

Signaling Design for Deterrence Strategies

Hiroto Sawada

Raphael Boleslavsky

Kristopher W. Ramsay

Mehdi Shadmehr*

Abstract

States and alliances spend enormous resources signaling resolve to deter adversaries. Two canonical strategies are burning money (sunk costs) and burning bridges (audience costs). Large audience costs can deter effectively, but they may be infeasible and, even when feasible, can increase the likelihood of war relative to burning money. We propose a signaling design in which the defender retains the flexibility to fight or back down but commits ex ante, before learning its resolve, to a rule mapping realized resolve into costly actions. When audience costs are limited, this design raises defender payoffs while reducing the likelihood of war. We discuss implementation through laws and automated algorithms. We also characterize defender payoffs when audience costs are chosen endogenously from a feasible set, showing that lowering the maximum feasible audience cost can raise defender payoffs, and we compare these payoffs with those under the standard burning-money setting without commitment.

JEL Codes: D82, D83, F51, F52

Keywords: Deterrence, Signaling, Commitment, Burning Money, Audience Costs

*Sawada: Department of Politics, Princeton University (e-mail: hsawada@princeton.edu); Boleslavsky: Department of Business Economics & Public Policy, Kelley School of Business, Indiana University (e-mail: rbole@iu.edu); Ramsay: Department of Politics, Princeton University (e-mail: kramsay@princeton.edu); Shadmehr: Department of Public Policy, University of North Carolina at Chapel Hill (e-mail: mshadmeh@gmail.com).

1 Introduction

Following Russia’s seizure of Crimea, Deputy Secretary of Defense Robert Work explained that the European Reassurance Initiative—which funded additional rotational deployments and infrastructure in Eastern Europe—was intended “to signal our resolve and capacity to deliver” (Work 2014). Two years later, Secretary of Defense Ash Carter similarly stated, “We’re also increasing military exercises with allies and partners to demonstrate resolve and build their resilience,” while also emphasizing their contribution to interoperability (Carter 2016). Similarly, as NATO prepared in 2016 to establish an enhanced Forward Presence (eFP) in the eastern part of the Alliance, Secretary General Jens Stoltenberg explained that it “shows our resolve. It enhances our deterrence” (Stoltenberg 2016). He likewise described the purpose of NATO’s Trident Juncture 2018 exercise not only as “to strengthen our ability to operate together,” but also as “to send a clear message. NATO has the capabilities and the resolve to protect all Allies against any threat” (Stoltenberg 2018). Following Russia’s invasion of Ukraine in 2022, NATO placed about 40,000 troops under NATO command in the eastern part of the Alliance and more than 100,000 troops on heightened alert. Stoltenberg explained: “This is to send a very clear message to Moscow and leave no room for miscalculation or misunderstanding: an attack on one NATO Ally will trigger the full response from the whole Alliance. This is deterrence” (Stoltenberg 2022).

Indeed, this logic is embedded in NATO doctrine. NATO’s *Allied Joint Doctrine* identifies credibility and communication among the five principles of deterrence (North Atlantic Treaty Organization 2022, p. 43). It states that NATO must be perceived as possessing both the “collective will to act” and the capability to do so, and that, in some circumstances, the joint force should take considered risks to “communicate the strength of the Alliance’s commitment” (p. 44); see also Janzen (2022, p. 19). NATO’s current deterrence guidance similarly states that exercises contribute to deterrence by communicating the Alliance’s “capability, readiness and resolve to potential adversaries” (North Atlantic Treaty Organization 2026).

As these examples highlight, to deter an adversary, a state must make its willingness to defend sufficiently credible. But assertions may not be credible: a state that would ultimately back down has the same incentive to say that it will fight. States can make such claims more credible in two fundamentally different ways. They can take costly actions before a challenge occurs—mobilizing forces, conducting military exercises, accumulating weapons, or undertaking other costly preparations. Or they can take actions that make backing down costly if

a challenge later occurs—for example, by publicly staking their reputation on a threat.¹ We call the first strategy *burning money* and the second *burning bridges* or *creating audience costs*. Fearon (1997) refers to the corresponding signaling mechanisms as “sinking costs” and “tying hands.”

Which is the better way to deter an adversary: burning money or burning bridges (creating audience costs)? Burning bridges has an apparent advantage. Money burned is lost regardless of what happens next. A cost created by burning bridges, by contrast, is incurred only if the state is challenged and subsequently backs down. If the threat succeeds in deterring the challenger, or if the state fights when challenged, the cost will not be paid. When creating audience costs is feasible, it is therefore a remarkably effective approach.

This logic underlies Fearon’s (1997) result that tying hands yields a higher expected payoff on average than sinking costs. An empirical literature finds evidence consistent with audience costs (Tomz 2007; Trager and Vavreck 2011; Levy et al. 2015; Li and Chen 2021; Gelpi and Griesdorf 2001; Weeks 2008; Takei 2024; Crisman-Cox and Gibilisco 2018). However, evidence that audience costs are typically large enough to generate commitment effects is thinner (Snyder and Borghard 2011; Katagiri and Min 2019; Trachtenberg 2012; Gartzke and Lupu 2012; Croco, Hanmer, and McDonald 2021); see Slantchev (2011, pp. 46-62) and Hoshiro (2026) for reviews and Tang 2026 for a meta-analysis. Moreover, when some challengers are undeterrable, even sufficiently large audience costs can increase the probability of war relative to standard sunk-cost signaling by locking low-resolve defenders into fighting when they would otherwise concede. This raises a natural question: are there better ways for states to signal resolve?

We introduce a new form of commitment into the classical burning-money approach that simultaneously increases the state’s payoff and (weakly) reduces the likelihood of war. What if a state can commit, *before it learns its resolve*, to how it will burn money once its resolve is realized? That is, before knowing how much it will value the issue at stake in a future confrontation, the state chooses a contingent signaling strategy that maps its subsequently realized resolve into observable costly actions.² The state does not commit to fight. It re-

1. States can also increase their strength. Even absent any information transmission incentives, states take costly actions to raise their preparedness. For example, if the state can virtually ensure that it will win a fight against an adversary, that may guarantee deterrence. Our focus is on what states can do conditional on a preparedness level. More generally, the marginal gains from preparedness actions eventually fall below their marginal costs. When information transmission incentives are present, states then may go beyond those levels to communicate information even though such costly actions have minimal effect on preparedness.

2. This approach applies Boleslavsky and Shadmehr 2025’s game theoretical framework to crisis-signaling

tains the choice between fighting and backing down if a challenge actually occurs. Instead, it commits to *how it will signal its willingness to fight* after learning its resolve but before the adversary decides whether to challenge. The next section discusses how laws and automated algorithms can create this commitment power.

This form of commitment is conceptually distinct from burning bridges. Burning bridges restricts the state's future choices by changing the consequences of backing down. Commitment to a burning-money strategy instead restricts the state's future signaling choice: it determines how the state will communicate after learning information but before the adversary acts. A state can therefore commit to its eventual response, or it can commit to the process by which its information is revealed. We characterize when each form is more valuable.

This form of commitment changes the logic of burning money. Without commitment, a costly signal must be sufficiently expensive to discourage imitation. A resolved state burns money to distinguish itself from an unresolved state, and the amount it burns must be large enough that the unresolved state prefers not to mimic it. The cost of the action is therefore precisely what makes the action informative.

With ex ante commitment, the costly action no longer has to perform this function. The state chooses the mapping from future resolve into actions before knowing which resolve will be realized. It can therefore assign different observable actions to its future types in advance. Conditional on being able to commit to that assignment, the actions need only be distinguishable; their difference in cost need not be large enough to deter imitation ex post. Therefore, the state uses the smallest available positive costly action. Commitment strictly improves the state's ex ante payoff because it preserves the informational benefit of separation while reducing the resources that must be burned.

This logic generates a striking reversal of the usual interpretation of costly signals. When the state is relatively unlikely to be highly resolved, the familiar pattern obtains: the high-resolve state takes the costly action and the low-resolve state does not. But when high resolve is sufficiently likely, the optimal assignment reverses. The high-resolve state takes the costless action and the relatively unlikely low-resolve state burns money. The absence of a costly action then signals *greater* resolve. We refer to this feature as *reversal of meaning*. This reversal of meaning is possible because the meaning of an action is determined by the ex ante commitment to the signaling rule, rather than by a type's ex post willingness to incur its cost.

models in international relations.

Does this strengthened form of burning money overturn the advantage of burning bridges (creating audience costs)? The answer depends on how large a bridge the state is capable of burning—how large an audience cost the state can create. We show that sufficiently large audience costs retain a powerful advantage over burning money. But that advantage can arise through two distinct mechanisms. In the first and most familiar mechanism, burning bridges does not reveal resolve; it creates *ex post* resolve. An audience cost can be large enough that even the low-resolve defender, who would otherwise back down, prefers to fight rather than incur the cost of retreat. Both types can then credibly stand firm. When all challengers are deterrable, this approach trivially creates the highest payoff for the defender. However, even when undeterrable challengers may arise, low-resolve types create no externality for high-resolve ones: once they create large enough audience costs to lock themselves into fighting, they will be treated as high-resolve by challengers.

In the second mechanism, burning bridges works as a signaling device. A sufficiently large prospective cost of backing down makes the low-resolve defender unwilling to imitate the high-resolve defender. The two types separate, but the audience cost is never actually paid on the equilibrium path: the high-resolve type fights if challenged, while the low-resolve type does not send the signal. Burning bridges therefore produces the same information as burning money without dissipating resources along the equilibrium path.

Whenever the defender can generate an audience cost large enough to achieve at least one of these two objectives—separating the types or committing the low-resolve type to fight—burning bridges yields a strictly higher *ex ante* payoff than burning money. This is the best the defender can hope for under standard conditions.³ Therefore, this ranking survives even with the proposed new form of commitment. *Ex ante* commitment makes burning money cheaper, but it cannot reproduce these advantages when sufficiently large audience costs are feasible.

Importantly, this ranking reverses when sufficiently large audience costs are infeasible. With low audience costs, the low-resolve defender now has an incentive to mimic the high-resolve defender and the equilibrium involves pooling or semi-separation: the low-resolve defender may bluff by taking the same action as the high-resolve defender and then back down when challenged. Consequently, when audience costs are too low to prevent bluffing (i.e., either to induce separation or to change subsequent behavior), committing to how money will

3. In particular, this holds when the defender's interim payoffs are convex, as we will show is the case with standard distributions of the challenger's valuation of the contested object when the expected value is large compared to the costs of conflict.

be burned provides the defender with a strictly higher ex ante payoff for every interior prior. The reason is not that the state can burn more money. In fact, commitment makes large expenditures unnecessary. By assigning actions to future types before learning its resolve, the state can use the smallest available costly action to communicate information. Even without commitment, burning money outperforms constrained audience costs if and only if the low-resolve conditions, and hence bluffing and backing down, arise with sufficient frequency.

The two signaling approaches also have different implications for war. Commitment to a burning-money strategy changes the cost—and sometimes even the meaning—of the signal, but it does not change the equilibrium probability of war. With and without commitment, burning money ultimately separates high and low resolve, so war occurs only when a high-resolve defender encounters an undeterrable challenger.

The audience-cost approach leads to war at least as often as burning money, but for different reasons. When a large audience cost separates the types without committing the low-resolve defender (second mechanism above), burning bridges and burning money generate the same probability of war. When the audience cost instead works by committing the low-resolve defender to fight (first mechanism), burning bridges generates strictly more war: some challenges that would have ended in acquiescence now lead to fighting. This is the familiar danger of tying hands emphasized by Fearon (1997)—the same commitment that strengthens deterrence can leave a state locked into fighting when deterrence fails.

When audience costs are too small to sustain separation, they also generate strictly more war, but for the opposite reason. They do *not* successfully tie the low-resolve defender's hands. Instead, they allow bluffing. Because a signal can come from either type, the challenger sometimes tests a defender it believes may be unresolved and instead encounters the high-resolve type, which fights. Thus, excess war under burning bridges can arise from two very different failures of discretion: sufficiently strong audience costs can commit too many types to fight, while insufficiently strong audience costs can leave too much uncertainty about which type the challenger faces.

This analysis points to a different way of thinking about the resources that support credible signaling. One might expect the choice between burning money and burning bridges (creating audience costs) to depend on both political institutions that create commitment powers (e.g., audience costs) and economic capacity that enables richer states to sink larger costs. However, once a sufficiently rich set of burning-money actions includes a small posi-

tive action, the ability to burn still larger amounts is not what makes this approach effective. With commitment, the state wants to use the cheapest distinguishable signal. Therefore, it is the defender’s ability to commit to a burning-money strategy, largely affected by political institutions, not the size of the money burned, that controls the efficacy of burning money for the defender. By contrast, the largest bridge the state can burn (the maximum feasible audience cost), also largely affected by political institutions, is decisive. Whether audience costs are large enough to sustain separation or alter subsequent incentives determines both the payoff ranking and, in part, the probability of war. The comparison, therefore, shifts attention from the sheer capacity to dissipate resources toward the institutional capacity.

Our main contribution is to introduce a new form of commitment to burning-money signaling protocols that is robustly better than signaling through audience costs in settings where states cannot create sufficiently large audience costs to eliminate bluffing—full commitment and full separation through audience costs are both infeasible. This advantage is robust in the sense that it applies to any (interior) prior and in the sense that it features a lower probability of war. We show how such a protocol leads to the reversal of the meaning of money burning, where low-resolve types burn money while high types do not. Our results apply the techniques developed in Boleslavsky and Shadmehr (2025) to crisis signaling settings—see Horz and Simpson (2025), Chen (2025), and Boleslavsky, Jungbauer, and Shadmehr (2026) for applications to law and online sales platforms. We discuss the feasibility of such commitment protocols in the next section.

Our secondary contribution is to characterize equilibrium payoffs in settings with limited audience costs and compare them with payoffs in burning-money settings with and without commitment. A large theoretical literature investigates various aspects of crisis signaling, including the endogenous determination of audience costs (Smith 1998; Acharya and Grillo 2019; Ashworth and Ramsay 2024) and variants of the canonical crisis signaling model with limited audience costs (Tarar and Leventoglu 2013; Kurizaki and Whang 2015). For example, Kurizaki and Whang (2015, p. 954) consider an exogenously fixed audience cost and show that as the audience cost shrinks, the equilibrium moves from separating to semi-separating to pooling. Our contribution here is to show that the defender’s payoff is non-monotone in the maximum feasible audience cost and that, when feasible audience costs are limited, the defender’s payoff under burning money without commitment exceeds that under audience costs if and only if the defender is a priori sufficiently likely to be low-resolve.

2 Commitment Mechanisms for Signaling Resolve

Our analysis highlights the value of constructing or strengthening mechanisms that enable commitment to mapping realized resolve into costly actions. In this section, we highlight two such mechanisms: laws and automated algorithms.

Laws The government can pass laws that map the state’s resolve (e.g., its valuation of the contested object or other parameters that determine its willingness to fight) into an action. States already employ systems that combine preestablished plans with observable and costly actions that aim, at least partly, to signal resolve to defend against attacks. For example, the U.S. DEFCON (defense readiness condition) system maps a level of threat to prespecified actions that are both costly and informative about U.S. intentions and resolve. Discussing DEFCON levels in a meeting of the Executive Committee of the National Security Council during the Cuban Missile Crisis, for example, Secretary of Defense Robert McNamara emphasized both the informational content and the material cost of U.S. readiness:⁴

any reduction in the state of readiness of U.S. forces would be a sign to the Soviet Union. The Russians would know immediately if our state of alert was reduced. He suggested that no reduction be made today. However, SAC [Strategic Air Command] should stand down as soon as possible because the present alert involves burning out large amounts of spare parts.

As another example, the federal regulation establishing Graduated Mobilization Response (GMR) states that “[t]he GMR system enables the nation to approach mobilization planning and actions as part of the deterrent response capability and to use it to reduce the probability of conflict.”⁵ The 1990 *National Security Strategy of the United States* similarly called for plans to “develop graduated responses that will themselves signal U.S. resolve and thus contribute to deterrence.”⁶ Other mappings include DoD Force Protection Conditions (FPCON), NATO Alert System during the Cold War, and Department of Energy Graded Protection.

These examples illustrate relevant institutional building blocks—preplanned contingent responses, differentiated observable actions, and costly measures explicitly understood to

4. *Foreign Relations of the United States*, Volume XI, Document 170.

5. 44 C.F.R. 334.3(c).

6. The White House, March 1990. *National Security Strategy of the United States*. Washington, DC: The White House, p. 27. See also National Research Council. 1991. *Graduated Mobilization Response: Forging a Strong Partnership with Industry*. Washington, DC: The National Academies Press.

convey information to an adversary. Our results push further in this direction. A state may benefit from committing—e.g., through legislation—to the mapping between subsequently realized conditions and costly signals, thereby surrendering some discretion over *how it communicates* while retaining discretion over whether ultimately to fight.⁷

Automated Algorithms The government can set up automated systems that evaluate resolve, map it into costly actions, and execute the actions. Automated systems have been in place since the Cold War and are analyzed particularly in the context of nuclear warfare. Schwartz and Horowitz 2026 discuss specific programs and provide evidence that automated systems are perceived as a commitment device. In the burning money setting, automated algorithms could, e.g., issue warnings that induce costly actions, or even expend ammunition and equipment without causing casualties.⁸

3 Model

We begin by outlining the model informally before turning to its formal exposition. We consider deterrence settings with two canonical forms: burning money and audience costs (or burning bridges). The two settings have the same structure and differ only in their payoffs. In both cases, a challenger seeks to obtain an object currently held by a defender. If the challenger attacks and the defender does not defend, the object passes from the defender to the challenger. If the defender instead chooses to defend, war follows. War is costly to both sides, its outcome is uncertain, and the winner obtains the contested object.

