Unattributable Coercion

September 3, 2023

Abstract

Can states use unattributable attacks to coerce others? Since Schelling (1960), the conventional wisdom is no. Explicit threats are necessary for coercive success. This wisdom underpins research into reputation, covert action, cyber-attacks, election meddling, and funding terror proxies. Using formal analysis and intuitive argumentation, I argue that unattributable coercion is achievable. Targets can learn from a history of harm that harm will come to them in the future, without learning who chose to harm them. Unattributable coercion is easiest when the harm inflicted is large, attacks are distinctive, the risk of an accident is small, and—most critically—there are many potential attackers. The results clarify why unattributable coercion was unobserved during the bi-polar Cold War, and also predicts that it will become more common as unattributable weapons proliferate to non-state actors and middle powers. I substantiate these claims by showing that still-unattributed Havana Syndrome attacks have influenced US policy.

Unattributable attacks—loosely, attacks where the victim cannot confidently identify the perpetrator—are common in the modern world.[1] Because of their strategic complexity and substantive importance attribution problems are fundamental to modern research into grey zone conflict (Schram 2022), cyber conflict (Gartzke and Lindsay 2015), covert action (Spaniel and Poznansky 2018), election meddling (Levin 2021), political misinformation (Lazer, Baum, Benkler, Berinsky, Greenhill, Menczer, Metzger, Nyhan, Pennycook, Rothschild, Schudson, Sloman, Sunstein, Thorson, Watts, and Zittrain 2018), mass atrocities (Krcmaric 2019) and funding non-state proxies (Canfil 2022).

Will the proliferation of unattributable weapons change coercive practices in world politics? More specifically, can perpetrators use unattributable attacks to coerce others into changing their policy choices?[2] Since Schelling (1960), the conventional wisdom is no. Explicit threats are necessary for coercive success. Coercion relies on the Target's (of an attack) anticipation of punishment for doing something Competitors[3] do not want her to do. If Competitors do not explain the conditions under which they will launch an attack, the Target cannot know whether their future choices will trigger an attack against them (Borghard and Lonergan 2017). Furthermore, Competitors must generate a reputation as resolved to make threats effective (see Dafoe, Renshon, and Huth 2014; Brutger and Kertzer 2018)

I argue that Competitors can use unattributable attacks to coerce rational Targets into altering their foreign policy choices. My mechanism hinges on a difference between how Targets form rational expectations about the consequences of their actions from a pattern of historical events, and how Targets identify the sponsors of attacks after they implement a policy and suffer harm. Attribution depends on inferences the Target can draw about a specific Competitor's involvement. Competitors can avoid attribution by concealing their attacks as accidents or blaming other potential attackers. However, a Target's expectation that she will face an attack if she implements a policy depends mainly on the historical pattern of attacks against her or against other Targets that implemented similar policies. If a Target realizes that a specific kind of attack has reliably followed a specific policy choice, she can assess with high confidence that she will suffer that same kind of harm if she implements that policy. She does not need to know who the perpetrator was. All she needs is a sufficiently long history to understand that someone will punish her if she chooses to act.

I identify the real-world conditions necessary to support unattributable coercion. Many variables are

[1] Brought on by emerging military technology (Horowitz 2020; Lin-Greenberg 2023).

[2] Axelrod and Iliev (2014); Baliga, Mesquita, and Wolitzky (2020) and others ask can uniformed states deter unattributable attacks against them. I ask can violent actors utilize unattributable attacks to alter others behavior.

[3] I use Competitor to describe actors who may launch attacks in secret. Since their type varies, and some never launch attacks, it is not appropriate to call them Attackers, Perpetrators, etc.

important. Unattributable coercion is easiest if the harm attacks inflict on the Target is large, the direct cost of an attack is low, there is not too much initial uncertainty about whether the Competitor is resolved or that attacks are actually accidents. But the number of Competitors stands out as critical. When there is only one Competitor, unattributable coercion can only be supported under implausible conditions. However, adding even a second Competitor generates empirically plausible conditions for unattributable coercion. If there are many Competitors unattributable coercion becomes easy.

