > \beta_H^*$$

(11)

Since by assumption $r_H - 1 + w > 0 > \pi_1 r_H + \pi_2 v - 1$, the left hand side of equation 11 is negative. Since $\beta_H^*$ is a proper probability, such an equality can never be satisfied and thus the defender would have an incentive to deviate.
• **Proof of proposition 4.** First, recall the equilibrium probabilities from the attribution game. Let:

\[
\begin{align*}
\beta_{H1} &= \frac{\pi_2 w}{\pi_1 r_H + \pi_2 w} \\
\beta_{H2} &= \frac{(1 - \pi_2)w}{(1 - \pi_1)r_H + (1 - \pi_2)w} \\
\beta_{L1} &= \frac{\pi_1 r_L + \pi_2 w}{\pi_2 w} \\
\beta_{L2} &= \frac{(1 - \pi_2)w}{(1 - \pi_1)r_L + (1 - \pi_2)w}
\end{align*}
\]

(12)

Since \(\pi_1 > \pi_2\) and \(r_H > r_L\), it can be shown that \(\beta_{H1} < \beta_{L1} < \beta_{H2} < \beta_{L2}\). By the same argument as in the attribution game, any equilibrium must have \(\beta_{H1} < \beta_H < \beta_{L2}\) and \(\beta_{H1} < \beta_L < \beta_{L2}\). The reason is that if any of the attacker’s randomization probabilities are outside these bounds, the defender’s best response, regardless of its type, is to either always retaliate or never retaliate, which cannot be part of an equilibrium (because then the attacker would no longer be willing to randomize). Given this fact, we will now prove each claim in the proposition.

1. **Proof of Part 1:** Under the strategy profile described in part 1, Bayes rule dictates that when the attacker receives signal \(s_H\), it knows with probability 1 that the defender is type \(H\). Therefore, when the attacker receives signal \(s_H\), any equilibrium must have the players play the attribution game and the attacker attacks with probability \(\beta_H = \beta_{H1}\) or \(\beta_H = \beta_{H2}\), depending on the sign of \(1 - \pi_1 c_H + \pi_2 v\). For the defender to be willing to randomize between signaling \(s_L\) and \(s_H\), it must be indifferent between sending the two signals. We examine the cases separately.

   - Suppose \(\beta_H = \beta_{H2}\). When the defender signals \(s_H\), it is indifferent between retaliating after \(a_1\) only and always retaliating. Thus, its expected utility is \(\beta_H(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w = \beta_H(r_H + w - 1) - w\). Let \(\beta_L\) be the probability the attacker attacks when it receives signal \(s_L\).

   - Suppose \(\beta_L < \beta_H\). If \(\pi_1 r_H + \pi_2 w - 1 < 0\), then when the defender of type \(H\) signals \(s_L\) and only attacks after \(a_1\), it earns \(\beta_L(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w\) which is strictly greater than its expected utility from signaling \(s_H\) because \(\beta_L < \beta_H\) and \(\pi_1 r_H + \pi_2 w - 1 < 0\) by assumption. Therefore, the attacker would not be willing to randomize between signals when it is type \(H\). If \(\pi_1 r_H + \pi_2 w - 1 > 0\), then the attacker would not be willing to signal \(s_L\) and only attack after \(a_1\) because \(\beta_L < \beta_H\), and the payoff for only attacking after \(a_1\) is increasing in \(\beta\). Therefore, the only way the defender can be indifferent between signaling \(s_H\) and \(s_L\) is if \(\beta_L\) is such that the defender’s best response is to never retaliate after signaling \(s_L\). This condition is given by

\[
\begin{align*}
-\beta_L &= \beta_H(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w \\
-\beta_L &= \beta_H(\pi_1 r_H + \pi_2 w - 1) - \beta_{H1}(\pi_1 r_H + \pi_2 w) \\
\beta_H - \beta_L &= \pi_1 r_H + \pi_2 w
\end{align*}
\]

(13)

For the condition in equation 13 to hold, it must be that \(\beta_L < \beta_{H1}\) since by assumption \(\pi_1 r_H + \pi_2 w > 1\). However, by the argument above, there is no equilibrium in which the attacker randomizes with a probability \(\beta_L < \beta_{H1}\) so there cannot be an equilibrium with \(\beta_L < \beta_H\).