The defender’s valuation of the object is private information. Some defender types value the object enough to fight for it, while others do not. The defender’s valuation can therefore be interpreted as its resolve to fight in order to preserve the status quo.

In the burning-money setting, the defender takes an early costly action that may signal how much it values the object. In the audience-cost setting, the defender instead takes an early action that becomes costly only if the challenger later attacks and the defender backs

7. In the burning-bridges approach, it is difficult to imagine how a country can commit to create audience costs, e.g., by requiring the government to make certain public statements, and no others, under certain conditions.

8. In other domains such as online sales platforms, computer algorithms are viewed as commitment mechanisms on the part of the seller (Boleslavsky, Jungbauer, and Shadmehr 2026).

down rather than fighting.

In the burning-money setting, the defender can commit in advance to an action that depends on its valuation of the object. For example, when the rule of law is strong in a country, it can pass laws requiring certain costly actions upon learning the value of the contested object. The executive agencies responsible for carrying out those actions then act accordingly. By contrast, it is harder to interpret such a commitment power in audience-cost settings.

Formally, there is a defender and a challenger. First, the defender chooses an action $s \in S = \{0, s_1, s_2, \dots, s_N\}$ with $0 < s_1 < s_2 < \dots < \bar{s} := s_N$. Next, the challenger observes the defender's action and then chooses $a \in \{0, 1\}$, where 1 means “attack” and 0 means “don't attack.” If $a = 0$, the game ends. If $a = 1$, then the defender chooses $b \in \{0, 1\}$, where 1 means “fight” and 0 means “don't fight,” payoffs are realized, and the game ends.

When we need to distinguish between the settings, we use the superscript b for the burning-money setting and t for the audience-cost setting. For example, S^b refers to the set of available actions (signals) and s_1^b to the smallest strictly positive action (signal) in the burning money setting, and $\bar{s}^t := s_N^t$ refers to the largest action (signal) in the audience cost setting.

There is a state of the world $(\omega, v_c) \in \{\omega^H, \omega^L\} \times \mathbb{R}_+$ with $0 < \omega^L < \omega^H$, where $\omega, v_c > 0$ are the values of the contested prize to the defender and challenger, respectively. We refer to the defender with $\omega = \omega^H$ as the high type and the one with ω^L as the low type. Let $u(a, b, v_c)$ be the challenger's payoff, $v^b(s, a, b, \omega)$ be the defender's payoff in the burning money setting, and $v^t(s, a, b, \omega)$ be the defender's payoff in the audience-cost setting:

$$u(a, b, v_c) = a(b((1-p)v_c - c_c) + (1-b)v_c) \quad (1)$$

$$v^b(s, a, b, \omega) = (1-a)\omega + ab(p\omega - c_d) - s \quad (2)$$

$$v^t(s, a, b, \omega) = (1-a)\omega + ab(p\omega - c_d) - a(1-b)s, \quad (3)$$

where $p \in (0, 1)$ is the probability that the defender wins a war with the challenger, and $c_d, c_c > 0$ are the costs of war to the defender and challenger, respectively.

The first component of the state, ω , is the private information of the defender, and its second component, v_c , is the private information of the challenger. ω and v_c are independent from each other, and players share a prior that $v_c \sim F$ and that $\Pr(\omega = \omega^H) = \mu_0 \in (0, 1)$. F is the CDF with density f , where $F(0) = 0$, and $F(x)$ is twice-differentiable and $f(x) > 0$ for $x \in (0, c_c/(1-p))$. We denote a posterior belief after a signal s by $\mu_s = \Pr(\omega = \omega^H | s)$. We

denote by $\hat{v}^b(\mu_s, s)$ and $\hat{v}^t(\mu_s, s)$ the defender's expected (interim) payoff, given a signal s with the associated posterior μ_s , in the burning-money and audience-cost settings, respectively.

The game without commitment proceeds as follows:

1. Nature determines ω and v_c .
2. The defender observes ω and chooses s .
3. The challenger observes v_c and s and chooses a .
4. The defender observes a and chooses b .
5. Payoffs are realized, and the game ends.

In the burning money setting, we further analyze an environment in which the defender, before learning the state ω , commits to a strategy of choosing s . That is, the defender, before observing ω , commits to a strategy $\pi(s|\omega) : \{\omega^L, \omega^H\} \rightarrow \Delta(S)$ that maps ω to a probability distribution over S .⁹ The timing of the game in this setting is as follows:

1. The defender commits to $\pi(s|\omega)$ and this mapping is observed by the challenger.
2. Nature determines ω and v_c .
3. The defender observes ω and chooses s according to $\pi(s|\omega)$.
4. The challenger observes v_c and s and chooses a .
5. The defender observes a and chooses b .
6. Payoffs are realized, and the game ends.

We focus on sender-optimal perfect Bayesian equilibria that satisfy the D1 criterion (Banks and Sobel 1987). Given parameters, a sender-optimal equilibrium is one that maximizes the sender's (defender's) ex ante payoff.

The following notation will simplify the exposition:

$$W^i = p\omega^i - c_d, i \in \{L, H\}, \delta = \omega^H - W^H, v_c^*(\mu) = \frac{\mu c_c}{1 - \mu p}, F_\mu = F(v_c^*(\mu)),$$

9. Commitment power in audience-cost settings seems substantively implausible. For theoretical completeness, the results for the audience-cost setting with commitment are in Proposition 11 in Appendix C.

where μ is the challenger's belief that $\omega = \omega^H$. W^i is a type i defender's payoff from war, and δ is the high type's expected marginal gain from deterring the challenger over going to war with the challenger. If $W^L, W^H > 0$, neither signaling through burning money nor creating audience costs has any value. If $W^L, W^H < 0$, signaling through burning money has no value and hence the comparison is straightforward. Thus, we focus on settings in which

$$W^L < 0 < W^H,$$

so that, under the default of $s = 0$ in (2)-(3), a high-type defender will defend if attacked, but the low-type defender will not.

We collect the assumptions used in the analysis below. We specify the parts that are used in each proposition.

Assumption 1. (i) *The defender's interim payoff in the burning-money setting, $\widehat{v}^b(\mu, s)$ in equation (6), is strictly convex in μ .*

(ii) *The smallest strictly positive signal in the burning-money setting is sufficiently small:*

$$s_1^b < \begin{cases} F_1 \omega^L & \text{(ii-a)} \\ (F_1 - F_{0.5}) \delta - F_{0.5} \omega^L & \text{(ii-b)} \\ \left(F_1 - \frac{\bar{s}^t}{\omega^L + \bar{s}^t} \right) \delta & \text{(ii-c)} \end{cases}$$

(iii) *The set of feasible signals in the burning-money setting is rich in the following sense:*

$$F_1 \omega^L \in S^b \quad \text{(iii-a)}$$

$$S^b \cap (s, s + (F_1 - F_{\mu_0})\delta) \neq \emptyset \text{ for any } s \in S^b \cap [0, F_{\mu_0} \omega^L] \quad \text{(iii-b)}$$

Assumption 1-(i) is satisfied with common distributions of the challenger's valuation v_c such as uniform, exponential, and lognormal in natural settings where the challenger's expected valuation normalized by the cost of war, $\mathbb{E}[v_c]/c_c$, is sufficiently large—see Proposition 9 in the Appendix.

The remaining assumptions are technical, concerning the richness of the action (signal) space in the burning-money setting S^b . We state them to invoke the exact requirements in each result; **1-(ii)** means that small actions (signals) are feasible. The right-hand side is strictly positive: $\omega^L, F_1 > 0$; **1-(ii-b)** is only invoked together with **1-(i)**, under which the right-hand side of **(ii-b)** is strictly positive (see Remark 1 in the Appendix); **1-(ii-c)** is only invoked when $\bar{s}^t$ is sufficiently small that the right-hand side is strictly positive. **1-(iii)** ensures that the results are not driven by the coarseness of the action (signal) space; e.g., **(iii-a)** ensures that the signal that maximizes the defender's payoff among separating equilibria is feasible.

4 Burning Money and the Reversal of Its Meaning

We begin with the burning-money setting. Let $\hat{a}(\mu_s, s, v_c)$ and $\hat{b}(\omega, s)$ be the equilibrium values of a and b , respectively. From (2) and $W^L < 0 < W^H$, the defender defends against an attack if and only if $\omega = \omega^H$:

$$\hat{b}(\omega, s) = \mathbb{1}(\omega = \omega^H). \quad (4)$$

Anticipating this, from (1), a challenger with a belief μ attacks if and only if $v_c > v_c^*(\mu)$. Therefore, from the perspective of the defender at the beginning of the game, upon observing a signal s , the challenger will attack with probability

$$1 - F_{\mu_s}, \quad (5)$$

where μ_s is the challenger's belief upon observing signal s . Combining (2), (4), and (5), the defender's interim payoff is

$$\begin{aligned} \hat{v}^b(\mu_s, s) &= \mathbb{E}_{\mu_s, v_c} \left[v^b(s, \hat{a}(\mu_s, s, v_c), \hat{b}(\omega, s), \omega) \right] \\ &= -s + (1 - F_{\mu_s}) W^H \mu_s + F_{\mu_s} (\mu_s \omega^H + (1 - \mu_s) \omega^L). \end{aligned} \quad (6)$$

Proposition 1 (Burning Money without Commitment). *Suppose Assumption 1-(iii-a) and (iii-b) hold. Consider the burning-money setting without commitment. In equilibrium, the defender chooses $s = F_1 \omega^L$ when $\omega = \omega^H$ and $s = 0$ when $\omega = \omega^L$; the challenger attacks*

when $s < F_1\omega^L$ or $v_c > v_c^*(1)$, upon which the defender fights if and only if $\omega = \omega^H$.

Proposition 1 shows that burning-money settings without commitment lead to separation, in which high types choose the least costly action that prevents low types from imitating. It is a canonical benchmark that appears in different forms in the literature (Fearon 1997, pp. 75–76; Reich 2026, p. 17; Slantchev 2011, p. 33). Other equilibria where low types pool with high types are not sustainable by D1 refinement, which requires the challenger to have reasonable beliefs off the equilibrium path, and by the sender optimal equilibrium selection, so that the low type separates when indifferent.¹⁰

The equilibrium minimizes the chances of war by deterring all deterrable challengers when the defender is of high type, and through concession when the defender is of low type. However, information transmission is costly as high types must burn just enough money to preclude imitation by low types.

The cost of this approach stems from the need for the high type to take sufficiently costly action. Our key observation is that if the defender could, ex ante, before learning its type, commit to a burning-money strategy, it could maintain separation while reducing the signaling costs of information transmission. This commitment amounts to the defender ex ante choosing a mapping $\pi(s|\omega)$ from its type $\omega \in \{\omega^L, \omega^H\}$ into a probability distribution over its actions S^b .

Figure 1 illustrates the logic. Dashed lines are the defender’s interim payoffs $\hat{v}^b(\mu, s)$ for $s \in S^b = \{0, 0.1, 0.15\}$. $\hat{v}^b(\mu, s)$ is the defender’s payoff if it takes action s and the belief that it is a high type is μ . Of course, not all such beliefs are feasible: a rational challenger makes inferences about the defender’s type upon observing the defender’s action s . The defender, of course, would like to somehow induce the challenger to believe that it is a high type by doing the free “action” of doing nothing $s = 0$ —it wants to be at the most extreme northeast corner: $\hat{v}^b(\mu = 1, s = 0)$. But it cannot: If the defender always does nothing, the challenger will maintain the prior μ_0 . What can the defender achieve and how? Proposition 2 describes the answer.

Proposition 2 (Burning Money with Commitment). *Suppose Assumption 1-(i), (ii-a), and (ii-b) hold. Consider the burning-money setting with commitment. In equilibrium,*

10. This latter choice is a technical assumption with no substantive consequences when the action (signal) space is sufficiently rich. Without it, high types would choose a larger signal than $F_1\omega^L$ to separate themselves.

- If $\mu_0 \leq 1/2$, the defender sends $s = s_1^b$ when $\omega = \omega^H$ and $s = 0$ when $\omega = \omega^L$. The challenger challenges if and only if $s = 0$ or $v_c > v_c^*(1)$, upon which the defender fights if and only if $\omega = \omega^H$.
- If $\mu_0 \geq 1/2$, the defender sends $s = 0$ when $\omega = \omega^H$ and $s = s_1^b$ when $\omega = \omega^L$. The challenger challenges if and only if $s = s_1^b$ or $v_c > v_c^*(1)$, upon which the defender fights if and only if $\omega = \omega^H$.

For any $\mu_0 \in (0, 1)$, commitment power strictly improves the defender's ex ante payoff.

Proposition 2 shows that the defender continues to separate with commitment power but uses very different signals to do so. When the high type is unlikely, as before, the low type does nothing, while the high type takes a costly action. However, this action is the least costly positive action. It no longer needs to be sufficiently costly to prevent imitation by the low type. It only needs to be an action different from the free “doing nothing” action of the low type. This already saves the defender the cost of money burning.

Remarkably, when the high type is likely, the defender's behavior changes starkly. Now, the high type will take no action, while the low type takes a costly action. That is, burning money now reveals that the defender has low resolve. Intuitively, if the defender wants separation, it only needs to ensure that different types take different actions. To save money, these actions should be either to do nothing or to take the least costly action. But which type should take which action? The type that is ex ante less likely should take the more expensive action, and the type that is ex ante more likely should take the less expensive action.

Proposition 2 applies the techniques developed in Boleslavsky and Shadmehr (2025). We now describe the logic of this result in more detail. Recall that the defender would like to burn no money $s = 0$ and at the same time induce the challenger to believe it is a high type. But it cannot; e.g., if all types always do nothing, then no information is transmitted, and the challenger's belief remains at the prior. What restrictions does the defender have, and how can it maximize its payoffs subject to those restrictions?

Adopting the analyses of Kamenica and Gentzkow (2011) and Boleslavsky and Shadmehr (2025), the defender faces two simple restrictions. First, beliefs and expectations must satisfy the law of iterated expectations. In particular, if action (signal) s induces a belief μ_s and if action s is taken with probability τ_s , then: $\mathbb{E}_\tau[\mu_s] = \mathbb{E}_\tau[\mathbb{E}[\mathbf{1}\{\omega = \omega^H\} | s]] = \mathbb{E}_\omega[\mathbf{1}\{\omega = \omega^H\}] = \mu_0$. Because action s induces μ_s , $(\tau_s)_{s \in S^b}$ is effectively a distribution over posteriors.

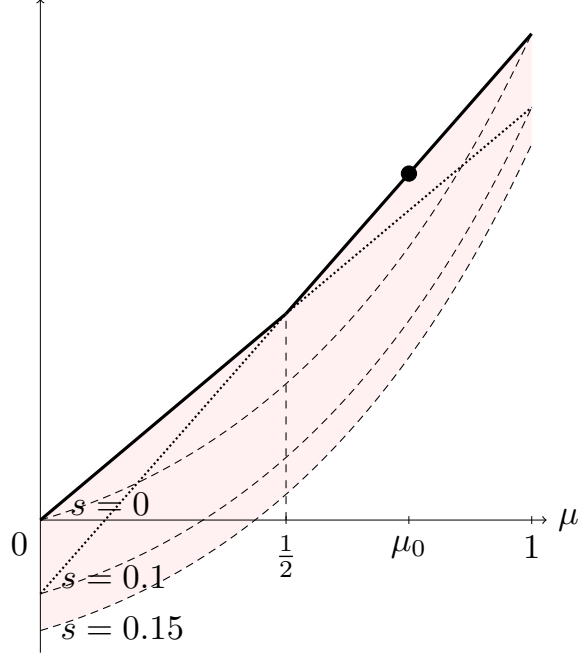


Figure 1: Burning Money with Commitment

Moreover, any combination of posteriors that satisfies the law of iterated expectations can be induced by an appropriately chosen mapping $\pi(s|\omega)$. In particular, by (ex ante) committing to a signaling strategy $\pi(s|\omega) : \{\omega^L, \omega^H\} \rightarrow \Delta(S^b)$, the defender can “split” any prior belief of the challenger μ_0 into several posterior beliefs:

$$\sum_{s \in S^b} \tau_s \mu_s = \mu_0 \quad \text{s.t.} \quad \sum_{s \in S^b} \tau_s = 1 \quad \text{and} \quad \tau_s \geq 0, \quad (7)$$

where τ_s here is the probability that action (signal) $s \in S^b$ is sent, and hence the posterior μ_s is induced. Second, each action $s \in S^b$ can induce only one posterior belief. It follows that the defender can produce any convex combination of its interim payoffs $\{\hat{v}^b(\mu_s, s) : s \in S^b\}$ that satisfies the law of iterated expectations:

$$\max_{(\mu_s, \tau_s)_{s \in S^b}} \sum_{s \in S^b} \tau_s \hat{v}^b(\mu_s, s) \quad \text{s.t.} \quad (7). \quad (8)$$

The optimization (8) is simply the convex combination of the set of all defender’s interim payoffs with the restriction that each interim payoff must be used at most once: a convex combination cannot use two points from the same interim payoff graph, because that would

imply that the same action (signal) induces two different posterior beliefs. This is the topological join of a set of interim payoff graphs, $\{\hat{v}^b(\mu_s, s) : s \in S^b\}$. The shaded area in Figure 1 shows the join. At any given prior $\mu = \mu_0$, the defender wants to create the maximum value of the join, the join envelope, shown in the figure by solid black lines. We note that, in contrast to the Bayesian persuasion literature, in which the convex hull and concave envelope are the relevant objects, the join set is not generally convex and the join envelope is not generally concave—e.g., it is convex here.