This result holds grim implications for the future of US national security. It illustrates how the unipolar moment could end even absent a pier-competitor (Wohlforth 2009). The proliferation of unattributable weapons in the modern world means that many small states and non-state actors can launch unattributable attacks against the US. So long as these attacks display a calling card feature, the US will eventually understand them, and alter its behavior. As demonstrated in section 3, the US is already influenced by attacks they cannot attribute. Unattributed Havana Syndrome attacks against US embassy staff have coerced the US to shut down consulates, repatriate CIA staff, delay vice presidential visits, and other changes.

I make two contributions to research into coercive threats where attribution is unclear beyond those stated above[4]. First, I advance a conceptually novel definition of plausible deniability (Poznansky 2022) that draws from research into electoral accountability in American politics (Ashworth 2012). I argue that audiences learn from strategic context and not only direct evidence, express non-certain confidence that a specific perpetrator harmed them, and then express outrage at the Competitor only if that confidence is sufficiently high. Second, I provide one reason why scholars who focus on state-specific reputations find mixed, unstable, or weak empirical results in cross-national studies (Huth and Russett 1993; Weisiger and Yarhi-Milo 2015; Wood 2012; Danilovic 2001; Schultz 2001; McManus 2014; Uzonyi, Souva, and Golder 2012; Weeks 2008), given the strength of the theoretical prediction and strong evidence in specific cases and experiments (Renshon, Dafoe, and Huth 2018; Levy, McKoy, Poast, and Wallace 2015; Kertzer 2017).[5] My reason is that states can affect coercion without cultivating a reputation at all. In a similar way, my finding similarly compliments recent into terrorism and credit-claiming (Kydd and Walter 2006) by explaining why terrorists may only claim 20% of attacks (Min 2022). The reason is that Targets may infer reasoning absent an explicit claim. Outside of international relations, my analysis of the multiple Competitor problem illuminates how type-uncertainty can resolve the volunteer's dilemma and other common coordination challenges

[4] See Carnegie (2021) for review and Kurizaki (2007); Carson (2018); Debs and Monteiro (2014); Wolford, Reiter, and Carrubba (2011); Arena and Wolford (2012) for additional arguments beyond those cited above

[5] (cf Press 2007).

in multi-actor games.

1 Unattributable Coercion: definitions and wisdoms

Schelling argued that unattributable coercion faced a paradox. Some strategic settings are so complicated that the Target could not infer who attacked her, or even if an attack occurred, with high confidence. Therefore, it was plausible that a Competitor could launch an attack and disclaim responsibility for it. However, in any setting this complicated, the Target is unlikely to understand if or why they were attacked. Since the Target could not appreciate the logic behind the attack, they would continue to act in the way that the Competitor did not want.

This basic conjecture persists. For example, after a comprehensive review, Borghard and Lonergan (2017) find that the entire "literature on coercion suggests that four fundamental conditions must be met for coercion to succeed: the coercive threat must be clearly communicated; it must be linked to a cost–benefit calculus such that the target's costs of conceding are less than the costs of not complying; it must be credible; and there must be an element of reassurance." Therefore, coercion "requires attribution to be effective." This argument is repeated in other recent, reviews of rationalist coercion (eg Greenhill and Krause 2018). It is offered as one of several arguments by those who believe that grey zone and cyber-coercion will be difficult to sustain (eg Lindsay 2015; Libicki 2009).[6] As a result of this persistently held belief, theorists of secrecy and coercion instead focus on accidental or imperfect attribution (Debs and Monteiro 2014; Baliga and Wolitzky 2018), credit claiming, or making threats public to some audiences but secret to others (Carson and Yarhi-Milo 2017; Kurizaki 2007). Others illuminate theoretically rational exceptions, such as deterrence via ransomware (Jun 2022). Otherwise scholars treat incomplete information as something to overcome to make threats credible and understandable (eg Gurantz and Hirsch 2017; Dafoe et al. 2014; Kydd and Walter 2006). Empirical scholars who study secrecy and coercion debate whether certain attacks are truly unattributable. If they agree that they are, they study them as a method for achieving brute force objectives rather than coercion objectives (eg Joseph and Poznansky 2018; Poznansky 2019).