- Suppose \(\beta_L > \beta_H\). This means that when the defender of type \(H\) signals \(s_L\) and the attacker randomizes with probability \(\beta_L\), the defender’s best response is to always retaliate. This implies that for all values of \(\beta\) such that \(\beta_H < \beta < \beta_L\), the defender’s expected utility is either strictly increasing or strictly decreasing in \(\beta\), depending on the sign of \(r_H + w - 1\). Due to strict monotonicity of the defender’s utility with respect to \(\beta\), it cannot be indifferent between signaling \(\beta_H\) and \(\beta_L\).
Since there cannot be an equilibrium with $\beta_H = \beta_{H2}$ and $\beta_L < \beta_H$ or $\beta_L > \beta_H$, there cannot be an equilibrium with $\beta_H = \beta_{H2}$.

* Suppose $\beta_H = \beta_{H1}$. In this case, the defender of type $H$ is indifferent between never retaliating and retaliating after $o_1$ only and earns an expected utility of $-\beta_H = \beta_H (\pi_1 r_H + \pi_2 w - 1) - \pi_2 w$. By the argument above $\beta_L$ cannot be less than $\beta_{H1}$. Therefore, suppose $\beta_L > \beta_H$. If $\pi_1 r_H + \pi_2 w - 1 > 0$, then the defender’s expected utility of signaling $s_L$ and only retaliating after $o_1$ is $\beta_L(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w$ which is greater than its expected utility of after signaling $s_H$. Therefore, the defender of type $H$ would not be willing to randomize between signaling $s_L$ and $s_H$. If $\pi_1 r_H + \pi_2 w - 1 < 0$, the only way the defender can be indifferent between signaling $s_H$ and $s_L$ is if it always retaliates after signaling $s_L$. This implies

$$-\beta_H = \beta_H (\pi_1 r_H + \pi_2 w - 1) - \pi_2 w = \beta_L (r_H + w - 1)$$

(14)

where the first equality comes from the fact that at $\beta_{H1}$ the attacker must be indifferent between never retaliating and retaliating after $o_1$ only. Now consider a defender of type $L$ that always signals it is $\beta_{L1}$. In this case, its utility is either $-\beta_L, \beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$ or $\beta_L(r_L + v - 1)$, depending on which of its strategies are optimal at $\beta_L$. If a defender of type $L$ instead signaled $s_H$ and never retaliated, its expected utility would be $-\beta_H = \beta_H(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w = \beta_L(r_H + v - 1)$. The following inequalities show that this regardless of which one of the defender’s strategies is optimal at $\beta_L$, there exists a profitable deviation where the defender of type $L$ signals $s_H$ and never retaliates:

$$-\beta_L < -\beta_H \quad \text{(By assumption)}$$

$$\beta_L(r_L + w - 1) - w < \beta_L(r_H + w - 1) \quad \text{(Because } r_H > r_L\text{)}$$

$$\beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w < \beta_H(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w$$

The last line follows because $r_H > r_L$ and $\pi_1 r_H + \pi_2 w - 1 < 0$. This shows that there cannot be a PBE where $\beta_H = \beta_{H1}$.

Since there cannot be an equilibrium where the attacker randomizes with either $\beta_{H1}$ or $\beta_{H2}$ after observing $s_H$, there cannot be a PBE where a defender of type $L$ always signals $s_L$ and a defender of type $H$ randomizes between signaling $s_L$ and $s_H$.

- **Proof of Part 2:** Under the strategy profile described in part 2, Bayes rule dictates that when the attacker receives signal $s_L$, it knows with probability 1 that the defender is type $L$. Therefore, when the attacker receives signal $s_L$, any equilibrium must have the players play the attribution game and the attacker attacks with probability $\beta_H = \beta_{L1}$ or $\beta_L = \beta_{L2}$, depending on the sign of $1 - \pi_1 c_1 + \pi_2 w$. For the defender to be willing to randomize between signaling $s_L$ and $s_H$, it must be indifferent between sending the two signals. We examine the cases separately.