Figure 1 provides a simple characterization of the defender’s optimal commitment strategy. Reading off the join envelope, the solid black line, when the high type is ex ante unlikely ($\mu_0 < 1/2$), the defender splits the prior into two posteriors $\mu = 0$ and $\mu = 1$: when $s = 0$, the posterior is $\mu_{s=0} = 0$; when $s = 0.1$, it is $\mu_{s=0.1} = 1$. That is, the low type burns no money, and the high type burns the minimum available amount $s_1^b = 0.1$. There is full separation, but at a minimal cost: because the defender has committed to the mapping $\pi(s = 0|\omega = \omega^L) = \pi(s = 0.1|\omega = \omega^H) = 1$, the low type is not able to imitate the high type. The cost of signaling here stems from the need to take an action different from the free action of the low type, not from the need to take an expensive enough action that prevents the low type’s imitation.

By contrast, when the high type is ex ante likely ($\mu_0 > 1/2$), the defender still splits the prior into two posteriors $\mu = 0$ and 1, but reverses the signals used to induce them: when $s = 0$, the posterior is $\mu_{s=0} = 1$, not 0; when $s = 0.1$, it is $\mu_{s=0.1} = 0$, not 1. That is, the high type burns no money, and the low type burns the minimum available amount $s_1^b = 0.1$. Because the defender has committed to the mapping $\pi(s = 0.1|\omega = \omega^L) = \pi(s = 0|\omega = \omega^H) = 1$, the low type cannot imitate the high type. The defender could still employ the strategy of having high types burn money, while low types do nothing. That would still save the defender money compared to the setting without commitment in which the high types would take even more costly actions. But the defender can do better by reversing the meaning of the actions: burning money means low type, while doing nothing means high type. The low type will have strictly negative payoffs: they take a costly action and back down. But they are unlikely to arise, and the costs saved by the likely high types more than compensate for those losses.

Through the reversal of the meaning of burning money, the defender can significantly reduce the costs of the burning-money approach and raise its payoff. As Figure 2 demonstrates, this commitment power generates most value exactly when the high type is more likely, and hence burning money without commitment, or even with commitment but without reversal

of meaning, is most expensive.

5 Audience Costs

We now analyze the audience-cost setting and then compare the defender's ex ante payoffs in the two settings. We distinguish between the settings in which audience costs can be large and those where they cannot.

Upon the challenger's attack, the defender fights if $\omega = \omega^H$ or $-s \leq p\omega^L - c_d$. Thus, the defender's interim payoff is

$$\begin{aligned} & \hat{v}^t(\mu_s, s) \\ = & \mathbb{1}(-s > p\omega^L - c_d) [F(v_c^*(\mu_s))(\mu_s\omega^H + (1 - \mu_s)\omega^L) + (1 - F(v_c^*(\mu_s)))(\mu_s(p\omega^H - c_d) - (1 - \mu_s)s)] \\ & + \mathbb{1}(-s \leq p\omega^L - c_d) [F_1(\mu_s\omega^H + (1 - \mu_s)\omega^L) + (1 - F_1)(\mu_s(p\omega^H - c_d) + (1 - \mu_s)(p\omega^L - c_d))]. \end{aligned} \tag{9}$$

5.1 Large Audience Costs

Consider settings where sufficiently large audience costs are feasible:

$$\bar{s}^t \geq s^* := \min \left\{ -p\omega^L + c_d, \frac{F_1}{1 - F_1} \omega^L \right\},$$

where we recognize that s^* is a function of p , c_d , ω^L , and F_1 . We use the convention that $\pm a/0 = \pm \infty$ for $a > 0$. Moreover, s^* is weakly increasing in F_1 , the likelihood that the challenger does not attack if it is sure that the defender will defend, with $\lim_{F_1 \rightarrow 0} s^* = 0$ and $\lim_{F_1 \rightarrow 1} s^* = -p\omega^L + c_d > 0$. In particular, if the challenger attacks almost surely even if it knows that the defender will defend upon an attack ($F_1 \approx 0$), then we are in this setting.

Proposition 3 (Large Audience Costs). *Suppose $\bar{s}^t \geq s^*$. In the audience-cost setting,*

- *If $-p\omega^L + c_d > \frac{F_1}{1 - F_1} \omega^L$, the defender sends some $s \geq \frac{F_1}{1 - F_1} \omega^L$ when $\omega = \omega^H$ and sends $s = 0$ when $\omega = \omega^L$; the challenger challenges when $s < \frac{F_1}{1 - F_1} \omega^L$ or $v_c > v_c^*(1)$, upon*

which the defender fights if and only if $\omega = \omega^H$.

- If $-p\omega^L + c_d < \frac{F_1}{1-F_1}\omega^L$, each type of the defender sends a signal $s \geq c_d - p\omega^L$. The challenger challenges when $s < c_d - p\omega^L$ or $v_c > v_c^*(1)$. Both types of the defender fight when challenged on the equilibrium path.

We note that in the second equilibrium, the two types of the defender do not have to send the same $s \geq c_d - p\omega^L$. Even if they send different s' and $s'' \neq s'$, these signals constitute an equilibrium that induces the same probability of attack and war, as long as $s', s'' \geq c_d - p\omega^L$.

Intuitively, the low type can choose $s \geq |p\omega^L - c_d|$ to commit itself to fighting if attacked, thereby transforming itself into the high type from the challenger's perspective. The low type will do that whenever

$$F_1\omega^L + (1 - F_1)(p\omega^L - c_d) \geq 0, \text{ i.e. } \frac{F_1}{1 - F_1}\omega^L \geq -p\omega^L + c_d, \quad (10)$$

where 0 is the low type's payoff if the challenger attacks and the defender does not defend. This logic has nothing to do with information transmission, and it is exactly the second equilibrium in Proposition 3. When (10) is reversed, the weak type does not want to commit. Instead, it wants to imitate the high type and reduce the probability of attack to $1 - F_1$, but backs down if the attack actually happens. But the high type can always choose a high enough signal to make imitation unattractive to the low type. If the high type takes $s \geq |p\omega^L - c_d|$, the low type would not imitate. In fact, even if the high type takes $s \geq \frac{F_1}{1-F_1}\omega^L$, the low type will not imitate because

$$F_1\omega^L + (1 - F_1)(-s) \leq F_1\omega^L + (1 - F_1)\left(-\frac{F_1}{1 - F_1}\omega^L\right) = 0.$$

That is, if $\bar{s}^t \geq \frac{F_1}{1-F_1}\omega^L$, the high type can choose a signal that the low type would want to imitate if and only if it wants to follow through and defend against an attack. Importantly, the low type would want to do so exactly when (10) holds. But then, the low type would not

need to pretend to be the high type; it could commit on its own by taking $s > |p\omega^L - c_d|$.

With full separation or the low type's commitment to defend if attacked, it follows that the defender's ex ante payoff (under the prior) is linear in the prior μ_0 . The red line in Panel (a) of Figure 2 demonstrates this.

Thus, in large audience-cost settings where $\bar{s}^t \geq s^*$, signaling has value if and only if $\frac{F_1}{1-F_1} \omega^L < -p\omega^L + c_d$. Otherwise, the commitment of burning bridges alone, without any signaling value, delivers the results. Moreover, when $\frac{F_1}{1-F_1} \omega^L < -p\omega^L + c_d$, so that signaling has value in the audience-cost setting, combining Propositions 1 and 3 reveals that signaling requires a more costly available signal in the audience-cost setting, $\frac{F_1}{1-F_1} \omega^L$, than in the burning-money setting where $F_1 \omega^L < \frac{F_1}{1-F_1} \omega^L$.

That is, while signaling costs are not paid in equilibrium in large audience-cost settings, they require a higher feasibility threshold: the same cost $F_1 \omega^L$ that delivers separation in burning-money settings will not deliver separation in audience-cost settings.

Combining Propositions 1, 2, and 3 allows us to compare the defender's ex ante payoffs across the audience-cost and burning-money settings.

Proposition 4 (Large Audience Costs versus Burning Money). *Suppose Assumption 1-(i), (ii-a), and (ii-b) hold and $\bar{s}^t \geq s^*$. For any $\mu_0 \in (0, 1)$, the defender's ex ante payoff is strictly higher in the audience-cost setting than in the burning-money setting with and without commitment.*

Without commitment, this result echoes the result in Fearon (1997). Moreover, the superiority of audience cost remains even when the defender has commitment power. That is, although Proposition 2 shows that commitment power strictly improves the defender's ex ante payoff from burning money, it is still strictly lower than that from signaling by audience cost.

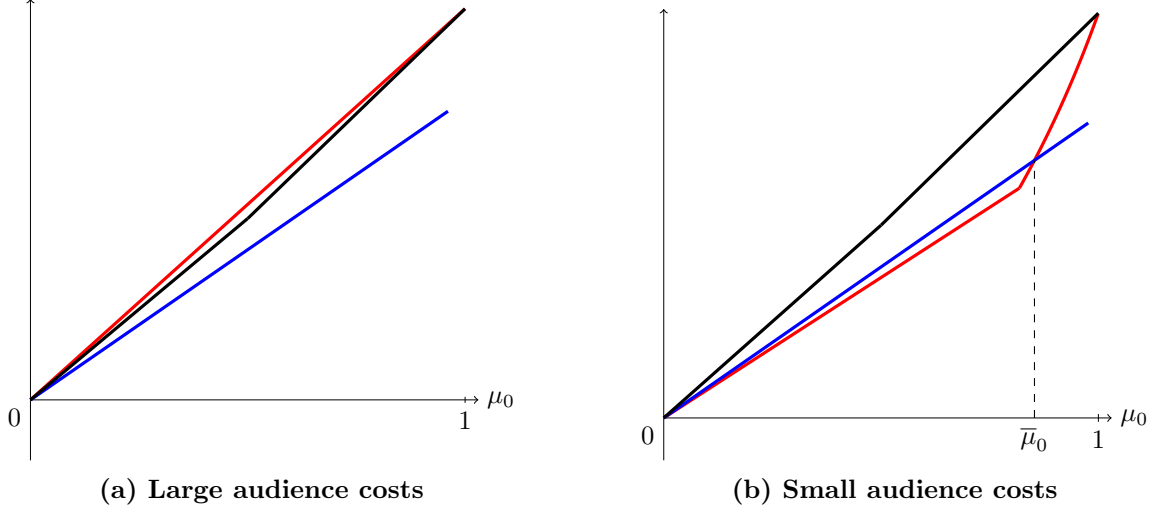


Figure 2: The defender's ex ante payoff under audience costs (red), burning money without commitment (blue), and burning money with commitment (black). In both panels, $\omega^L = 0.8$, $\omega^H = 3$, $p = 0.6$, $c_d = 1.5$, $c_c = 0.4$, and $S^b = S^t = \{0, 0.1, \dots, 0.6\}$, so $s_1^b = 0.1$ and $\bar{s}^t = 0.6$. In panel (a), $v_c \sim U[0, 2.5]$, $F_1 = \frac{2}{5}$, and $s^* = \frac{8}{15}$. In panel (b), $v_c \sim U[0, 1.5]$, $F_1 = \frac{2}{3}$, and $s^* = \frac{51}{50} > \bar{s}^t$. Vertical scales differ across panels.

5.2 Small Audience Costs

If the defender can create a sufficiently large audience cost, it achieves the best possible payoff: all deterrable challengers will be deterred at no realized cost. However, the literature highlights that the ability to create large audience costs is rare (Katagiri and Min 2019; Snyder and Borghard 2011; Trachtenberg 2012). We thus consider possibly more realistic settings where the ability to create audience costs is limited:

$$\bar{s}^t < s^*.$$

Proposition 5 characterizes the equilibrium.

Proposition 5 (Small Audience Costs). *Suppose $\bar{s}^t < s^*$. In the audience cost setting,*

- *If $\frac{F_{\mu_0}}{1-F_{\mu_0}} \omega^L < \bar{s}^t < s^*$, the defender sends $s = \bar{s}^t$ when $\omega = \omega^H$. He mixes $s = \bar{s}^t$ and $s = 0$ when $\omega = \omega^L$. The challenger challenges when $s < \bar{s}^t$ or $v_c > v_c^*(\mu_{s=\bar{s}^t})$, upon which the defender fights if and only if $\omega = \omega^H$.*

- If $\bar{s}^t \leq \frac{F\mu_0}{1-F\mu_0} \omega^L$, both types of the defender send $s = \bar{s}^t$. The challenger challenges when $s < \bar{s}^t$ or $v_c > v_c^*(\mu_0)$, upon which the defender fights if and only if $\omega = \omega^H$.

With limited audience costs, the low type cannot commit to defending if attacked, and the high type cannot fully separate. The low type tries to bluff, pretending to be a high type, but backs down when the challenger attacks. As the maximum feasible audience cost falls, equilibrium behavior moves from fully separating to semi-separating to pooling.

As the maximum feasible audience cost falls, less information is transmitted. However, the defender's payoff is generally non-linear. To see this, suppose $\bar{s}^t < \frac{F\mu_0}{1-F\mu_0} \omega^L$, so that all defender types pool on $\bar{s}^t$, and the low type will back down if attacked. Conditional on pooling behavior, higher audience costs have no informational effect and only reduce the defender payoffs, as they will be incurred by low types. By contrast, when $\bar{s}^t$ is large enough that the equilibrium is semi-separating, higher feasible audience costs reduce the probability of imitation, raising deterrence and the defender's ex ante payoffs.¹¹ When equilibrium behavior moves from semi-separating to pooling as the feasible audience cost falls, the informational channel and the direct channel coexist and push the relative payoffs in opposite directions.

Figure 3 illustrates these effects. From Proposition 5, $\mu_0^{\bar{s}^t} := \frac{F^{-1}\left(\frac{\bar{s}^t}{\omega^L + \bar{s}^t}\right)}{c_c + pF^{-1}\left(\frac{\bar{s}^t}{\omega^L + \bar{s}^t}\right)}$ is the threshold on prior belief at which the equilibrium switches from semi-separating to pooling. When $\mu_0 < \mu_0^{0.01}$, the equilibrium is semi-separating under both $\bar{s}^t = 0.01$ and 0.1 , and raising $\bar{s}^t$ increases the defender's payoffs. When $\mu_0 > \mu_0^{0.1}$, the equilibrium is pooling under both $\bar{s}^t = 0.01$ and 0.1 and raising $\bar{s}^t$ reduces the defender's payoffs. In between, the equilibrium is semi-separating with $\bar{s}^t = 0.1$ and pooling with $\bar{s}^t = 0.01$. The shaded part indicates where the payoff increases as the feasible audience cost falls, so that the equilibrium switches from semi-separating to pooling.

Having characterized the equilibria under limited audience costs, we now compare the equilibrium payoffs with those in burning-money settings.

11. We note that when mixing, the low type remains indifferent between imitating the high type and sending no signal, so its payoff does not change. But the high type benefits from the additional deterrence.

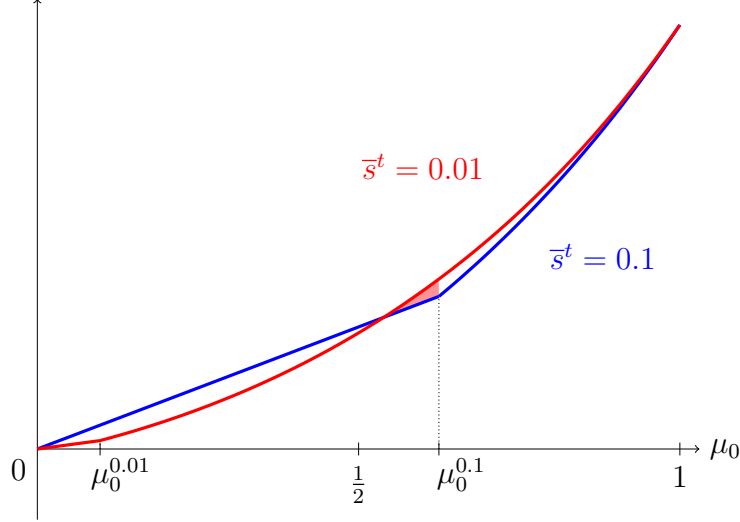


Figure 3: The defender's ex ante payoff in the audience-cost setting with $\bar{s}^t = 0.1$ (blue) and $\bar{s}^t = 0.01$ (red). Given $\bar{s}^t$, the equilibrium is semi-separating at $\mu_0 < \mu_0^{\bar{s}^t}$ and pooling at $\mu_0 > \mu_0^{\bar{s}^t}$. Other parameters are the same as in panel (b) of Figure 2.