Schelling's paradox underpins this research trend, but it has not been subject to scrutiny. In the next section I study a model of unattributable coercion. While some worry that formal models are difficult to interpret, and rely on unrealistic assumptions that are too abstract to apply in real life, a model is useful

[6] These authors also point to the difficulty of repeaeting attacks, among other factors.

here for two reasons. First, the 60-year-old conventional wisdom holds that unattributable coercion is not rationalizable. Thus, if I can rationalize it in a simple model, it provides scope for more intuitive theorizing. As we shall see in section 2.1.2, the model not only departs from Schelling's conventional wisdom,[7] it yields clues about why the result likely extends to more realistic settings. Second, even overt coercion is hard to observe within and across cases (Press 2007). When attacks are secretive, observing coercion is even harder (Carnegie 2021). As I later explain, the model reveals that unattributable coercion arises under conditions that are not the same as overt coercion. Similarly, the observable indicators are different. The theory helps empirical scholars by providing clear predictions about where to look for unattributable coercion, and how it manifests.

Before we can understand the model and its results, we need to define rationalist, unattributable coercion. In short, I mean a strategic setting that can sustain *coercion* as the result of the Competitors' *unattributable attack*. The remainder of this section details what I mean by (a) coercion and (b) unattributable attacks.

1.1 Coercion

The study of rationalist coercion[8] assumes a repeated, strategic interaction between a Target (of an attack) and a Competitor (of a policy position) (see Dafoe et al. 2014). In it, the Target is given an opportunity to revise the status quo or not. Regardless of what the Target does, the Competitor is given the opportunity to harm the Target or not. This interaction repeats. Following recent rationlist research, I define coercion as a specific equilibrium within this strategic setting. In the coercion equilibrium the Competitor harms the Target if and only if the Target plays revision (i.e. chooses to revise the status quo). The Target would play revision if it did not raise the risk of harm. However, the Target avoids revision because she knows that harm will come to her otherwise. As a result, we observe that the Target does not enacts revision, and the Competitor does not harm her. But if the Target sought revision, the Competitor would punish her.

If the Competitor's cost to punish the Target is sufficiently low, and the Target's sensitivity to the Competitor's harm is sufficiently high, a coercion equilibrium exists in a repeated interaction. If states are initially uncertain about the Target's resolve, they arrive at coercion over time. Since this result is well known (Gu-

[7] Some game theorists could likely intuit that unattributable coercion is rationalizable. The simple model is valuable because it clearly explains the logic to non-formal readers, identifies the precise conditions under which it works, and make sure that strategic challenges, such as the volunteer's dilemma, do not ruin the result.

[8] As Kydd and McManus (2017) argues, in abstract, rationalist models the difference between compellance and deterrence is semantic. Consider I can deter you from taking an action and compel you not to take the same action. The difference is meaningful once we introduce psychological variables. But since I use a rational model to address a conventional wisdom about rational deterrence, I use the words coercion and deterrence interchangeably.

rantz and Hirsch 2017), we don't describe it in detail. But the core model is the basis for our study. Thus, we re-derive this classic result in Appendix A. We treat the conditions where it arises (sensitivities to costs and repeated play) as our initial scope condition.