* Suppose $\beta_L = \beta_{L2}$. In this case, the defender of type $L$ is indifferent between retaliating after $o_1$ only and always retaliating and earns $\beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w = \beta_L(r_L + w - 1) - w$. There cannot be an equilibrium where $\beta_H > \beta_L$ because then the defender’s best response would be to always retaliate regardless of its type. Therefore, it is sufficient to only consider the case where $\beta_H < \beta_L$. If $\pi_1 r_L + \pi_2 w - 1 < 0$, then the defender of type $L$ has a profitable deviation to signal it is type $H$ and only retaliates after $o_1$ and earns $\beta_H(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$ which is greater than $\beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$. Now suppose $\pi_1 r_L + \pi_2 w - 1 > 0$. Since there is no equilibrium where the attacker ever randomizes with a probability $\beta < \beta_{H1}$, it must be that $\beta_{H1} < \beta_H < \beta_L$. This means that the attacker’s best response at $\beta_H$ is either to retaliate after $o_1$ only or always retaliates. However, since $\pi_1 r_L + \pi_2 w - 1 > 0$ then $\pi_1 r_H + \pi_2 w - 1 > 0$ and $r_H + \pi_2 w - 1 > 0$ which means that the expected utility of the defender of type $H$ is increasing in $\beta$ and thus a defender of type $H$ would prefer to signal $s_L$. Consequently, there cannot be an equilibrium where $\beta_L = \beta_{L2}$.

* Suppose $\beta_L = \beta_{L1}$. In this case, a defender of type $L$ is indifferent between never retaliating and retaliating after $o_1$ only and earns $-\beta_L = \beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$. Suppose $\beta_H < \beta_L$, then there defender of type $L$ would not be willing to randomize between $s_L$ and $s_H$ because
it can signal $s_H$, never retaliate and earn $\beta_H$, which is greater than its expected utility of $-\beta_L$ by signaling $s_L$. Now suppose $\beta_H > \beta_L$. If $\pi_1 r_L + \pi_2 w - 1 > 0$, the defender of type $L$ can earn $\beta_H(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$ when it signals $s_H$ and only retaliates after $o_1$. Since this is higher than its maximum expected utility of $\beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$ when it signals $s_L$, a defender of type $L$ would not be willing to randomize its signals. Lastly, if $\pi_1 r_L + \pi_2 w - 1 < 0$, the defender can only be indifferent between signaling $s_H$ and $s_L$ if its best response is to always retaliate when it is type $L$ and signals $s_H$ (because its expected utility of only retaliating after $o_1$ is strictly monotonic in $\beta$). However, if the defender’s of type $L$’s best response to an attacker randomizing with probability $\beta_H$ is to always retaliate, it is also a defender of type $H$’s best response to always retaliate. Since the defender always retaliating after a signal cannot be part of an equilibrium, there cannot be an equilibrium where $\beta_H > \beta_L$.

Since there cannot be an equilibrium where the attacker randomizes with either $\beta_{L1}$ or $\beta_{L2}$ after observing $s_L$, there cannot be an equilibrium where a defender of type $H$ always signals $s_H$ and a defender of type $L$ randomizes between signaling $s_L$ and $s_H$.

\begin{itemize}
\item Proof of Proposition 5. Begin by considering the attacker. If the attacker receives signal $s_L$, then Bayes rules necessitate it knows the defender is type $L$ and thus the attacker and defender play the attribution game. As proposition 1 shows, there is an equilibrium in the attribution game where the attacker randomizes with probability $\frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$ and the defender retaliates with probability $\frac{1 - \pi_2 w}{\pi_1 r_L + \pi_2 w}$, which by assumption 2 is a proper probability. Now consider the attacker that receives signal $s_H$. For it to be willing to randomize, it must be indifferent between attacking and not attacking. Assuming the defender retaliates regardless of its type, this condition is given by:

\[
\frac{P(s_H|H)P(H)(1-c_H)}{P(s_H|H)P(H) + P(s_H|L)P(L)} + \frac{P(s_H|L)P(L)(1-c_L)}{P(s_H|L)P(H) + P(s_H|L)P(L)} = -v
\]

By assumption, $1 - c_H + v$ and $c_L - v - 1$ are both less than 0, so the second fraction in equation 15 is positive. By assumption, 5, the entire right hand side of equation 15 is less than 1 and thus a proper probability.