Proposition 6 (Small Audience Costs versus Burning Money without Commitment). *Suppose Assumption 1-(i), (iii-a), and (iii-b) hold, and $\bar{s}^t < s^*$. Let $D := (1 - p)\omega^H + c_d - \omega^L$.*

- *If $\frac{F_1 D}{\delta - F_1 D} \omega^L < \bar{s}^t$, then the audience-cost setting yields a strictly higher defender ex ante payoff than the burning-money setting without commitment.*
- *If $\bar{s}^t < \frac{F_1 D}{\delta - F_1 D} \omega^L$, then there is a threshold $\bar{\mu}_0 \in (0, 1)$ such that the burning-money setting without commitment yields a strictly higher defender ex ante payoff than the audience-cost setting if and only if $\mu_0 < \bar{\mu}_0$.*

Intuitively, in the pooling equilibrium of the audience-cost setting, a lower probability of the high type (smaller μ_0) makes costly bluffing and backing down by the low type more frequent. At the same time, money is burned less frequently in the burning-money setting. Both effects therefore make the burning-money setting relatively more attractive as the high type becomes less likely. In the semi-separating equilibrium, as discussed above, a lower maximum feasible audience cost (smaller $\bar{s}^t$) reduces the defender's payoff by weakening deterrence, while leaving the burning-money setting unaffected. Once $\bar{s}^t$ falls below a threshold,

the defender's payoff in the audience cost setting falls below that in the burning-money setting. This is the intuition for the second part of Proposition 6, illustrated in Panel (b) of Figure 2. When $\bar{s}^t$ is above this threshold, as in the first part of the proposition, the audience cost setting instead yields the defender a higher payoff than the burning money setting.

Proposition 7 (Small Audience Costs versus Burning Money with Commitment). *Suppose Assumption 1-(i), (ii-b), and (ii-c) hold, and $\bar{s}^t < s^*$. For any $\mu_0 \in (0, 1)$, the defender's ex ante payoff is strictly lower in the audience-cost setting than in the burning-money setting with commitment.*

Intuitively, when the large audience costs are infeasible, the equilibria are not separating. Moreover, the low type bluffs and sometimes incurs the audience cost. Both effects reduce the value of the audience-cost setting to the defender. By contrast, in burning money with commitment, not only can the defender separate, but it can also do so at a minimal cost as long as a sufficiently small burning-money action is available.

6 Probability of War

We analyzed the equilibria and compared defender payoffs under different signaling settings. While the defender maximizes its payoffs, from a social planner's perspective, it may not internalize all the costs of war. We now compare the probability of war under different signaling settings.

Proposition 8 (Equilibrium Probability of War). *Suppose Assumption 1-(i), (ii-a), (ii-b), (iii-a), and (iii-b) hold. In equilibrium, the burning-money setting has the same ex ante probability of war with and without commitment. Moreover,*

- *When $\bar{s}^t \geq s^*$, the audience-cost setting yields a strictly greater probability of war than the burning-money setting if $-p\omega^L + c_d < \frac{F_1}{1-F_1} \omega^L$ and $F_1 < 1$. If $-p\omega^L + c_d > \frac{F_1}{1-F_1} \omega^L$ or $F_1 = 1$, the two settings have the same equilibrium probability of war.*

- When $\bar{s}^t < s^*$, the audience-cost setting yields a strictly greater probability of war than the burning-money setting.

Separating equilibria minimize the probability of war because war happens only when an undeterrable challenger attacks the high type. The equilibria of the burning-money setting have this feature. By contrast, only when both large audience costs are feasible and the low type does not want to commit to war do audience-cost settings deliver the same minimal probability of war. In all other circumstances, the probability of war is strictly higher in audience-cost settings.

When large audience costs are feasible and the low type commits to fight, this increases the probability of fighting because an undeterrable challenger will attack and will be fought not only by high types but also by low types. When large audience costs are infeasible, the low type will never fight, but because separation is not possible, challengers who would not attack a high type now attack because they believe that there is a chance that the defender is a bluffing low type pretending to be the high type.

7 Conclusion

Academics and practitioners view signaling resolve as a key element of deterrence strategies. Two primary classes of signaling are burning money and burning bridges or creating audience costs. When the defender can create sufficiently large audience costs, it effectively commits to fighting if challenged. We proposed a new signaling approach, burning money with commitment, in which the defender maintains its flexibility to fight or back down, but ex ante commits to *how* it burns money to signal resolve. We showed that this new approach can robustly improve the defender's payoffs and at the same time reduce the likelihood of war and discussed how this approach can be implemented by law and preestablished plans or automated algorithms. We also compared the payoffs of audience-cost and burning-money settings without commitment, characterizing when each dominates and highlighting the non-

monotonic relationship between the defender’s payoff and its capacity to create audience costs: the ability to create more audience costs can harm the defender.

Two directions for future research stand out. First, we focused on the canonical model of Fearon 1997. It will be beneficial to analyze settings with more general forms of payoffs than (1)-(3). In the same vein, it may be helpful to allow for information processing costs (Sims 2003; Caplin and Dean 2015; Denti 2022; Pomatto, Strack, and Tamuz 2023) so that actions that are too similar to each other may be misinterpreted by adversaries. Second, it is valuable to analyze settings with imperfect commitment in which the defender is able to alter its preestablished plan ex post or intervene in the automated algorithm with some likelihood or to some extent. These directions are left to future research.

References

- Acharya, Avidit, and Edoardo Grillo. 2019. “A Behavioral Foundation for Audience Costs.” *Quarterly Journal of Political Science* 14 (2): 159–190.
- Ashworth, Scott, and Kristopher W. Ramsay. 2024. “The accountability of politicians in international crises and the nature of audience cost.” *Political Science Research and Methods* 12 (1): 1–26.
- Banks, Jeffrey S., and Joel Sobel. 1987. “Equilibrium Selection in Signaling Games.” *Econometrica* 55 (3): 647–661.
- Boleslavsky, Raphael, Thomas Jungbauer, and Mehdi Shadmehr. 2026. “Algorithm Transparency and Search Manipulation: Steering vs. Persuasion.” Working paper. arXiv: [2608.12558](https://arxiv.org/abs/2608.12558) [econ.TH]. <https://arxiv.org/abs/2608.12558>.
- Boleslavsky, Raphael, and Mehdi Shadmehr. 2025. “Signaling with Commitment.” Working paper. <https://dx.doi.org/10.2139/ssrn.4433764>.
- Caplin, Andrew, and Mark Dean. 2015. “Revealed Preference, Rational Inattention, and Costly Information Acquisition.” *American Economic Review* 105 (7): 2183–2203.
- Carter, Ash. 2016. *Remarks on “Ensuring Continued Excellence in Defense at a Time of Strategic Transition”*. Reagan National Defense Forum, Ronald Reagan Presidential Library, Simi Valley, California. U.S. Department of Defense, December. <https://www.defense.gov/News/Speeches/Speech/Article/1019479/remarks-on-ensuring-continued-excellence-in-defense-at-a-time-of-strategic-transition/>.
- Chen, Ben. 2025. “Socially Optimal Ex-ante Adjudication.” *American Law and Economics Review*, <https://doi.org/10.1093/aler/ahaf007>.
- Crisman-Cox, Casey, and Michael Gibilisco. 2018. “Audience Costs and the Dynamics of War and Peace.” *American Journal of Political Science* 62 (3): 566–580.
- Croco, Sarah E., Michael J. Hanmer, and Jared A. McDonald. 2021. “At What Cost? Reexamining Audience Costs in Realistic Settings.” *The Journal of Politics* 83, no. 1 (January): 8–22.
- Denti, Tommaso. 2022. “Posterior Separable Cost of Information.” *American Economic Review* 112 (10): 3215–3259.
- Fearon, James D. 1997. “Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs.” *Journal of Conflict Resolution* 41, no. 1 (February): 68–90.
- Gartzke, Erik, and Yonatan Lupu. 2012. “Still Looking for Audience Costs.” *Security Studies* 21, no. 3 (July): 391–397.
- Gelpi, Christopher F., and Michael Griesdorf. 2001. “Winners or Losers? Democracies in International Crisis, 1918–94” [in en]. *American Political Science Review* 95, no. 3 (September): 633–647.

- Horz, Carlo M., and Hannah K. Simpson. 2025. "Community Interventions in the Administration of Justice." *American Journal of Political Science* 69 (3): 1047–1063.
- Hoshiro, Hiroyuki. 2026. "Thirty Years of the Audience Cost Theory: Correlations, Mechanisms, and Scientific Realism." *Political Studies Review* 24, no. 3 (August): 524–542.
- Janzen, Jay H. 2022. "The Allied Command Operations Cognitive Campaign for Deterrence and Defence." *The Three Swords*, no. 38, 18–24.
- Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review* 101, no. 6 (October): 2590–2615.
- Katagiri, Azusa, and Eric Min. 2019. "The Credibility of Public and Private Signals: A Document-Based Approach." *American Political Science Review* 113, no. 1 (February): 156–172.
- Kurizaki, Shuhei, and Taehee Whang. 2015. "Detecting Audience Costs in International Disputes." *International Organization* 69, no. 4 (October): 949–980.
- Levy, Jack S., Michael K. McKoy, Paul Poast, and Geoffrey P.R. Wallace. 2015. "Backing Out or Backing In? Commitment and Consistency in Audience Costs Theory." *American Journal of Political Science* 59 (4): 988–1001.
- Li, Xiaojun, and Dingding Chen. 2021. "Public opinion, international reputation, and audience costs in an authoritarian regime." *Conflict Management and Peace Science* 38, no. 5 (September): 543–560.
- North Atlantic Treaty Organization. 2022. *Allied Joint Doctrine*. Edition F, Version 1. NATO Standardization Office, December. https://assets.publishing.service.gov.uk/media/659ea238e96df5000df843f3/AJP_01_EdF_with_UK_elements.pdf.
- . 2026. *Deterrence and Defence*. NATO Topic. Updated June 25, 2026, June. <https://www.nato.int/en/what-we-do/deterrence-and-defence/deterrence-and-defence>.
- Pomatto, Luciano, Philipp Strack, and Omer Tamuz. 2023. "The Cost of Information: The Case of Constant Marginal Costs." *American Economic Review* 113 (5): 1360–1393.
- Reich, Noam. 2026. "When Can States Signal with Sunk Costs?" Working paper. <https://www.noamreich.com/wp-content/uploads/2026/01/SunkCosts0126.pdf>.
- Schwartz, Joshua A., and Michael C. Horowitz. 2026. "Delegating Destruction: Coercive Threats and Automated Nuclear Systems." *International Organization* 80 (1): 179–204.
- Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–690.
- Slantchev, Branislav. 2011. *Military Threats: The Costs of Coercion and the Price of Peace*. Cambridge: Cambridge University Press.
- Smith, Alastair. 1998. "International Crises and Domestic Politics." *American Political Science Review* 92 (3): 623–638.

- Snyder, Jack, and Erica D. Borghard. 2011. "The Cost of Empty Threats: A Penny, Not a Pound." *American Political Science Review* 105, no. 3 (August): 437–456.
- Stoltenberg, Jens. 2016. *The Warsaw Summit: Strengthening NATO in Turbulent Times*. Speech at Warsaw University, Warsaw, Poland. NATO Secretary General, May. <https://www.nato.int/en/news-and-events/events/transcripts/2016/05/31/the-warsaw-summit-strengthening-nato-in-turbulent-times>.
- . 2018. *Remarks at Trident Juncture*. NATO transcript. Trondheim, October 30, 2018, October. <https://www.nato.int/en/news-and-events/events/transcripts/2018/10/30/remarks>.
- . 2022. *Securing Europe*. Panel discussion at the World Economic Forum Annual Meeting, Davos, Switzerland. NATO Secretary General, May. <https://www.nato.int/en/news-and-events/events/transcripts/2022/05/24/securing-europe>.
- Takei, Makito. 2024. "Audience Costs and the Credibility of Public versus Private Threats in International Crises." *International Studies Quarterly* 68, no. 3 (September): sqae091.
- Tang, Shiping. 2026. "The Empirical Evidence of Audience Costs Theory: A Quasi-meta-analysis of Case Evidence Based on Bayesian Logic." *European Journal of International Relations* (July): 1–32.
- Tarar, Ahmer, and Bahar Leventoglu. 2013. "Limited Audience Costs in International Crises." *The Journal of Conflict Resolution* 57 (6): 1065–1089. Accessed August 23, 2026.
- Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61 (4): 821–840.
- Trachtenberg, Marc. 2012. "Audience Costs: An Historical Analysis." *Security Studies* 21, no. 1 (January): 3–42.
- Trager, Robert F., and Lynn Vavreck. 2011. "The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party." *American Journal of Political Science* 55 (3): 526–545.
- Weeks, Jessica L. 2008. "Autocratic Audience Costs: Regime Type and Signaling Resolve." *International Organization* 62, no. 1 (January): 35–64.
- Work, Robert O. 2014. *CSIS Global Security Forum*. Remarks by the Deputy Secretary of Defense, U.S. Department of Defense. Delivered at the Center for Strategic and International Studies, Washington, DC, November 12, 2014, November. <https://www.defense.gov/News/Speeches/Speech/Article/606631/csis-global-security-forum/>.

Appendix

A Additional Model Details

Here, we provide a sufficient condition for Assumption 1-(i), which requires that the defender's interim payoff is strictly convex.

Proposition 9. *The interim payoff $\widehat{v}^b(\mu, s)$ is strictly convex in $\mu \in (0, 1)$ if $\inf_x (\log(f(x)))' \geq -2(1-p)p/c_c$, where $x \in (0, \frac{c_c}{1-p})$. Moreover, this condition holds in the following cases.*

1. (Uniform) Let $v_c \sim U[0, \bar{v}_c]$ with $\bar{v}_c > \frac{c_c}{1-p}$ so that $f(x) = 1/\bar{v}_c$ and $f'(x) = 0$ for $x \in (0, \frac{c_c}{1-p})$. Then, the interim payoffs are convex.
2. (Exponential) Let $f(x) = \lambda e^{-\lambda x}$, where $x \geq 0$ and $\lambda > 0$. Then, the interim payoffs are convex if $1 \leq 2(1-p)p(1/\lambda)/c_c = 2(1-p)p \mathbb{E}[v_c]/c_c$.
3. (Lognormal) Let $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log x - m)^2}{2\sigma^2}\right\}$, where $x > 0$ and m is the mean of $\log v_c$. Then, the interim payoffs are convex if $1 \leq 2(1-p)p \sigma^2 e^{m-\sigma^2+1}/c_c = 2(1-p)p \sigma^2 e^{(1-\sigma^2/2-\sigma^2)} \mathbb{E}[v_c]/c_c$.

Proof. From equation (6), it suffices to examine $\widehat{v}^b(\mu, 0)$. Let $g(\mu) = \mu c_c/(1-\mu p)$, so that $g'(\mu) = c_c/(1-\mu p)^2 > 0$ and $g''(\mu) = 2pg'(\mu)/(1-\mu p) > 0$. The first derivative is

$$\begin{aligned} \frac{d}{d\mu} \widehat{v}^b(\mu, s=0) &= \underbrace{((1-p)\omega^H - \omega^L + c_d)}_{>0 \text{ by } W^L < 0 < W^H} \left[\underbrace{F(g(\mu))}_{\geq 0} + \underbrace{\mu f(g(\mu))g'(\mu)}_{\geq 0} \right] \\ &\quad + \omega^L f(g(\mu))g'(\mu) + \underbrace{p\omega^H - c_d}_{>0} \\ &> 0. \end{aligned}$$

Next, the second derivative is

$$\begin{aligned} \frac{d^2}{d\mu^2} \widehat{v}^b(\mu, s=0) &= (\mu((1-p)\omega^H - \omega^L + c_d) + \omega^L) [f'(g(\mu))(g'(\mu))^2 + f(g(\mu))g''(\mu)] \\ &\quad + 2(\omega^H(1-p) - \omega^L + c_d)f(g(\mu))g'(\mu). \end{aligned}$$

Rearranging this and substituting $g'(\mu) = c_c/(1-\mu p)^2$ and $g''(\mu) = 2c_c p/(1-\mu p)^3$, we have

$$\frac{d^2}{d\mu^2} \widehat{v}^b(\mu, s=0) \geq 0 \text{ if}$$

$$\frac{f'\left(\frac{\mu c_c}{1-\mu p}\right)}{f\left(\frac{\mu c_c}{1-\mu p}\right)} \geq -2(1-\mu p) \frac{p}{c_c} \left[1 + \frac{1-\mu p}{p} \frac{(1-p)\omega^H - \omega^L + c_d}{\mu((1-p)\omega^H - \omega^L + c_d) + \omega^L} \right] \quad (11)$$

for all μ . Considering the right-hand side of inequality (11), we have

$$\begin{aligned} -2 \frac{(1-p)p}{c_c} \left[1 + \frac{1-p}{p} \frac{(1-p)\omega^H - \omega^L + c_d}{(1-p)\omega^H + c_d} \right] &\leq -2 \frac{(1-p)p}{c_c} \left[1 + \frac{(1-p)^2}{p} \frac{\omega^H - \omega^L}{\omega^H} \right] \\ &\leq -\frac{2(1-p)p}{c_c}, \end{aligned}$$

where the first and second inequalities follow from $W^L < 0 < W^H$. Thus, inequality (11)

holds if $\min_x \frac{f'(x)}{f(x)} \geq -\frac{2(1-p)p}{c_c}$.

Now consider the case of a uniform distribution $v_c \sim U[0, \bar{v}_c]$ with $\bar{v}_c > \frac{c_c}{1-p}$. Then, for $x \in \left(0, \frac{c_c}{1-p}\right)$, we have $f(x) = 1/\bar{v}_c$, $f'(x) = 0$, and thus $(\log f(x))' = 0 \geq -2(1-p)p/c_c$.