1.2 Unattributable attacks

Consistent with Poznansky (2022) I say that the Competitor has launched an unattributable attack if the Competitor inflicts harm on the Target, the Target knows that she suffers harm, but the Target or any other relevant audience cannot say that a specific Competitor has inflicted harm upon her with sufficiently high confidence. Consistent with research into backlash for exposed attacks, the relevant audience is any actor that would impose costs on the Competitor if they learned that the Competitor sponsored secretive attack. This could represent the Target who faces strategic incentives to retaliate once an attack against them is attributed (Baliga et al. 2020), or domestic audiences who are put-off when they learn of their leader's controversial secretive actions (Kurizaki 2007),[9] or the international community that express backlash against secret violent actions that circumvent laws and norms that underpin world order (see Krcmaric 2019; Poznansky 2019; Colgan 2021, for evidence).

The critical difference between attributed and unattributed attacks is plausible deniability (see Carnegie 2021). As others have noted, plausible deniability assumes that the Competitor leaves no direct evidence of his sponsorship (Joseph and Poznansky 2018).[10] But this is not enough. It must be true that the relevant audiences cannot use strategic inferences to figure out the Competitor was the perpetrator (Axelrod and Iliev 2014).

If we accept that attribution includes strategic inferences, we face definitional challenges. If the Target observes harm, she will suspect that the Competitor harmed her with some probability (possibly very low). How can we distinguish between attributable and unattributable attacks in theory? Following the American politics literature on electoral accountability (Ashworth 2012), I model attribution of an attack by embedding the Target's beliefs into the Competitor's utility function. I assume a belief threshold that defines how confident the Target (or any audience) must be before she says that attribution has happened. I say that attribution has happened (and the Competitor suffers retaliatory costs) if and only if the Target's beliefs that the Competitor is responsible surpass this belief threshold. For brevity I write the Target's beliefs. But since

[9] cf Myrick (2020).

[10] In an exhaustive review, Poznansky (2022, 523-524) identifies three 'threats to plausible deniability' at the state-level: leaks, rival intelligence, electronic recording. All are variants of direct evidence.

updating follows from observable events in my theory (whether the Target suffered harm), then all relevant audiences would update in the same way. Specifically, if I included a third-party who would inflict harm if they were sufficiently confident that the Competitor launched an attack, I would achieve identical results. For a technical definition of attribution with or without coercion see the end of section B.1.

Defining attribution as beliefs relative to a threshold provides intuitive explanatory gains. For example, Poznansky and Perkoski (2018) wrestle with a case where the US assessed with confidence that China hacked the Office of Personal and Management. The US had no direct evidence, and China denied their involvement. The US arrived at their high-confidence estimate based on their knowledge of China's strategic aims, and the information that was stolen. Under the direct evidence definition, China's attack was unattributed. I code this case as attributed because the US was so confident that China was involved that they were willing to impose counter-measures.[11] In a case like this my definition is attractive because it requires the level of attribution to be meaningful for the attacker. This is likely consistent with China's reasoning. China likely did not care much about whether the US would render a low or moderate confidence estimate of their sponsorship. What they likely cared about was whether the US would retaliate, and how severe the retaliation would be. The answer to these questions largely depended upon US estimates of China's sponsorship.

Of course, my definition cannot tell us the theoretically "correct" level of confidence to set the threshold at. In any real-life context, the true attribution threshold varies depending on the audience, the nature of the harm, and the cost of retaliation against a Competitor. What an analysis of the strategic model can tell is the smallest attribution necessary to support unattributable coercion give the other parameters and the coercion equilibrium we are analyzing. It is hard to explain whether an assumed threshold is empirically plausible without first detailing the equilibrium. Thus, I first present the necessary thresholds derived from the strategic model, then provide a conceptual interpretation of whether that threshold is empirically plausible.