Now consider the defender. To begin, consider a defender of type $L$. A necessary condition for the defender of type $L$ to be willing to randomize between signaling 1) $s_L$ and earning a payoff of $\frac{-\pi_2 w}{\pi_1 r_L + \pi_2 w}$ and randomizing between never retaliating and retaliating after $o_1$ only and 2) $s_H$, inducing the attacker to retaliate with probability $\beta_H$, and always retaliating, it must be indifferent between the two signals. This implies

\[
\frac{-\pi_2 w}{\pi_1 r_L + \pi_2 w} = \beta_H(r_L + w - 1) - w
\]

Since, $\pi_1 r_L + \pi_2 w < 1$ by assumption, the defender’s expected utility from retaliating after $o_1$ only is decreasing in $\beta$ and thus, the defender’s best response at $\beta_H$ is to always retaliate. Therefore, the defender of type $L$ is indifferent between signaling $s_H$ and $s_L$. Finally, consider the defender of type $H$. When the attacker randomizes with probability $\beta_H$, because it is a defender’s of type $L$ best response to always retaliate, it must also be a defender of type $H$’s best response to always retaliate. The defender’s payoff from always retaliating is $\beta_H(r_H + w - 1) - w = -\beta_L + \beta_H(r_H - r_L)$. 

\end{itemize}
What remains to be shown is that a defender of type $H$ does not have an incentive to signal $s_L$. If the defender signals $s_L$ and always retaliates, its payoff must be less than if it signals $s_H$ because its payoff to always retaliating is increasing in $\beta$. It’s payoff by signaling $s_L$ and never retaliating is

$$\frac{-\pi_2 w}{\pi_1 r_L + \pi_2 w} = \beta_H(r_L + w - 1) - w < \beta_H(r_H + w - 1) - w.$$  

Finally, it’s payoff of signaling $s_L$ and retaliating after $o_1$ only is

$$\beta_L(\pi_1(r_H - r_L) - 1)$$

which is strictly less than $-\beta_L + \beta_H(r_H - r_L)$. Therefore, the defender does not have an incentive to change its strategy.

No other semi-separating equilibrium. First, we will show that there is no equilibrium in which the defender randomizes its signal for each of its types. Then we will show that there is no equilibrium in which the defender randomizes only when it is type $H$.

For contradiction, suppose there is a a signaling equilibrium where the defender randomizes its signal for each of its types and induces the attacker to randomize with probability $\beta_H$ and $\beta_L$ and without loss of generality, assume $\beta_H > \beta_L$. There is no equilibrium where $\beta_L$ is so low that the attacker of type $H$ would never retaliate. Therefore, the defender of type $H$ must be indifferent between retaliating after $o_1$ only and always retaliating. This implies

$$\beta_L(\pi_1 r_H + \pi_2 w - 1) = \pi_2 w = \beta_H(r_H + w - 1) - w$$

(17)

of course, this can only happen when $(\pi_1 r_H + \pi_2 w - 1) < 0$ and $(r_H + w - 1)$. For a defender of type $L$ to be indifferent, there are two cases.

- Consider the case where the defender of type $L$ is indifferent between retaliating only after $o_1$ when the attacker attacks with probability $\beta_L$ and always retaliating when the attacker attacks with probability $\beta_H$. This implies

$$\beta_L(\pi_1 r_L + \pi_2 w - 1) = \pi_2 w = \beta_H(r_L + w - 1) - w$$

(18)

However, solving for equations 17 and 18 yields $\beta_H = \pi_1 \beta_L$ which violates the fact that $\beta_H > \beta_L$.