Next, consider the case of the exponential distribution. Observe $\log f(x) = \log \lambda - \lambda x$ and thus $(\log f(x))' = -\lambda = \min_x (\log f(x))'$. Hence, by $\mathbb{E}[v_c] = 1/\lambda$, rearranging $\min_x (\log f(x))' = -\lambda \geq -\frac{2(1-p)p}{c_c}$ leads to

$$1 \leq \frac{2(1-p)p}{c_c} \mathbb{E}[v_c].$$

Finally, consider the case of the lognormal distribution. By $\log f(x) = -\log x - \log \sigma - \frac{1}{2} \log(2\pi) - \frac{(\log x - m)^2}{2\sigma^2}$, we have $(\log f(x))' = -\frac{1}{x} - \frac{\log x - m}{\sigma^2 x} = \frac{m - \sigma^2 - \log x}{\sigma^2 x}$, and thus $(\log f(x))'' =$

$\frac{\log x - m + \sigma^2 - 1}{\sigma^2 x^2}$. Here, note that $\lim_{x \rightarrow 0} (\log f(x))' = +\infty$ and $\lim_{x \rightarrow +\infty} (\log f(x))' = 0$. Because $(\log f(x))'$ is decreasing for $x < \exp\{m - \sigma^2 + 1\}$ and increasing for $x > \exp\{m - \sigma^2 + 1\}$, we obtain $\arg \min (\log f(x))' = \exp\{m - \sigma^2 + 1\}$ and $\min_x (\log f(x))' = -\frac{1}{\sigma^2 \exp\{m - \sigma^2 + 1\}}$. Thus, by $\mathbb{E}[v_c] = \exp\{m + \sigma^2/2\}$, substituting this into and rearranging $\min_x (\log f(x))' \geq -2(1-p)p/c_c$ yields $1 \leq \frac{2(1-p)p}{c_c} \sigma^2 e^{1-3\sigma^2/2} \mathbb{E}[v_c]$. $\square$

Moreover, we can use the same condition for the audience-cost setting for $s < -p\omega^L + c_d$.

Proposition 10. $\widehat{v}^t(\mu, s)$ for $s < -p\omega^L + c_d$ is strictly convex in μ if $\inf_{x < c_c/(1-p)} (\log f(x))' \geq -2(1-p)p/c_c$.

Proof. By $\widehat{v}^t(\mu, 0) = \widehat{v}^b(\mu, 0)$ and Proposition 9, $\widehat{v}^t(\mu, 0)$ is strictly convex if $\min_x (\log f(x))' \geq -\frac{2(1-p)p}{c_c}$. Thus, consider the case of $s \in (0, -p\omega^L + c_d)$. Using $g(\mu) = \frac{\mu c_c}{1-\mu p}$, $g'(\mu) = \frac{c_c}{(1-\mu p)^2}$, and $g''(\mu) = \frac{2c_c p}{(1-\mu p)^3}$ by Proposition 9, we can rewrite equation (9) as

$$\widehat{v}^t(\mu, s) = \mu(p\omega^H - c_d + s) - s + F(g(\mu)) [\mu(D - s) + \omega^L + s],$$

where $D = (1-p)\omega^H + c_d - \omega^L$. Differentiating this twice with respect to μ yields

$$\frac{d^2}{d\mu^2} \widehat{v}^t(\mu, s) = [\mu(D - s) + \omega^L + s] [f'(g(\mu))(g'(\mu))^2 + f(g(\mu))g''(\mu)] + 2(D - s)f(g(\mu))g'(\mu).$$

The assumption of $s < -p\omega^L + c_d$ here implies

$$\begin{aligned} D - s &> D - (-p\omega^L + c_d) \\ &= (1-p)(\omega^H - \omega^L) > 0. \end{aligned}$$

Then, because $f(g(\mu)) > 0$, $g'(\mu) > 0$, and $\mu(D - s) + \omega^L + s > 0$, the second derivative is strictly positive if

$$\frac{f'(g(\mu))}{f(g(\mu))} > -2(1-\mu p) \frac{p}{c_c} \left[1 + \frac{1-\mu p}{p} \frac{D-s}{\mu(D-s) + \omega^L + s} \right]. \quad (12)$$

The right-hand side of inequality (12) satisfies

$$\begin{aligned} -2(1-\mu p) \frac{p}{c_c} \left[1 + \frac{1-\mu p}{p} \frac{D-s}{\mu(D-s) + \omega^L + s} \right] &< -2(1-\mu p) \frac{p}{c_c} \\ &\leq -\frac{2(1-p)p}{c_c}. \end{aligned}$$

Because $g(\mu) \in (0, c_c/(1-p))$ for any $\mu \in (0, 1)$, $\min_{x \in (0, c_c/(1-p))} (\log f(x))' \geq -\frac{2(1-p)p}{c_c}$ implies inequality (12). Thus, $\hat{v}^t(\mu, s)$ is strictly convex in μ for every $s \in [0, -p\omega^L + c_d]$. $\square$

Next, we show that the right-hand side of Assumption 1-(ii-b) is strictly positive when Assumption 1-(i) holds.

Remark 1. Under Assumption 1-(i) we have $\left(F_1 - F\left(\frac{c_c}{2-p}\right)\right) \delta - F\left(\frac{c_c}{2-p}\right) \omega^L > 0$.

Proof. By Assumption 1-(i), $\hat{v}^t(\mu, 0) = \hat{v}^b(\mu, 0)$ is strictly convex in μ . Thus, $\hat{v}^b(\frac{1}{2}, 0) < \frac{1}{2}\hat{v}^b(0, 0) + \frac{1}{2}\hat{v}^b(1, 0)$. Because $\hat{v}^b(0, 0) = 0$ from expression (6) and $F(0) = 0$, this is equivalent to $2\hat{v}^b(\frac{1}{2}, 0) < \hat{v}^b(1, 0)$. By expression (6), we obtain

$$2\hat{v}^b(1/2, 0) = F(v_c^*(1/2))(\omega^H + \omega^L) + (1 - F(v_c^*(1/2)))(p\omega^H - c_d) \quad (13)$$

$$\hat{v}^b(1, 0) = F_1 \omega^H + (1 - F_1)(p\omega^H - c_d). \quad (14)$$

Substituting $v_c^*(1/2) = \frac{c_c}{2-p}$ into (13) and (14), we have: $2\hat{v}^b(\frac{1}{2}, 0) < \hat{v}^b(1, 0)$ if and only if $\left(F_1 - F\left(\frac{c_c}{2-p}\right)\right) \delta - F\left(\frac{c_c}{2-p}\right) \omega^L > 0$. $\square$

B Proofs

Proof of Proposition 1. Consider the following cases.

Step 1 (Separating). First, consider the case in which the high type of the defender sends $s = s'$ and the low type sends $s = s''$, with off-path beliefs $\mu_{s < s'} = 0$ and $\mu_{s > s'} = 1$. Observe that, in any separating equilibrium, we must have $s'' = 0$. Otherwise, the low type would

receive a strictly negative payoff and have an incentive to deviate to $s = 0$. The low type does not have the incentive to deviate to s' when

$$\begin{aligned} 0 &\geq F_1\omega^L + (1 - F_1) \times 0 - s' \\ s' &\geq F_1\omega^L. \end{aligned}$$

Because of $\mu_{s < s'} = 0$, the low type does not deviate to any $s \in (0, s')$. Also, while $\mu_{s > s'} = 1$, the low type does not have the incentive to deviate to $s > s'$ given $F_1\omega^L - s < F_1\omega^L - s' \leq 0$.

Under the above off-path beliefs, the high type does not have the incentive to deviate to a larger s because $\mu_{s=s'}$ is already 1. Because $\mu_s = 0$ for any $s < s'$, if there is a profitable deviation from s' , it must be $s = 0$. The high type does not benefit from deviating to $s = 0$ when

$$\begin{aligned} F_1\omega^H + (1 - F_1)(p\omega^H - c_d) - s' &\geq p\omega^H - c_d \\ s' &\leq F_1\delta. \end{aligned}$$

Thus, s' must be in the interval $[F_1\omega^L, F_1\delta]$. By $W^L < 0 < W^H$, we have $\delta > \omega^L$, and thus, this interval is nonempty. Also, observe that the high type's payoff is decreasing in s' . Thus, the optimal signal is $s' = F_1\omega^L$, which is in S^b by Assumption 1-(iii-a).

For any off-path signal $s \in (0, F_1\omega^L)$, the D1 criterion assigns probability zero to the high type, and thus requires $\mu_s = 0$. For $s > F_1\omega^L$, neither type can benefit even under $\mu_s = 1$, and thus the criterion imposes no restriction. Thus, the off-path beliefs supporting the separating equilibrium satisfy the D1 criterion.

Step 2 (Pooling on $s = 0$). Next, consider a pooling equilibrium where both types send $s = 0$ with off-path beliefs $\mu_{s > 0} \leq \mu_0$. Under the pessimistic off-path beliefs, neither type has the incentive to deviate to strictly positive s . Thus, we examine the D1 criterion. Observe that we can express the challenger's mixed best response as the probability of not challenging, $F(v_c^*(\mu_s)) \in [0, F_1]$. Because the equilibrium payoff and the payoff from deviating to an

off-path signal s for the low type are $F_{\mu_0}\omega^L$ and $F(v_c^*(\mu_s))\omega^L - s$, respectively, the low type has a weak incentive to deviate when

$$F(v_c^*(\mu_s)) \geq F_{\mu_0} + \frac{s}{\omega^L}.$$

On the other hand, the high type's equilibrium payoff and the payoff from deviation are $F_{\mu_0}\omega^H + (1 - F_{\mu_0})(p\omega^H - c_d)$ and $F(v_c^*(\mu_s))\omega^H + (1 - F(v_c^*(\mu_s)))(p\omega^H - c_d) - s$, respectively. Thus, the high type has a strict incentive to deviate to an off-path $s > 0$ when

$$F(v_c^*(\mu_s)) > F_{\mu_0} + \frac{s}{\delta}.$$

Because $F_{\mu_0} + s/\omega^L > F_{\mu_0} + s/\delta$ by $W^L < 0 < W^H$, the high type has a strict incentive to deviate whenever the low type has a weak incentive to deviate. The D1 criterion requires $\mu_s = 1$ for such s . Because this off-path signal yields $F(v_c^*(\mu_s)) = F_1$, the high type gains $(p\omega^H - c_d + F_1\delta - s) - (p\omega^H - c_d + F_{\mu_0}\delta) = \Delta\delta - s$ from deviating, where $\Delta = (F_1 - F_{\mu_0})$. Thus, the high type would deviate to any feasible $s \in (0, \Delta\delta)$. By Assumption 1-(iii-b), there exists a feasible signal in the interval. Hence, this equilibrium violates the D1 criterion.

Step 3 (Pooling on $s > 0$). Consider a case in which both types of the defender send $s = s' > 0$ with off-path beliefs $\mu_{s \neq s'} = 0$. Given the off-path beliefs, the low type does not have the incentive to deviate to $s = 0$ when $F_{\mu_0}\omega^L - s' \geq 0$, i.e., $s' \leq F_{\mu_0}\omega^L$. Similarly, the condition for the high type is $F_{\mu_0}\omega^H + (1 - F_{\mu_0})(p\omega^H - c_d) - s' \geq p\omega^H - c_d$, i.e., $s' \leq F_{\mu_0}((1 - p)\omega^H + c_d)$, which is always true whenever the low type does not deviate. For a given prior μ_0 , this pooling equilibrium exists when $s' \leq F_{\mu_0}\omega^L$ because the low type would otherwise deviate to $s = 0$.

We now examine the D1 criterion. The low type's equilibrium payoff and the payoff from deviating to some $s > s'$ are $F_{\mu_0}\omega^L - s'$ and $F(v_c^*(\mu_s))\omega^L - s$, respectively. Hence, the low

type is weakly incentivized to deviate to such s when

$$F(v_c^*(\mu_s)) \geq F_{\mu_0} + \frac{s - s'}{\omega^L}.$$

The high type's equilibrium payoff and the payoff from deviating to $s > s'$ are, respectively, $F_{\mu_0}\omega^H + (1 - F_{\mu_0})(p\omega^H - c_d) - s'$ and $F(v_c^*(\mu_s))\omega^H + (1 - F(v_c^*(\mu_s)))(p\omega^H - c_d) - s$. Thus, a deviation to $s > s'$ is strictly beneficial for the high type when

$$F(v_c^*(\mu_s)) > F_{\mu_0} + \frac{s - s'}{\delta}.$$

By $W^L < 0 < W^H$, we know that $F_{\mu_0} + (s - s')/\omega^L > F_{\mu_0} + (s - s')/\delta$, and thus, the high type has a strict incentive to deviate to $s > s'$ whenever the low type has a weak incentive to do so. The D1 criterion assigns $\mu_s = 1$ to such s . Because we have $F(v_c^*(\mu_s)) = F_1$ for such s , the high type's benefit from deviating from $s' > 0$ to $s > s'$ is $\Delta\delta - s + s'$. Therefore, the high type has a strictly profitable deviation to any feasible $s \in (s', s' + \Delta\delta)$. Assumption 1-(iii-b) guarantees that the interval has a feasible signal for any $s' \in (0, F_{\mu_0}\omega^L]$. Thus, a pooling equilibrium with $s' > 0$ violates the D1 criterion.

Step 4 (Semi-separating). Consider a case in which the high type sends $s = s' > 0$, and the low type sends $s = s'$ with probability $q \in (0, 1)$ and $s = 0$ with probability $1 - q$ with off-path beliefs $\mu_{s < s'} = 0$ and $\mu_{s > s'} = 1$. It is easy to see that the signal only from the low type must be $s = 0$ because the posterior is zero; for any positive s , the low type would deviate to a smaller s . We show that a semi-separating equilibrium in which the low type mixes violates the D1 criterion.

Because the low type's indifference condition between sending $s = s'$ and $s = 0$ is $F(v_c^*(\mu_{s=s'}))\omega^L - s' = 0$, we obtain $s' = F(v_c^*(\mu_{s=s'}))\omega^L$. Also observe that, by Bayes' rule, the posterior $\mu_{s=s'} = \frac{\mu_0}{\mu_0 + (1 - \mu_0)q} \in (\mu_0, 1)$ implies $s' = F(v_c^*(\mu_{s=s'}))\omega^L \in (F_{\mu_0}\omega^L, F_1\omega^L)$. Consider an off-path signal $s = F_1\omega^L$, which is feasible by Assumption 1-(iii-a). The low

type has a weakly profitable deviation to $s = F_1\omega^L$ when

$$F(v_c^*(\mu_{s=F_1\omega^L})) \geq F(v_c^*(\mu_{s=s'})) + \frac{F_1\omega^L - s'}{\omega^L}$$

because the equilibrium payoff and the payoff from deviation are $F(v_c^*(\mu_{s=s'}))\omega^L - s'$ and $F(v_c^*(\mu_{s=F_1\omega^L}))\omega^L - F_1\omega^L$, respectively. The high type's equilibrium payoff and the payoff from deviation are $F(v_c^*(\mu_{s=s'}))\omega^H + (1 - F(v_c^*(\mu_{s=s'})))(p\omega^H - c_d) - s'$ and $F(v_c^*(\mu_{s=F_1\omega^L}))\omega^H + (1 - F(v_c^*(\mu_{s=F_1\omega^L}))(p\omega^H - c_d) - F_1\omega^L$. Hence, the high type has a strictly profitable deviation to $s = F_1\omega^L$ when

$$F(v_c^*(\mu_{s=F_1\omega^L})) > F(v_c^*(\mu_{s=s'})) + \frac{F_1\omega^L - s'}{\delta}.$$

Because we know $s' < F_1\omega^L$ and $\delta > \omega^L$ by $W^L < 0 < W^H$, we have $F(v_c^*(\mu_{s=s'})) + \frac{F_1\omega^L - s'}{\omega^L} > F(v_c^*(\mu_{s=s'})) + \frac{F_1\omega^L - s'}{\delta}$. Thus, the high type has a strictly profitable deviation to $s = F_1\omega^L$ whenever the low type has a weakly profitable deviation. Thus, by the D1 criterion, we obtain $\mu_{s=F_1\omega^L} = 1$. Then, the high type's profit from deviating to $s = F_1\omega^L$ is

$$(F_1 - F(v_c^*(\mu_{s=s'})))\delta - F_1\omega^L + s' = (F_1 - F(v_c^*(\mu_{s=s'}))) (\delta - \omega^L) > 0.$$

Because the high type has an incentive to deviate, a semi-separating equilibrium in which the low type mixes violates the D1 criterion.

Next, consider a semi-separating equilibrium in which the low type sends $s = s'$, and the high type sends $s = s'$ with probability $q \in (0, 1)$ and $s = s''$ with probability $1 - q$, with off-path beliefs $\mu_{s < s'} = 0$, $\mu_{s \in (s', s'')} = \mu_{s=s'}$, and $\mu_{s > s''} = 1$. We show that a semi-separating equilibrium in which the high type mixes violates the D1 criterion.