The theory that follows formalizes three potential mechanisms through which a Competitor can conceal his sponsorship behind an attack. First, the Target may suffer harm for reasons unrelated to her policy choices. The Target may be uncertain if an attack happened at all if the Competitor disguises his attack as an accident. One way to do this is to make an attack appear like no intervention took place. For example, an assassination that looks like a heart attack, or making mis-information appear like organic electoral

[11] The US does not publicise the actions that they took. But my off-record interviews suggest that the US has responded to the OPM hack.

discussion (Levin 2021). Another way to do this is to sponsor proxies that want to harm the Target for a different reason (Canfil 2022). For brevity, I refer to both mechanisms as an accident, meaning there is a random chance that the Target may suffer harm. The Target must weigh this risk when he evaluates whether a specific Competitor was responsible. Second, if the Target has many adversaries, the Target may know that she was attacked, but may not know who is responsible. Any Competitor can launch an attack and claim that someone else was responsible. Finally, the Target is initially uncertain about each Competitor's intentions and capabilities. Potentially, Competitors that are initially perceived as unlikely attackers can more easily avoid detection (cf Baliga et al. 2020).

It is not clear how these mechanisms fit Schelling's paradox. On the one hand, each source of uncertainty makes it easier for the Competitor to conceal his sponsorship. But they also make it more difficult for the Target to learn the logic behind an attack. In what follows, we use formal analysis to better understand the specific role that each plays.

2 Rationalizing Unattributable Coercion

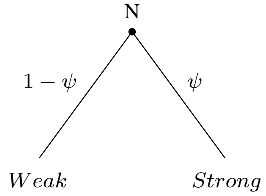
Table 1 presents a formalization for many of these intuitive concepts. The technical set-up is in B.1. The model includes the complicating factors that supply Schelling's paradox. It allows for multiple Competitors, where each Competitor varies in her level of resolve. It also assumes that harm could come to the Target as the result of an accident λ, or a strategic attack from one of up to J Competitors. However, it formally disentangles any Competitors' choice to harm the Target, and whether the Target or any Competitor observes harm come to the Target (only x_t is public). It also assumes all attacks are costly, but disentangles the direct costs that a Competitor suffers from launching an attack k_d, and the costs a Competitor only accrues if an attack is attributed to them (k_i). The model focuses on attacks that inflict harm on the Target, and do not directly deny the Target their policy goals. Adding in a denial affect only makes unattributable coercion easier to sustain.[12] I focus on punitive harm because it is the theoretically difficult test, and is most consistent with modern unattributable weaponry (Libicki 2009). Finally, the payoff and belief structures explicitly assume that attribution can occur from strategic inferences, and absent direct evidence of an attack. This is a critical feature of testing Schelling's argument because it assumes attribution under a far broader set of conditions.

[12] Specifically, we can support unattributable attacks that inflict less harm, and with a high chance of an accident.

Table 1: Summary of formalization

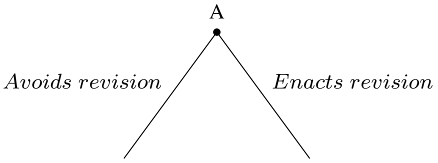
Player Notation: $A =$ Target, $B_j =$ Competitor 'j' (there are J Competitors)

Begin of game: Every Competitors' Resolve is chosen

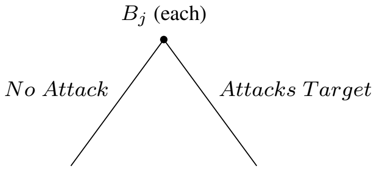


Nature determines if each Competitors is strongly/weakly resolved (private for each Competitor). Highly resolved Competitors value issues $\pi = \pi_H$. Weakly resolved Competitors value issues $\pi = \pi_L$

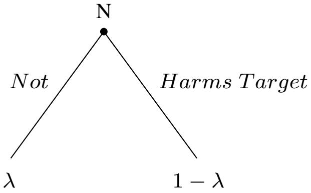
One period of the model (repeats infinitely). Arbitrary period denoted t



The Target is faced with the choice to enact a favourable foreign policy revision $a_t = r$ or not $a_t = nr$. Target's choice is public.



Each Competitor (There are $J \geq 1$ of them) simultaneously decides to harm the Target $j_t = h$ or not $j_t = nh$. Each Competitor's choice is private.