- Consider the case where the defender of type $L$ is indifferent between never retaliating when the attacker attacks with probability $\beta_L$ and always retaliating when the attacker attacks with probability $\beta_H$. This implies

$$-\beta_L = \beta_H(r_L + w - 1) - w$$

(19)

Equation 17 and 19 together imply that

$$\beta_L = \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w} + (r_H - r_L)(\beta_H - \pi_1 \beta_L)$$

(20)

However, the solution to equation 20 yields a value of $\beta_L > \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$, which cannot be part of an equilibrium because at such a value of $\beta_L$, the defender would prefer to retaliate after $o_1$.

Now consider the semi-separating strategy where the defender randomizes its signal when it is type $H$ and always signals $s_L$ when it is type $L$. If the attacker randomizes its signal when it is type $H$, then Bayes rule will dictate that when the attacker receives signal $s_H$, it knows the attacker is type $H$ with certainty: Let $\beta_H$ be the attacker’s randomization probability when it receives signal $s_H$. Since the attacker knows the defender’s type when the defender signals $s_H$, the players play the attribution game and therefore the attacker either randomizes with

$$\beta_H = \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$$

or

$$\beta_H = \frac{(1 - \pi_2) w}{(1 - \pi_1) r_H + (1 - \pi_2) w}.$$  

We consider these cases separately.

- Suppose $\beta_H = \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$ where the defender of type $H$ is indifferent between never retaliating and retaliating after $o_1$ only. For the defender of type $H$ to be willing to randomize its signal, it must be that the defender is indifferent between signaling $s_H$ and signaling $s_L$, inducing the attacker to attack with probability $\beta_L$ and always retaliating. However, at such a $\beta_L$, the defender of type $L$’s expected utility for any of it strategies is strictly less than its expected utility from signaling $s_H$ and never retaliating, therefore, the defender would never be willing to signal $s_L$.

- Suppose $\beta_H = \frac{(1 - \pi_2) w}{(1 - \pi_1) r_H + (1 - \pi_2) w}$ where the defender of type $H$ is indifferent between retaliating after $o_1$ only and always retaliating. For the defender of type $H$ to be willing to randomize its signal, it must be
that the defender is indifferent between signaling $s_H$ and signaling $s_L$, inducing the attacker to attack with probability $\beta_L$ and never retaliating. However, at such a value of $\beta_L$, the defender of type $H$ or type $L$ would ever retaliate and thus the attacker wouldn’t be willing to randomize but instead would attack with probability 1. Therefore, there can’t be an equilibrium in which $\beta_H = \frac{(1-\pi_2)_w}{(1-\pi_1)r_H+(1-\pi_2)_w}$.

Finally consider the semi-separating strategy where the defender randomizes its signal when it is type $L$ and always signals $s_H$ when it is type $H$. If the attacker randomizes its signal when it is type $L$, then Bayes rule will dictate that when the attacker receives signal $s_L$, it knows the attacker is type $L$ with certainty. Let $\beta_L$ be the attacker’s randomization probability when it receives signal $s_L$. Since the attacker knows the defender’s type when the defender signals $s_L$, the players play the attribution game and therefore the attacker either randomizes with $\beta_L = \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$ or $\beta_H = \frac{(1-\pi_2)_w}{(1-\pi_1)r_L+(1-\pi_2)_w}$. Our main proposition showed that there can be an equilibrium when $\beta_L = \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$ so here, we consider the case where $\beta_L = \frac{(\pi_2 - \pi_1)_w}{\pi_1 r_L + \pi_2 w}$.

The only way the defender can be indifferent between the attacker attacking with probability $\beta_L$ and $\beta_H$ is if $\pi_1 r_L + \pi_2 w - 1 > 0$ and the defender of type $L$ never retaliates at $\beta_H < \beta_L$. However, since $\pi_1 r_L + \pi_2 w - 1 > 0$, the attacker of type $H$’s expected utility is increasing in $\beta$ and therefore would prefer to signal $s_L$ and not $s_H$.

All of the cases show that there are no other semi-separating equilibria.