By Bayes' rule, we have $\mu_{s=s'} = \frac{\mu_0 q}{\mu_0 q + (1 - \mu_0)} < \mu_0$ and thus $F(v_c^*(\mu_{s=s'})) < F_{\mu_0}$. The

indifference condition of the high type yields

$$\begin{aligned} F(v_c^*(\mu_{s=s'}))\omega^H + (1 - F(v_c^*(\mu_{s=s'}))) (p\omega^H - c_d) - s' &= F_1\omega^H + (1 - F_1) (p\omega^H - c_d) - s'' \\ s'' &= s' + (F_1 - F(v_c^*(\mu_{s=s'}))) \delta. \end{aligned}$$

We consider an off-path signal $s \in (s', s'')$. Assumption 1-(iii-b) guarantees that there exists a feasible signal in the interval (s', s'') . To see this, observe that $s' \leq F(v_c^*(\mu_{s=s'}))\omega^L < F_{\mu_0}\omega^L$ by $F(v_c^*(\mu_{s=s'}))\omega^L - s' \geq 0$ and $F(v_c^*(\mu_{s=s'})) < F_{\mu_0}$. Also observe that $s'' = s' + (F_1 - F(v_c^*(\mu_{s=s'}))) \delta > s' + \Delta\delta$ by $F(v_c^*(\mu_{s=s'})) < F_{\mu_0}$. Thus, we have $(s', s' + \Delta\delta) \subset (s', s'')$. Then, there is some s in (s', s'') by Assumption 1-(iii-b).

The low type's equilibrium payoff and the payoff from deviating to $s \in (s', s'')$ are $F(v_c^*(\mu_{s=s'}))\omega^L - s'$ and $F(v_c^*(\mu_{s \in (s', s'')}))\omega^L - s$, respectively. Thus, the low type has a weak incentive to deviate to such s when

$$F(v_c^*(\mu_{s \in (s', s'')})) \geq F(v_c^*(\mu_{s=s'})) + \frac{s - s'}{\omega^L}.$$

On the other hand, the high type's equilibrium payoff and the payoff from deviation are respectively

$$\begin{aligned} &F(v_c^*(\mu_{s=s'}))\omega^H + (1 - F(v_c^*(\mu_{s=s'}))) (p\omega^H - c_d) - s' \text{ and} \\ &F(v_c^*(\mu_{s \in (s', s'')}))\omega^H + (1 - F(v_c^*(\mu_{s \in (s', s'')}))) (p\omega^H - c_d) - s. \end{aligned}$$

Hence, the high type has a strictly profitable deviation to $s \in (s', s'')$ when

$$F(v_c^*(\mu_{s \in (s', s'')})) > F(v_c^*(\mu_{s=s'})) + \frac{s - s'}{\delta}.$$

By $\delta > \omega^L$ and $s > s'$, we have $F(v_c^*(\mu_{s=s'})) + \frac{s-s'}{\omega^L} > F(v_c^*(\mu_{s=s'})) + \frac{s-s'}{\delta}$. Hence, whenever the low type has a weakly profitable deviation to $s \in (s', s'')$, the high type has a strictly profitable

deviation to such s . This yields $\mu_{s \in (s', s'')} = 1$. By deviating to $s \in (s', s'')$, the high type gains

$$(F_1 - F(v_c^*(\mu_{s=s'})))\delta - (s - s').$$

This is strictly positive by $s < s'' = s' + (F_1 - F(v_c^*(\mu_{s=s'})))\delta$. Thus, because the high type deviates to $s \in (s', s'')$, a semi-separating equilibrium in which the high type mixes violates the D1 criterion.

Finally, consider a case where both types send $s = s'$ and $s = s''$ with positive probability, where $s' < s''$. Suppose that such an equilibrium exists. Then, the indifference condition of the low type requires $F(v_c^*(\mu_{s=s'}))\omega^L - s' = F(v_c^*(\mu_{s=s''}))\omega^L - s''$, and thus

$$s'' - s' = (F(v_c^*(\mu_{s=s''})) - F(v_c^*(\mu_{s=s'})))\omega^L. \quad (15)$$

Next, the high type's indifference condition yields

$$F(v_c^*(\mu_{s=s'}))\omega^H + (1 - F(v_c^*(\mu_{s=s'})))W^H - s' = F(v_c^*(\mu_{s=s''}))\omega^H + (1 - F(v_c^*(\mu_{s=s''})))W^H - s''$$

$$s'' - s' = (F(v_c^*(\mu_{s=s''})) - F(v_c^*(\mu_{s=s'})))\delta. \quad (16)$$

By $\delta > \omega^L$ and equalities (15) and (16), the two indifference conditions cannot hold simultaneously, a contradiction.

Step 5 (defender-optimal equilibrium). Because only the separating equilibrium in Step 1 satisfies our equilibrium refinement, the D1 criterion, it is the defender-optimal equilibrium. $\square$

Proof of Proposition 2. We first specify the signal structure in the case with commitment power. Then, we show that the defender's ex ante payoff under this signal structure is the highest possible payoff in the setting, and thus commitment power is strictly beneficial for the defender.

Step 1 (Signal structure). First, we specify the signal structure in the case with commitment power and the ex ante payoff of the defender. When μ_0 is small, we have $\mu_{s=0} = 0, \mu_{s=s_1^b} = 1$. Bayes-Plausibility requires $\tau(\mu_{s=0})\mu_{s=0} + (1 - \tau(\mu_{s=0}))\mu_{s=s_1^b} = \mu_0$, hence

$$\tau(\mu_{s=0}) = 1 - \mu_0 \text{ and } \tau(\mu_{s=s_1^b}) = \mu_0.$$

The defender's signal structure follows from the above values:

$$\pi(0|\omega^H) = 0, \pi(s = s_1^b|\omega^H) = 1, \pi(0|\omega^L) = 1, \text{ and } \pi(s = s_1^b|\omega^L) = 0. \quad (17)$$

When μ_0 is large, we have $\mu_{s=0} = 1$ and $\mu_{s=s_1^b} = 0$. Bayes-Plausibility yields

$$\tau(\mu_{s=0}) = \mu_0 \text{ and } \tau(\mu_{s=s_1^b}) = 1 - \mu_0.$$

Hence, we obtain

$$\pi(0|\omega^H) = 1, \pi(s_1^b|\omega^H) = 0, \pi(0|\omega^L) = 0, \text{ and } \pi(s_1^b|\omega^L) = 1. \quad (18)$$

By (17), the ex ante payoff when μ_0 is small is

$$\mu_0 [F_1\omega^H + (1 - F_1)W^H - s_1^b] + (1 - \mu_0) \times 0 = \mu_0 (\widehat{v}^b(1, 0) - s_1^b),$$

whereas the payoff when μ_0 is large is

$$\mu_0 [F_1\omega^H + (1 - F_1)W^H] - (1 - \mu_0)s_1^b = \mu_0\widehat{v}^b(1, 0) - (1 - \mu_0)s_1^b.$$

The above payoffs become equal at $\mu_0 = 1/2$, and thus, the ex ante defender payoff in the

burning-money setting with commitment is expressed as

$$\mu_0 \widehat{v}^b(1, 0) - \min \{ \mu_0, 1 - \mu_0 \} s_1^b = V^{jo}(\mu_0 | (\widehat{v}_s^b)_{s \in S^b}) . \quad (19)$$

Step 2 (Payoff optimality). We want to show that $V^{jo}(\mu_0 | (\widehat{v}_s^b)_{s \in S^b})$ in (19) is the highest possible payoff the defender can achieve. That is, we want to show

$$\sum_{s \in S^b} \tau(\mu_s) \widehat{v}^b(\mu_s, s) \leq V^{jo}(\mu_0 | (\widehat{v}_s^b)_{s \in S^b})$$

for any Bayes-plausible signal structure. First, consider any $s > 0$ and $\mu \in [0, 1]$. Then, by Assumption 1-(i), we obtain

$$\widehat{v}^b(\mu, s > 0) = \widehat{v}^b(\mu, 0) - s \leq \mu \widehat{v}^b(1, 0) + (1 - \mu) \underbrace{\widehat{v}^b(0, 0)}_{=0} - s_1^b.$$

Next, consider $s = 0$ and $\mu_{s=0} \leq 1/2$. First, by Assumption 1-(ii-b), observe

$$\begin{aligned} 2\widehat{v}^b\left(\frac{1}{2}, 0\right) &= 2 \times \left(\frac{1}{2} (W^H + F_{0.5}\delta) + \frac{1}{2} F_{0.5}\omega^L \right) \\ &= W^H + F_1\delta - [(F_1 - F_{0.5})\delta - F_{0.5}\omega^L] \\ &< \widehat{v}^b(1, 0) - s_1^b. \end{aligned} \quad (20)$$

Thus, for any $\mu_{s=0} \leq 1/2$, we have

$$\widehat{v}^b(\mu_{s=0}, 0) \leq 2\mu_{s=0} \widehat{v}^b\left(\frac{1}{2}, 0\right) + (1 - 2\mu_{s=0}) \widehat{v}^b(0, 0) \leq \mu_{s=0} (\widehat{v}^b(1, 0) - s_1^b),$$

by Assumption 1-(i) and $\widehat{v}^b(0, 0) = 0$. Then, by Bayes-plausibility, we obtain the ex ante

payoff

$$\begin{aligned}
\sum_{s \in S^b} \tau(\mu_s) \widehat{v}^b(\mu_s, s) &\leq \sum_{s \in S^b} \tau(\mu_s) \mu_s (\widehat{v}^b(1, 0) - s_1^b) \\
&= \mu_0 (\widehat{v}^b(1, 0) - s_1^b).
\end{aligned} \tag{21}$$

Lastly, consider $s = 0$ and $\mu_{s=0} \geq 1/2$. By Assumption 1-(i) and inequality (20), we have

$$\begin{aligned}
\widehat{v}^b(\mu_{s=0}, 0) &\leq 2(1 - \mu_{s=0}) \widehat{v}^b\left(\frac{1}{2}, 0\right) + (2\mu_{s=0} - 1) \widehat{v}^b(1, 0) \\
&\leq (1 - \mu_{s=0}) (\widehat{v}^b(1, 0) - s_1^b) + (2\mu_{s=0} - 1) \widehat{v}^b(1, 0) \\
&= \mu_{s=0} \widehat{v}^b(1, 0) - (1 - \mu_{s=0}) s_1^b.
\end{aligned}$$

Then, by Bayes-plausibility, we obtain the ex ante payoff

$$\begin{aligned}
\sum_{s \in S^b} \tau(\mu_s) \widehat{v}^b(\mu_s, s) &\leq \sum_{s \in S^b} \tau(\mu_s) (\mu_s \widehat{v}^b(1, 0) - (1 - \mu_s) s_1^b) \\
&= \mu_0 \widehat{v}^b(1, 0) - (1 - \mu_0) s_1^b.
\end{aligned} \tag{22}$$

In sum, every Bayes-plausible signal structure satisfies at least one of (21) and (22). Hence, we have

$$\begin{aligned}
\sum_{s \in S^b} \tau(\mu_s) \widehat{v}^b(\mu_s, s) &\leq \max \{ \mu_0 (\widehat{v}^b(1, 0) - s_1^b), \mu_0 \widehat{v}^b(1, 0) - (1 - \mu_0) s_1^b \} \\
&= V^{jo}(\mu_0 | (\widehat{v}_s^b)_{s \in S^b})
\end{aligned}$$

for any Bayes-plausible signal structure.

Finally, the above signal structures strictly improve the defender's ex ante payoff, compared to the case without commitment. We already know that the ex ante payoff is bounded by $V^{jo}(\mu_0 | (\widehat{v}_s^b)_{s \in S^b})$. Then, we show that the case without commitment power cannot attain it and, thus, yields a strictly smaller payoff. First, suppose $\mu_0 \leq 1/2$. The defender obtains

the ex ante payoff $\mu_0 (\hat{v}^b(1, 0) - s_1^b)$ by sending $s = s_1^b$ when $\omega = \omega^H$ and sending $s = 0$ when $\omega = \omega^L$. If the defender does not have commitment power, the low type can profitably deviate from $s = 0$ to s_1^b because $F_1 \omega^L - s_1^b > 0$ by Assumption 1-(ii-a). Second, suppose $\mu_0 \geq 1/2$. The defender achieves the ex ante payoff $\mu_0 \hat{v}^b(1, 0) - (1 - \mu_0) s_1^b$ by sending $s = 0$ when $\omega = \omega^H$ and sending $s = s_1^b$ when $\omega = \omega^L$. If the defender does not have commitment power, then the low type can profitably deviate from $s = s_1^b > 0$ to $s = 0$. In sum, the signal structures in (17) and (18) cannot be part of an equilibrium in the burning-money setting without commitment. Thus, the case without commitment power does not yield the same ex ante payoff as the setting with commitment. $\square$

Proof of Proposition 3. We first consider a special case of $F_1 = 1$. Because we have adopted the interpretation that $1/(1 - F(x)) = \infty$ when $F(x) = 1$, we have $-p\omega^L + c_d < \frac{F_1}{1-F_1} \omega^L = \infty$. Because this implies that the challenger never challenges if the defender fights for sure, both types of the defender send some $s \geq -p\omega^L + c_d$. Consequently, the high and low types receive ω^H and ω^L , and they do not have the incentive to deviate to any $s < -p\omega^L + c_d$. Moreover, because ω^H and ω^L are the defender's highest possible payoffs, this equilibrium satisfies the D1 criterion.

Henceforth, we focus on the case of $F_1 < 1$ and consider the following cases.

Case 1 ($-p\omega^L + c_d \geq \frac{F_1}{1-F_1} \omega^L$). Consider a separating equilibrium where the high-valuation defender sends $s = s'$ and the low type sends $s = s''$. As shown in Step 1 of the proof of Proposition 1, we have $s'' = 0$ in any separating equilibrium. Because $\mu_{s=s'} = 1$ and $\mu_{s=0} = 0$, the low type does not have the incentive to deviate to $s' > 0$ as long as

$$\underbrace{0}_{\text{Equilibrium payoff}} \geq \underbrace{F_1 \omega^L - (1 - F_1) s'}_{\text{Deviation to } s=s'}$$

$$s' \geq \frac{F_1}{1 - F_1} \omega^L.$$

Observe that the equilibrium payoff for the low type is zero because, by $F(0) = 0$, the chal-

lenger challenges for sure when $\mu_{s=0} = 0$. When $-p\omega^L + c_d \geq \frac{F_1}{1-F_1}\omega^L$, the low type does not have the incentive to deviate to $s \geq -p\omega^L + c_d$ because the inequality is equivalent to $0 \geq F_1\omega^L + (1-F_1)(p\omega^L - c_d)$. Next, the high type's payoff from sending $s = s'$ is $F_1\omega^H + (1-F_1)(p\omega^H - c_d)$, which does not depend on s' . Thus, any feasible $s' \geq \frac{F_1}{1-F_1}\omega^L$ is optimal for the high type. Our assumption $\bar{s}^t \geq s^* = \min \left\{ -p\omega^L + c_d, \frac{F_1}{1-F_1}\omega^L \right\}$ guarantees that such s' is feasible when $-p\omega^L + c_d \geq \frac{F_1}{1-F_1}\omega^L$. Off-path beliefs $\mu_{s < \frac{F_1}{1-F_1}\omega^L} = 0$ and $\mu_{s \geq \frac{F_1}{1-F_1}\omega^L} = 1$ support this equilibrium.

This equilibrium satisfies the D1 criterion. First, consider an off-path signal $s < \frac{F_1}{1-F_1}\omega^L$. The high type's payoff from deviation under the most favorable posterior is $F_1\omega^H + (1-F_1)(p\omega^H - c_d)$, which is equal to his equilibrium payoff. On the other hand, the low type would receive $F_1\omega^L - (1-F_1)s > 0$ under the most favorable posterior. Thus, whenever the high type has a weakly profitable deviation, the low type has a strictly profitable deviation. Thus, the D1 criterion assigns $\mu_s = 0$ for such an off-path signal. The challenger would always challenge after $s < \frac{F_1}{1-F_1}\omega^L$, lowering both types' payoffs. Second, consider an off-path signal $s \in \left(\frac{F_1}{1-F_1}\omega^L, -p\omega^L + c_d \right)$. As in the first case, this deviation does not improve the high type's payoff. For the low type, a deviation to $s \in \left(\frac{F_1}{1-F_1}\omega^L, -p\omega^L + c_d \right)$ would yield at most $F_1\omega^L - (1-F_1)s < 0$. Hence, neither type can have a strict incentive to deviate. Third, consider an off-path signal $s \geq -p\omega^L + c_d \geq \frac{F_1}{1-F_1}\omega^L$. This deviation never strictly improves the defender's payoff. In sum, the separating equilibrium satisfies the D1 criterion.

We show that, when $c_d - p\omega^L > \frac{F_1}{1-F_1}\omega^L$, any equilibrium satisfying the D1 criterion must take the form of the above equilibrium, (a) which is a separating equilibrium, (b) in which the low type sends $s = 0$, and (c) in which the high type sends some $s \geq \frac{F_1}{1-F_1}\omega^L$. First, suppose to the contrary that feature (a) does not hold: consider a semi-separating or a pooling equilibrium. This implies that some common s is sent by both types with positive probability. Observe that this common signal must satisfy $s < \frac{F_1}{1-F_1}\omega^L$. The low type never sends $s > \frac{F_1}{1-F_1}\omega^L$. Also, the low type does not send $s = \frac{F_1}{1-F_1}\omega^L$, either, because a common signal induces $\mu_s < 1$, which gives the low type a strictly negative payoff. Because the low

type does not send $s \geq \frac{F_1}{1-F_1} \omega^L$, a common signal must be smaller than $\frac{F_1}{1-F_1} \omega^L$. Next, fix some feasible $s' \geq \frac{F_1}{1-F_1} \omega^L$. Because the low type never sends such s' , we have $\mu_{s=s'} = 1$ if the high type sends it by Bayes' rule, and the high type would achieve his highest possible payoff. However, an equilibrium that is not separating implies that some common s is sent by both types with positive probability. This common signal leads to $\mu_s < 1$, implying that the high type's payoff becomes lower than the highest possible one. Thus, the high type cannot send s' by the indifference condition. Thus, s' must be an off-path signal when feature (a) does not hold. From this large off-path signal s' , the low type never benefits even under the most favorable posterior, $\mu_{s=s'} = 1$. On the other hand, the high type would strictly benefit from such s' . Then, the D1 criterion imposes $\mu_{s=s'} = 1$ where $s' \geq \frac{F_1}{1-F_1} \omega^L$, and the high type deviates to the large signal, a contradiction.