Nature inflicts harm on Target or not. Completely unobserved.

Final Revelation: If either at least one Competitor strategically, or Nature (λ) randomly, inflicts harm on Target, Nature reveals to all that Target was harmed $x_t = h$. Otherwise, Nature reveals that Target was not harmed $x_t = nh$.

Summary of one-period (t) payoffs:

$$\text{Target's payoff} \quad \underbrace{(a_t = r)}_{\text{A selects revision}} \times 1 - \overbrace{(x_t = h)}^{\text{A suffers harm}} \times c$$

$$\text{One Competitor's (indexed } j\text{) payoff :} \quad \underbrace{(a_t = nr)}_{\text{A avoids revision}} \times \pi - \overbrace{(j_t = h)}^{B^j \text{ inflicts harm}} \times k_d - \underbrace{(\beta^j_t > \hat{\beta})}_{\text{Attribution has happened}} \times k_i$$

Attribution and cost parameters:

β^j_t	A's belief that the specific Competitor j has the interest/ability to harm her.
$\hat{\beta}$	The Attribution threshold. When $\beta^j_t > \hat{\beta}$, that Competitor suffers audience costs k_i for attacks against A, regardless of whether they attack.
c	How sensitive A is to harm.
k_d	Direct resource and attention cost to B for launching attack
k_i	Indirect cost from audience backlash if B is blamed for attack

I want to know: is unattributable coercion theoretically possible? If it is, what is the rationalist mechanism that can support it? To answer these questions, I solve several variants of the model in search of Perfect Bayesian Equilibria that match my definition of unattributable coercion. As a reminder, this is an equilibrium wherein the anticipation of harm eventually coerces the Target from future policy revision, but the Target cannot determine with confidence above the attribution threshold who is responsible for the attacks (or even certainty that the attacks are not a result of an accident).

Result 1: Unattributable coercion is possible. It easier to support if the harm inflicted on the Target is high, the direct cost an attack is low, the risk of an accident is not too high, the attribution threshold is sufficiently high and there are more Competitors.

I now informally describe the simplest strategies[13] and rationalist mechanism that supports this results. For clarity, I focus on the one Competitor case. As stated in result 1, increasing the number of Competitors influences the conditions under which we can avoid attribution and achieve coercion. However, the core mechanism for unattributable coercion emerges even in the one Competitor case. Thus, focusing here allows me to detail the mechanism with the fewest moving parts. For technical readers, a preliminary analysis of reversion strategies necessary to support the result, and a formal statement of strategies I use to derive results 1 is described in section B.3. The equilibrium analysis used to support my description of the mechanism appears in proposition B2.

The Competitor's strategies share many features with the classic logic of attributed coercion in studies of reputation and resolve (Gurantz and Hirsch 2017). Competitors come in two types, those who are highly resolved to stop the Target and those who are not. In equilibrium, the weakly resolved Competitors never inflict harm on the Target. Any strongly resolved Competitor inflicts harm on the Target with positive probability if and only if (a) the Target pursues revision; and (b) that Competitor can avoid attribution. That is, a specific Competitor must avoid the perception that he is responsible for the attacks.

The Target's most basic choice is whether to pursue revision or not. She weighs her expectation of suffering harm in a given period against the value she gets from revision. When the Target believes that pursuing revision holds little increased risk of harm in that period, she does it. The Target's expectation of harm always captures three relatively fixed variables: (a) her initial beliefs that Competitors want to harm her; (b) her beliefs that harm comes as the result of an accident; (c) her knowledge of the Competitor's strategy. Over time, her perceptions of harm will change because she gathers information about the Competitor's

[13] That is, pure strategy equilibria. The Appendix reports others. But the basic logic I describe is common to them.

interests and abilities as she pursues policies and suffers harm (or doesn't).