Second, suppose that feature (b) does not hold: consider a separating equilibrium where the low type sends $s > 0$. If $s < -p\omega^L + c_d$, he would receive $-s < 0$. If $s \geq -p\omega^L + c_d$, he would receive $F_1\omega^L + (1 - F_1)(p\omega^L - c_d) < 0$. Thus, this case cannot constitute an equilibrium because the low type has the incentive to deviate to $s = 0$.

Third, suppose that feature (c) does not hold: consider a separating equilibrium where the high type sends $s < \frac{F_1}{1-F_1} \omega^L$. Then, because $\mu_s = 1$, the low type has the incentive to imitate and receive $F_1\omega^L - (1 - F_1)s > 0$. In sum, any equilibrium must take the form of the above separating equilibrium, implying that it is defender-optimal.

Case 2 ($-p\omega^L + c_d \leq \frac{F_1}{1-F_1} \omega^L$). Next, consider a case in which both types of the defender send some $s \geq -p\omega^L + c_d$ with off-path beliefs $\mu_{s < -p\omega^L + c_d} = 0$, where both types are willing to fight on the equilibrium path. By sending some $s \geq c_d - p\omega^L$, the high type obtains the highest possible payoff $F_1\omega^H + (1 - F_1)(p\omega^H - c_d)$ because the high type never pays the cost of the signal. Now the low type is also willing to fight and does not pay audience costs. Sending $s \geq c_d - p\omega^L$ generates a weakly higher payoff for the low type than sending $s = 0$

and then acquiescing when

$$\begin{aligned} F_1 \omega^L + (1 - F_1) (p \omega^L - c_d) &\geq 0 \\ -p \omega^L + c_d &\leq \frac{F_1}{1 - F_1} \omega^L, \end{aligned}$$

which is our assumption here.

This equilibrium satisfies the D1 criterion. First, any off-path signal such that $s \geq -p \omega^L + c_d$ does not change the defender's payoff. Second, consider an off-path signal $s < -p \omega^L + c_d$. The high type would receive at most his equilibrium payoff from this deviation. Under the most favorable belief, the low type would receive $F_1 \omega^L - (1 - F_1)s > 0$. The benefit for the low type from this deviation is

$$(F_1 \omega^L - (1 - F_1)s) - (F_1 \omega^L + (1 - F_1)(p \omega^L - c_d)) = (1 - F_1)(c_d - p \omega^L - s) > 0.$$

Thus, whenever the high type weakly prefers deviation to $s < -p \omega^L + c_d$, the low type has a strict incentive to deviate to such s . The D1 criterion imposes $\mu_s = 0$. This implies that the high type and the low type would receive $p \omega^H - c_d$ and $-s$, respectively, by deviating to $s < -p \omega^L + c_d$, both of which are strictly smaller than their equilibrium payoffs. Thus, this equilibrium satisfies the D1 criterion.

There is no other equilibrium with a higher defender payoff. In any equilibrium, the high type must achieve the highest possible payoff, $F_1 \omega^H + (1 - F_1)(p \omega^H - c_d)$, because he can always commit to fighting and receive this payoff. First, the two types of the defender never send a common $s < -p \omega^L + c_d$ with positive probability. This is because such pooling leads to $\mu_s < 1$, which lowers the high type's payoff. Second, given the first argument, the low type sending $s < -p \omega^L + c_d$ cannot be part of an equilibrium because it would yield $\mu_s = 0$ and the low type would receive $-s \leq 0$, which is strictly smaller than the payoff he can guarantee by committing to fighting. Thus, the low type sends only $s \geq -p \omega^L + c_d$. Lastly, consider any case where the high type sends $s < -p \omega^L + c_d$ with positive probability and the low type

sends only $s \geq -p\omega^L + c_d$. Such an equilibrium does not exist because the low type would profitably deviate to the high type's signal, as $F_1\omega^L - (1 - F_1)s > F_1\omega^L + (1 - F_1)(p\omega^L - c_d)$ whenever $s < -p\omega^L + c_d$. Thus, the remaining candidate is the above case in which both types send only $s \geq -p\omega^L + c_d$.

Case 3 ($-p\omega^L + c_d = \frac{F_1}{1-F_1}\omega^L$). When $-p\omega^L + c_d = \frac{F_1}{1-F_1}\omega^L$, both types of the above equilibria are defender-optimal. Moreover, in this knife-edge case, there is another class of equilibria. Consider a case in which the high type sends some $s' \geq -p\omega^L + c_d$ and the low type sends $s = s'$ with probability q and $s = 0$ with probability $1-q$. While $\mu_{s=s'} \in [0, 1]$ should follow from Bayes' rule, we know that both types commit to fighting by sending $s' \geq -p\omega^L + c_d$. This implies that the challenger with $v_c > \frac{c_c}{1-p}$ challenges and that the probability of challenge is $1 - F_1$. Consequently, the low type's payoff from sending $s = s'$ when $-p\omega^L + c_d \leq \frac{F_1}{1-F_1}\omega^L$ is

$$F_1\omega^L + (1 - F_1)(p\omega^L - c_d) = F_1\omega^L - (1 - F_1)\frac{F_1}{1 - F_1}\omega^L = 0.$$

Because the payoff from sending $s = 0$ is zero, we have established the low type's indifference condition. The low type may use an arbitrary $q \in [0, 1]$. Similarly, by committing to fight, the high type obtains the highest possible payoff, $F_1\omega^H + (1 - F_1)(p\omega^H - c_d)$, regardless of q .

This class of equilibrium satisfies the D1 criterion. Consider an off-path signal $s \in (0, -p\omega^L + c_d)$. First, the high type cannot strictly improve his payoff. Second, because the low type would acquiesce to a challenge after sending $s \in (0, -p\omega^L + c_d)$ and because his equilibrium payoff is zero, he has a strict incentive to deviate to $s \in (0, -p\omega^L + c_d)$ when

$$F(v_c^*(\mu_s)) > \frac{s}{\omega^L + s}.$$

Thus, whenever the high type has a weak incentive to deviate to $s \in (0, -p\omega^L + c_d)$ (which occurs only when $F(v_c^*(\mu_s)) = F_1$), the low type has a strict incentive to do so. The D1 criterion thus assigns $\mu_s = 0$ for such s . With $\mu_s = 0$, both types would receive payoffs strictly smaller than the equilibrium payoff. Therefore, neither type deviates, and thus this

class of equilibria satisfies the D1 criterion. $\square$

Proof of Proposition 4. We compare burning money with commitment and audience cost. By Proposition 3, the defender's ex ante payoff in the audience-cost setting is at least $\mu_0 [F_1 \omega^H + (1 - F_1) (p \omega^H - c_d)]$. On the other hand, by Proposition 2, which relies on Assumption 1-(i), (ii-a), and (ii-b), the burning-money setting with commitment yields either

$$\mu_0 [F_1 \omega^H + (1 - F_1) (p \omega^H - c_d) - s_1^b] \quad \text{or} \quad \mu_0 [F_1 \omega^H + (1 - F_1) (p \omega^H - c_d)] - (1 - \mu_0) s_1^b.$$

It is straightforward to see that both of them are strictly smaller than the ex ante payoff in the audience-cost case for any $\mu_0 \in (0, 1)$.

Proposition 2 shows that commitment power improves the defender's ex ante payoff. Thus, the payoff from burning money without commitment must be smaller than that from audience cost. $\square$

Proof of Proposition 5. Consider the following cases.

Step 1 (Separating). By the assumption that $\bar{s}^t < s^*$, there does not exist a separating equilibrium. This is because the low type would profitably deviate from 0 to a positive s that the high type sends, and he would receive $F_1 \omega^L - (1 - F_1) s > 0$.

We first show that any equilibrium in this case is either a pooling equilibrium or a semi-separating equilibrium in which the high type sends $s = s' > 0$ with probability one and the low type mixes $s = s'$ and $s = 0$. First, if the high type mixes, there cannot be a signal that comes only from the high type. This is because, if a signal s comes only from the high type, the posterior becomes $\mu_s = 1$. Then, because the high type never pays audience costs in equilibrium, a signal affects the high type's payoff only through posterior beliefs. Thus, the high type would strictly prefer s such that $\mu_s = 1$ and send it with probability one. Second, when the low type mixes, the signal sent only by the low type must be $s = 0$ because such a signal leads to a posterior $\mu_s = 0$ and any $s > 0$ would result in a strictly negative payoff.

Third, suppose that two different signals are sent by both types with positive probability, denoted by $s' < s''$. Because the high type never pays audience costs on any equilibrium path, any difference in an equilibrium payoff must come from different posteriors. However, the high type's indifference condition imposes $\mu_{s=s'} = \mu_{s=s''}$. Then, because the posterior is the same, the low type would strictly prefer s' to s'' by $s' < s''$. Consequently, any equilibrium must be a pooling equilibrium or a semi-separating equilibrium in which the high type sends some $s = s' > 0$ and the low type mixes $s = s'$ and $s = 0$.

Step 2 (Semi-separating). Consider a semi-separating equilibrium where the high-valuation defender sends $s = s' > 0$ and the low type sends $s = s'$ with probability $q \in (0, 1)$ and $s = 0$ with probability $1 - q$. The low type's indifference condition yields

$$\begin{aligned} F(v_c^*(\mu_{s=s'})) \omega^L - (1 - F(v_c^*(\mu_{s=s'}))) s' &= 0 \\ F(v_c^*(\mu_{s=s'})) &= \frac{s'}{\omega^L + s'}. \end{aligned}$$

We first show that the D1 criterion requires $s' = \bar{s}^t$. Suppose $s' < \bar{s}^t$, and consider $s = \bar{s}^t$ as an off-path signal. Because the low type's equilibrium payoff is zero, he weakly benefits from deviating to $s = \bar{s}^t$ if $F(v_c^*(\mu_{s=\bar{s}^t})) \geq \frac{\bar{s}^t}{\omega^L + \bar{s}^t}$. On the other hand, because the high type never pays audience costs in any equilibrium, he strictly benefits from deviating to $s = \bar{s}^t$ if $F(v_c^*(\mu_{s=\bar{s}^t})) > F(v_c^*(\mu_{s=s'})) = \frac{s'}{\omega^L + s'}$. By $\frac{s'}{\omega^L + s'} < \frac{\bar{s}^t}{\omega^L + \bar{s}^t}$, the high type strictly benefits from this deviation whenever the low type weakly benefits from it. Then, the D1 criterion assigns $\mu_{s=\bar{s}^t} = 1$. By $\bar{s}^t < \frac{F_1}{1-F_1} \omega^L$, the high type profitably deviates to $s = \bar{s}^t$ because we have $\frac{s'}{\omega^L + s'} < \frac{\bar{s}^t}{\omega^L + \bar{s}^t} < F_1$. Thus, any semi-separating equilibrium satisfying the D1 criterion must have $s' = \bar{s}^t$.

Next, by Bayes' rule, we have $\mu_{s=\bar{s}^t} = \frac{\mu_0}{\mu_0 + (1-\mu_0)q} \in (\mu_0, 1)$ for $q \in (0, 1)$. Thus, for this semi-separating equilibrium to exist, it must be that $F(v_c^*(\mu_{s=\bar{s}^t})) = \frac{\bar{s}^t}{\omega^L + \bar{s}^t} \in (F_{\mu_0}, F_1)$. These conditions are equivalent to $\frac{F_{\mu_0}}{1-F_{\mu_0}} \omega^L < \bar{s}^t < \frac{F_1}{1-F_1} \omega^L$. By $f(x) > 0$ and $\mu_0 < 1$, we have $F_{\mu_0} < F_1$, and thus $1 - F_{\mu_0}$ is strictly positive. Thus, the threshold $\frac{F_{\mu_0}}{1-F_{\mu_0}} \omega^L$ is well defined;

by the assumption that $\bar{s}^t < s^*$, the second inequality holds.

Finally, we verify that the semi-separating equilibrium with $s' = \bar{s}^t$ satisfies the D1 criterion. Consider any off-path signal $s < \bar{s}^t$. The high type weakly benefits from a smaller s if $F(v_c^*(\mu_{s < \bar{s}^t})) \geq \frac{\bar{s}^t}{\omega^L + \bar{s}^t}$. The low type strictly benefits from deviating to a smaller s if $F(v_c^*(\mu_{s < \bar{s}^t})) > \frac{s}{\omega^L + s}$. By $\frac{s}{\omega^L + s} < \frac{\bar{s}^t}{\omega^L + \bar{s}^t}$ for $s < \bar{s}^t$, the low type has a strictly profitable deviation to a small signal whenever the high type has a weakly profitable deviation. Thus, the D1 criterion imposes $\mu_s = 0$ for $s < \bar{s}^t$. The high type would receive $p\omega^H - c_d$ from this deviation, which is strictly smaller than his equilibrium payoff, $\frac{\bar{s}^t}{\omega^L + \bar{s}^t} \omega^H + \left(1 - \frac{\bar{s}^t}{\omega^L + \bar{s}^t}\right) (p\omega^H - c_d)$. Under $\mu_{s < \bar{s}^t} = 0$, the low type would receive $-s \leq 0$. Thus, neither type profitably deviates, and the semi-separating equilibrium satisfies the D1 criterion.

Step 3 (Pooling). Consider a pooling equilibrium in which both types send $s = s' \in S^t$. On the equilibrium path, the posterior is $\mu_{s=s'} = \mu_0$. We first show that the D1 criterion requires $s' = \bar{s}^t$. Suppose to the contrary that $s' < \bar{s}^t$, and consider the off-path signal $s = \bar{s}^t$. The low type weakly benefits from deviating to $\bar{s}^t$ if

$$\begin{aligned} F(v_c^*(\mu_{s=\bar{s}^t})) \omega^L - (1 - F(v_c^*(\mu_{s=\bar{s}^t}))) \bar{s}^t &\geq F_{\mu_0} \omega^L - (1 - F_{\mu_0}) s' \\ F(v_c^*(\mu_{s=\bar{s}^t})) &\geq \frac{F_{\mu_0} \omega^L - (1 - F_{\mu_0}) s' + \bar{s}^t}{\omega^L + \bar{s}^t} \\ &= F_{\mu_0} + \frac{(1 - F_{\mu_0})(\bar{s}^t - s')}{\omega^L + \bar{s}^t} > F_{\mu_0}. \end{aligned}$$

The high type strictly benefits from deviating to $\bar{s}^t$ if $F(v_c^*(\mu_{s=\bar{s}^t})) > F_{\mu_0}$. Thus, the high type has a strictly profitable deviation to $\bar{s}^t$ whenever the low type weakly benefits from the deviation. The D1 criterion imposes $\mu_{s=\bar{s}^t} = 1$. Then, the high type strictly benefits by deviating to $s = \bar{s}^t$. As a result, any pooling equilibrium satisfying the D1 criterion must have $s' = \bar{s}^t$.

Because the low type can guarantee a payoff of zero by sending $s = 0$, a pooling equilibrium in which both types send $s = \bar{s}^t$ exists when $F_{\mu_0} \omega^L - (1 - F_{\mu_0}) \bar{s}^t \geq 0$, which is equivalent to $\bar{s}^t \leq \frac{F_{\mu_0}}{1 - F_{\mu_0}} \omega^L$.

Lastly, we show that the pooling equilibrium with $s' = \bar{s}^t$ satisfies the D1 criterion. Con-

sider an off-path signal $s < \bar{s}^t$. The high type weakly benefits from deviating if $F(v_c^*(\mu_s)) \geq F_{\mu_0}$. The low type strictly benefits from deviating to $s < \bar{s}^t$ if

$$F(v_c^*(\mu_s)) > F_{\mu_0} - \frac{(1 - F_{\mu_0})(\bar{s}^t - s)}{\omega^L + s}.$$

Because $F_{\mu_0} - \frac{(1 - F_{\mu_0})(\bar{s}^t - s)}{\omega^L + s} < F_{\mu_0}$, the low type has a strictly profitable deviation to $s < \bar{s}^t$ whenever the high type has a weakly profitable deviation to such s . Thus, the D1 criterion assigns $\mu_s = 0$. Under this belief, the high type would receive $p\omega^H - c_d$, which is strictly smaller than his equilibrium payoff, $F_{\mu_0}\omega^H + (1 - F_{\mu_0})(p\omega^H - c_d)$. The low type would receive $-s \leq 0$, which never exceeds his equilibrium payoff, $F_{\mu_0}\omega^L - (1 - F_{\mu_0})\bar{s}^t \geq 0$. Thus, this pooling equilibrium satisfies the D1 criterion.