Under interesting[14] conditions, the Target's initially estimates that it is unlikely that there is a Competitor out there that wants to punish her. Therefore, the Target is willing to pursue revision[15] in the first few periods. In any period that she pursues revision and suffers harm, she does not know if that harm followed from an accident or punishment. However, after several periods she can look back and recall how frequently she pursued revision and whether she suffered harm. If she looks back through history and realizes that she suffered harm far more frequently than she expects an accident to occur, she grows confident that someone is punishing her because she pursues revision. She stops future revision, and the interaction enters the coercion phase.

To be clear, the Target performs this historical analysis without knowing whether the harm she suffered was the result of an accident, or whether it was an attack. Similarly, even if the Target knows that she was attacked, she can perform this historical analysis even if she does not know who was behind it, or the reason for it. Furthermore, no Competitor need develop a specific reputation for resolve. The Target simply develops the expectation that harm will come to her for future actions based on a long enough history of harm coming shortly after a foreign policy choice.

Of course, the Target is also trying to figure out who is harming her. Even though the attacks are not attributed, the Target (and other audiences) is drawing strategic inferences about who the likely sponsor is. After all, the Target is fully aware of the strategic incentives of the Competitor, and holds expectations that the Competitor is sufficiently resolved to inflict harm upon her.

Thus, we can re-frame Schelling's paradox as a race in two kinds of beliefs that are increasing over time and approaching a critical threshold. The Target's belief that (a) harm will come to her in the future if she selects another revisionist policy; and (b) a specific Competitor is causing her harm. To achieve unattributable coercion, the Target must grow confident in (a) faster than (b).

The question, then, is what determines how quickly these different beliefs arrive at their respective thresholds? In what follows I analyze these two thresholds individually. I start with the Target's perception that someone is harming her because of her policy choices.

[14] When the Target's prior belief that Competitors are highly resolved is too high we never see the Target select revision because she is too confident she will be attacked.

[15] As noted above, revision is short-hand for the policy choice that Competitors do not like. But the model is abstract. It applies in any setting where the Target is confronted with two options, one that Competitors like more than the other.

Result 2: The time it takes the Target to realize someone is harming her strategically is increasing in both the risk of an accident, the initial expectation that the Competitors are weakly resolved; and decreasing in the amount of harm that comes to the Target as the result of an attack.

The logic of these results is straightforward, so I leave much of the discussion to the Appendix B.4.3.[16] But the results are valuable because they can help us understand the combination of conditions that are necessary for learning in a period of time that is empirically plausible.

I now turn to the second part of the race: the Target's ability to attribute harm to a specific Competitor. My theory explicitly allows each Competitor to utilize several mechanisms—accident, uncertainty about the Competitors resolve, and variation in the number of Competitors who could launch an attack—to disclaim sponsorship. To some degree, all of these play a role in concealing the Competitor's sponsorship. But one stands out as critical for resolving Schelling's paradox: the number of Competitors. This is important because many past studies of unattributable coercion assume but one Competitor.

It turns out that the one-Competitor model is a special case. The parameters that support it, we believe, rarely arise in real life. To understand why, it is useful to focus on the one-competitor model and examine the two belief thresholds in our race, relative to each other. Recall, the attribution threshold is exogenously set. It represents the level of confidence the Target needs that a specific Competitor is responsible to retaliate against that Competitor. When it is met, we say attribution occurs, and the Competitor incurs an attribution cost. The second threshold represents the minimum level of confidence the Target must hold that any Competitor is attacking her for the Target to prefer the status quo over taking the revision opportunity. As stated above, this threshold arises endogenously.

Result 3a: When there is only one Competitor the attribution threshold necessary to sustain unattributable coercion must be at least as high as the belief threshold necessary to coerce the Target from future revision.