Step 4 (defender-optimal equilibrium). By Step 1, we know that the only candidates for defender-optimal equilibria are a semi-separating equilibrium where the low type strictly mixes between $s = \bar{s}^t$ and $s = 0$, and a pooling equilibrium where both types of the defender send $s = \bar{s}^t$. Step 2 shows that the semi-separating equilibrium exists if and only if $\bar{s}^t > \frac{F_{\mu_0}}{1 - F_{\mu_0}} \omega^L$. We also know from Step 3 that the pooling equilibrium exists if and only if $\bar{s}^t \leq \frac{F_{\mu_0}}{1 - F_{\mu_0}} \omega^L$. Hence, these equilibria are defender-optimal whenever they exist. $\square$

Proof of Proposition 6. First, by Proposition 5, the ex ante payoffs from the semi-separating and pooling equilibria are, respectively,

$$\mu_0 \left[\frac{\bar{s}^t}{\omega^L + \bar{s}^t} \omega^H + \left(1 - \frac{\bar{s}^t}{\omega^L + \bar{s}^t} \right) (p\omega^H - c_d) \right] + (1 - \mu_0) \times 0 \text{ and} \quad (23)$$

$$\mu_0 [F_{\mu_0}\omega^H + (1 - F_{\mu_0})(p\omega^H - c_d)] + (1 - \mu_0) [F_{\mu_0}\omega^L - (1 - F_{\mu_0})\bar{s}^t]. \quad (24)$$

These values become equal at μ_0 such that $F_{\mu_0} = \frac{\bar{s}^t}{\omega^L + \bar{s}^t}$. The former value is greater if and only if $F_{\mu_0} < \frac{\bar{s}^t}{\omega^L + \bar{s}^t}$, which is equivalent to the condition under which the semi-separating equilibrium in the audience-cost setting with $\bar{s}^t < s^*$ exists, i.e., $\bar{s}^t > \frac{F_{\mu_0}}{1 - F_{\mu_0}} \omega^L$.

We now compare the above payoffs with the payoff from the separating equilibrium in

the burning-money setting without commitment. By Proposition 1, observe that the ex ante payoff in the burning-money setting without commitment is

$$\mu_0 [F_1 \omega^H + (1 - F_1) (p \omega^H - c_d) - F_1 \omega^L] + (1 - \mu_0) \times 0. \quad (25)$$

Rearranging the payoffs in expressions (25) and (23) shows that the payoff from the semi-separating equilibrium in the small-audience-cost setting is strictly greater than the payoff from the separating equilibrium in the burning-money setting without commitment if and only if $\bar{s}^t > \frac{F_1 D}{\delta - F_1 D} \omega^L$. Because the payoff from the pooling equilibrium with $s = \bar{s}^t$ in the small-audience-cost setting is equal to or strictly greater than the payoff from the semi-separating equilibrium whenever the former exists, the audience-cost setting yields a strictly higher defender ex ante payoff than the burning-money setting without commitment if $\bar{s}^t > \frac{F_1 D}{\delta - F_1 D} \omega^L$.

Next, suppose $\bar{s}^t < \frac{F_1 D}{\delta - F_1 D} \omega^L$. Because the separating equilibrium yields a strictly greater ex ante payoff than the semi-separating equilibrium for any $\mu_0 > 0$, we compare the payoffs from the separating equilibrium in the burning-money setting without commitment and the pooling equilibrium in the small-audience-cost setting. Observe that, by expressions (6) and (9), the interim payoffs of the two settings are equal when $s = 0$, i.e., $\hat{v}^b(\mu, 0) = \hat{v}^t(\mu, 0)$. Thus, by expression (24), we can rewrite the ex ante payoff in the pooling equilibrium as $\hat{v}^t(\mu_0, \bar{s}^t) = \hat{v}^b(\mu_0, 0) - (1 - \mu_0)(1 - F_{\mu_0})\bar{s}^t$. By expression (25), the ex ante payoff in the burning-money setting without commitment can be rewritten as $\mu_0 (\hat{v}^b(1, 0) - F_1 \omega^L)$. Then, by rearranging these payoffs, we have

$$\frac{\hat{v}^b(\mu_0, 0)}{\mu_0} = \hat{v}^b(1, 0) - F_1 \omega^L + \frac{1 - \mu_0}{\mu_0} (1 - F_{\mu_0}) \bar{s}^t$$

when the payoffs are equal. By Assumption 1-(i), the left-hand side is strictly increasing in μ_0 . On the other hand, the right-hand side is strictly decreasing in μ_0 because F_{μ_0} is increasing and $\frac{1 - \mu_0}{\mu_0}$ is decreasing in μ_0 . Because $\hat{v}^t(0, \bar{s}^t) = -\bar{s}^t < 0$ and $\hat{v}^t(1, \bar{s}^t) = \hat{v}^b(1, 0) >$

$1 \times (\widehat{v}^b(1, 0) - F_1 \omega^L)$, there is a unique $\bar{\mu}_0 \in (0, 1)$ at which the two payoffs are equal and above which the audience-cost setting yields a strictly greater ex ante payoff.

Finally, suppose $\bar{s}^t = \frac{F_1 D}{\delta - F_1 D} \omega^L$. In this case, observe that the ex ante payoff in the burning-money setting without commitment is equal to that of the semi-separating equilibrium in the audience-cost setting. Hence, the two settings yield the same ex ante payoff for $\mu_0 \in [0, \mu'_0]$ where $F_{\mu'_0} = \frac{\bar{s}^t}{\omega^L + \bar{s}^t}$, and the audience-cost setting achieves a strictly higher payoff for $\mu_0 > \mu'_0$. $\square$

Proof of Proposition 7. By Proposition 5, the defender-optimal equilibrium in the audience-cost setting with $\bar{s}^t < s^*$ is either a semi-separating equilibrium or a pooling equilibrium in which both types of the defender send $s = \bar{s}^t$.

First, suppose that the semi-separating equilibrium is defender-optimal in the small-audience-cost setting. By Proposition 5, we know that the payoff from the semi-separating equilibrium is linear in μ_0 . In the burning-money setting with commitment, we also know that, for any pair of $\mu'_0 \leq 1/2$ and $\mu''_0 > 1/2$, the slope of $V^{jo}(\mu''_0 | (\widehat{v}_s^b)_{s \in S^b})$ is greater than that of $V^{jo}(\mu'_0 | (\widehat{v}_s^b)_{s \in S^b})$. Hence, burning money with commitment is strictly more beneficial for any $\mu_0 \in (0, 1)$ whenever it is beneficial for $\mu \in (0, 1/2]$. This is true when

$$\begin{aligned} & \overbrace{V^{jo}(\mu_0 \leq \frac{1}{2} | (\widehat{v}_s^b)_{s \in S^b})} \\ & \mu_0 [F_1 \omega^H + (1 - F_1)(p\omega^H - c_d) - s_1^b] + (1 - \mu_0) \times 0 \\ & > \underbrace{\mu_0 \left[\frac{\bar{s}^t}{\omega^L + \bar{s}^t} \omega^H + \left(1 - \frac{\bar{s}^t}{\omega^L + \bar{s}^t} \right) (p\omega^H - c_d) \right]}_{\text{Semi-separating in audience costs}} + (1 - \mu_0) \times 0. \end{aligned}$$

This condition simplifies to

$$s_1^b < \left(F_1 - \frac{\bar{s}^t}{\omega^L + \bar{s}^t} \right) ((1 - p)\omega^H + c_d),$$

which is Assumption 1-(ii-c).

Second, suppose the pooling equilibrium is defender-optimal in the small-audience-cost

setting. By the proof of Proposition 6, we know that $\hat{v}^t(\mu_0, \bar{s}^t) = \hat{v}^b(\mu_0, 0) - (1 - \mu_0)(1 - F_{\mu_0})\bar{s}^t < \hat{v}^b(\mu_0, 0)$. Moreover, Proposition 2 shows that the ex ante payoff in the burning-money setting with commitment is strictly greater than $\hat{v}^b(\mu_0, 0)$ for any $\mu_0 \in (0, 1)$ under Assumption 1-(i) and (ii-b). Therefore, the defender's ex ante payoff of the burning-money setting with commitment is strictly greater than that of the audience-cost setting with $\bar{s}^t < s^*$ for any $\mu_0 \in (0, 1)$. $\square$

Proof of Proposition 8. First, consider the burning-money setting. The results in Propositions 1 and 2 use Assumption 1-(i), (ii-a), (ii-b), (iii-a), and (iii-b). By Propositions 1 and 2, the defender conveys his type in the defender-optimal equilibrium in the burning-money setting with and without commitment. Thus, war occurs when a challenger with $v_c > v_c^*(1)$ challenges and the high type fights. The ex ante probability of war in this case is $\mu_0 \times (1 - F_1)$.

Second, consider the audience-cost setting with $\bar{s}^t \geq s^*$. When $-p\omega^L + c_d \geq \frac{F_1}{1-F_1} \omega^L$, by Proposition 3, a separating equilibrium is defender-optimal. Thus, as in the burning-money setting, the ex ante probability of war is $\mu_0(1 - F_1)$, which is equal to the probability of war in the burning-money setting.

When $-p\omega^L + c_d \leq \frac{F_1}{1-F_1} \omega^L$, the defender-optimal equilibrium becomes the one in which both types commit to fighting by sending $s \geq -p\omega^L + c_d$. Thus, war erupts when a challenger with $v_c > v_c^*(1)$ challenges. Then, the ex ante probability of war is $1 - F_1$. Therefore, when $-p\omega^L + c_d < \frac{F_1}{1-F_1} \omega^L$, the probability of war in the large audience-cost setting is strictly greater than that of the burning-money setting for any $\mu_0 < 1$ and any $F_1 < 1$. The special case of $F_1 = 1$ belongs to the condition $-p\omega^L + c_d \leq \frac{F_1}{1-F_1} \omega^L$ because $\frac{F_1}{1-F_1} \omega^L = \infty$. In this case, the burning-money and audience-cost settings yield the same ex ante probability of war by $\mu_0(1 - F_1) = 1 - F_1 = 0$ when $F_1 = 1$.

When $-p\omega^L + c_d = \frac{F_1}{1-F_1} \omega^L$, both of the above equilibria are optimal. Moreover, there also exist semi-separating equilibria in which the low type mixes $s \geq -p\omega^L + c_d$ and $s = 0$. Case 3 of the proof of Proposition 3 shows that the low type may commit to fighting with

any probability $q \in [0, 1]$, whereas the high type commits to fighting for sure. Consequently, the ex ante probability of war is $(\mu_0 + (1 - \mu_0)q)(1 - F_1) \in [\mu_0(1 - F_1), 1 - F_1]$. Therefore, when $-p\omega^L + c_d = \frac{F_1}{1-F_1} \omega^L$, the audience-cost setting yields the same probability of war as the burning-money setting if $q = 0$ and a strictly greater probability of war otherwise.

Third, consider the audience-cost setting with $\bar{s}^t < s^*$. By Proposition 5, when $\bar{s}^t \in \left(\frac{F_{\mu_0}}{1-F_{\mu_0}} \omega^L, s^*\right)$, the defender-optimal equilibrium is a semi-separating equilibrium where the low type mixes between $s = 0$ and $s = \bar{s}^t$. By $F(v_c^*(\mu_{s=\bar{s}^t})) = \frac{\bar{s}^t}{\omega^L + \bar{s}^t}$ as shown in Proposition 5, the ex ante probability of war is $\mu_0(1 - F(v_c^*(\mu_{s=\bar{s}^t}))) = \frac{\omega^L}{\omega^L + \bar{s}^t} \mu_0$. This is strictly greater than the probability of war in the burning-money setting because we have $1 - F(v_c^*(\mu_{s=\bar{s}^t})) > 1 - F_1$ when $\bar{s}^t < \frac{F_1}{1-F_1} \omega^L$.

By Proposition 5, the defender-optimal equilibrium is a pooling equilibrium with $s = \bar{s}^t$ when $\bar{s}^t \leq \frac{F_{\mu_0}}{1-F_{\mu_0}} \omega^L$. Thus, war occurs when the challenger with $v_c > v_c^*(\mu_0)$ challenges and then the high type fights. Hence, the probability of war is $\mu_0(1 - F_{\mu_0})$, which is greater than the probability of war in the burning-money setting, $\mu_0(1 - F_1)$, by $F_{\mu_0} < F_1$. In sum, when $\bar{s}^t < s^*$, the ex ante probability of war is strictly greater in the audience-cost setting than in the burning-money setting for any $\mu_0 \in (0, 1)$. $\square$

C Audience Costs with Commitment

For completeness, here we characterize the audience-cost setting with commitment power.

Proposition 11. *Suppose Assumption 1-(i) holds.*

- When $-p\omega^L + c_d \geq \frac{F_1}{1-F_1} \omega^L$ or $\bar{s}^t < -p\omega^L + c_d$, the defender sends some $s > 0$ when $\omega = \omega^H$ and sends $s = 0$ when $\omega = \omega^L$. The challenger challenges when $s = 0$ or $v_c > \frac{c_c}{1-p}$, upon which the defender fights if and only if $\omega = \omega^H$.
- When $-p\omega^L + c_d \leq \frac{F_1}{1-F_1} \omega^L$ and $\bar{s}^t \geq -p\omega^L + c_d$, both types of the defender send some $s \geq c_d - p\omega^L$. The challenger challenges when $s < c_d - p\omega^L$ or $v_c > \frac{c_c}{1-p}$. Both types of

the defender fight when challenged on the equilibrium path.

Proof of Proposition 11. We first specify the signal structure and then show that it achieves the defender's highest possible ex ante payoff.

Step 1 (Signal structure). First, suppose $-p\omega^L + c_d \geq \frac{F_1}{1-F_1} \omega^L$ or $\bar{s}^t < -p\omega^L + c_d$. Consider a separating signal structure in which the low type sends $s = 0$ and the high type sends a fixed feasible signal $s > 0$. The corresponding posteriors are $\mu_{s=0} = 0$ and $\mu_{s>0} = 1$. Bayes plausibility requires

$$\begin{aligned} \tau(\mu_{s=0})\mu_{s=0} + (1 - \tau(\mu_{s=0}))\mu_{s>0} &= \mu_0 \\ \implies \tau(\mu_{s=0}) &= 1 - \mu_0 \text{ and } \tau(\mu_{s>0}) = \mu_0. \end{aligned}$$

The defender's signal structure follows from the above values:

$$\pi(0|\omega^H) = 0, \pi(s > 0|\omega^H) = 1, \pi(0|\omega^L) = 1, \text{ and } \pi(s > 0|\omega^L) = 0.$$

As a result, the challenger challenges when $s = 0$ or $v_c > \frac{c_c}{1-p}$, upon which the high-valuation defender fights.

Next, suppose $-p\omega^L + c_d \leq \frac{F_1}{1-F_1} \omega^L$ and $\bar{s}^t \geq -p\omega^L + c_d$. As in the case without commitment (Proposition 3), the defender is better off committing to fighting ($s \geq -p\omega^L + c_d$) rather than separating. That is, when

$$\begin{aligned} & \overbrace{F_1 (\mu_0 \omega^H + (1 - \mu_0) \omega^L) + (1 - F_1) (p (\mu_0 \omega^H + (1 - \mu_0) \omega^L) - c_d)}^{\text{Committing to fighting}} \\ & \geq \underbrace{\mu_0 (F_1 \omega^H + (1 - F_1) \omega^L)}_{\text{Committing to separating}}, \end{aligned}$$

committing to fighting is weakly better than the separating signal structure. Rearranging, we obtain the condition $-p\omega^L + c_d \leq \frac{F_1}{1-F_1} \omega^L$.

Step 2 (Payoff optimality). The above signal structures yield the highest possible ex

ante payoff in each case. First, suppose $-p\omega^L + c_d \geq \frac{F_1}{1-F_1} \omega^L$ or $\bar{s}^t < -p\omega^L + c_d$. The ex ante payoff in the case with commitment is $\mu_0 \hat{v}^t(1, 0) = \mu_0 \hat{v}^b(1, 0)$. By $\hat{v}^t(\mu, s) \leq \hat{v}^b(\mu, 0)$ for $s < -p\omega^L + c_d$, Assumption 1-(i) guarantees $\hat{v}^t(\mu, s) \leq \hat{v}^b(\mu, 0) \leq \mu \hat{v}^b(1, 0)$. Then, by Bayes plausibility, we obtain the ex ante payoffs

$$\sum_{s \in S^t} \tau(\mu_s) \hat{v}^t(\mu_s, s) \leq \sum_{s \in S^t} \tau(\mu_s) \mu_s \hat{v}^t(1, 0) = \mu_0 \hat{v}^t(1, 0).$$

Next, suppose $-p\omega^L + c_d \leq \frac{F_1}{1-F_1} \omega^L$ and $\bar{s}^t \geq -p\omega^L + c_d$. Because $F_1 \omega^L + (1 - F_1) W^L \geq 0$ in this case, it is straightforward to see

$$\hat{v}^t(\mu, s) \leq \mu \hat{v}^t(1, 0) \leq \mu \hat{v}^t(1, 0) + (1 - \mu) (F_1 \omega^L + (1 - F_1) W^L).$$

Then, by Bayes plausibility, we have the ex ante payoffs

$$\begin{aligned} \sum_{s \in S^t} \tau(\mu_s) \hat{v}^t(\mu_s, s) &\leq \sum_{s \in S^t} \tau(\mu_s) \mu_s \hat{v}^t(1, 0) + \sum_{s \in S^t} \tau(\mu_s) (1 - \mu_s) (F_1 \omega^L + (1 - F_1) W^L) \\ &= \mu_0 \hat{v}^t(1, 0) + (1 - \mu_0) (F_1 \omega^L + (1 - F_1) W^L), \end{aligned}$$

which establishes the payoff optimality. When $-p\omega^L + c_d = \frac{F_1}{1-F_1} \omega^L$, both signaling structures are optimal. $\square$