See section B.4.3 for technical support. Here we informally describe the results. When there is only one Competitor, the Target's belief that someone will harm her is equivalent to her belief that the lone Competitor harmed her. It follows that the threshold necessary to avoid attribution must be as high as her belief that someone will attack her. We visualize how this would unfold over time in Figure 1(a). The Figure assumes an attribution threshold that can support result 1. It then plots how the Target's equilibrium beliefs change over time under in the sub-game where each time the Target seeks revision she suffers harm. Based on the cost of harm, the risk of an accident, etc. the Target is not deterred until he is 85% confident that *some*

[16] To be clear, result 1 and 2 are derived from the same equilibrium analyzed in proposition B.2. Result 2 describes the time it takes to arrive at the on-path coercion sub-game.

Competitor is out there inflicting harm. The game begins and the Target is far less confident that harm will come to her because she is uncertain about the Competitors' resolve. The red triangles plot changes in the Target's beliefs that harm will come to her as she selects revision and suffers harm. After four revisions that lead to harm, the Target's belief exceeds the 85% threshold, and the Target stops. The game enters the coercion phase where there is no more revision and therefore the Target's beliefs do not change (We plot beliefs in this phase as grey triangles). The blue squares plot the Target's beliefs that a specific Competitor is inflicting harm. Because there is only one Competitor, both kinds of beliefs move at the same pace. Thus, to avoid attribution in period 4, the attribution threshold must also exceed 85%.

This has an important substantive implication. In real life, I expect the attribution threshold rarely exceeds the belief threshold necessary to sustain coercion. When it does, I expect that attribution is not the major issue complicating coercion. This is especially true when the Target is the one that inflicts harm on the Competitor. The reason is that if a Target is so sure that a specific Competitor is inflicting harm on her that she is usually willing to alter the course of her foreign policy, then she is also likely willing to retaliate against that Competitor. For example, suppose that each time the US broadcasts anti-regime statements inside Iran, mobs of Iraqi civilians stormed the US embassy in Iraq, leading to US diplomatic casualties. After a few protest events, the CIA may draw an estimate that Iran was responsible. To sustain unattributable coercion it would need to be the case that (a) the US was so sure Iran was responsible that they stopped broadcasting anti-regime sentiment, but (b) not sure enough that they were willing to retaliate against Iran for the harm they caused. It could be the case that the president is unwilling to punish Iran. But it unlikely that attribution is the issue. More than likely, the president weighs the costs and benefits of retaliation and concludes it is not useful. This reflects a more basic question of coercion, and not one of attribution.

The one-Competitor result helps clarify why Schelling's conjecture has persisted since 1960. Schelling theorized about one Competitor and one Target to match the Cold War context. Subsequent empirical studies of coercion frequently draw from this case (Nye 2017). Consistent with this logic, I find that unattributable coercion is incredibly hard to achieve in this case. Thus, it is reasonable that we have not found strong evidence for it in Cold War history.

But the one-Competitor case is not the likely scenario that western powers, especially the United States, will face in the future. The US, for example, is concerned that middle powers, or non-state actors are harnessing unattributable attacks. What happens if there is more Competitors?

Result 3b: The attribution threshold necessary to sustain unattributable coercion is decreasing in the number of Competitors. When the number of Competitors grows very large, the Target cannot attribute any attack to a specific Competitor. Thus, the Target draws no inference about which specific Competitor is responsible beyond the Target's prior belief about each Competitor's likely culpability.

Result 3b describes the easy of unattributable coercion as the number of Competitors increases. But is useful to think about the one Competitor problem as a special case, and two or more Competitors as sharing many similar properties. There are two reasons. First, as soon as we add a second Competitor it is reasonably simple to overcome Schelling's paradox. The technical analysis for the two Competitor model appears in Appendix B.5. But Figure 1(b) visualizes the results using the same parameters values as panel (a). Contrasting the red triangles across panels (a) and (b), we see that the belief necessary to sustain deterrence is insensitive to the number of Competitors. Thus, the Target is also deterred when she is 85% confident that some Competitor is out there harming her. What is more, the Target's belief that *some* Competitor is harming her increases at the same rate. As in the one-Competitor case, it takes four periods to achieve coercion.

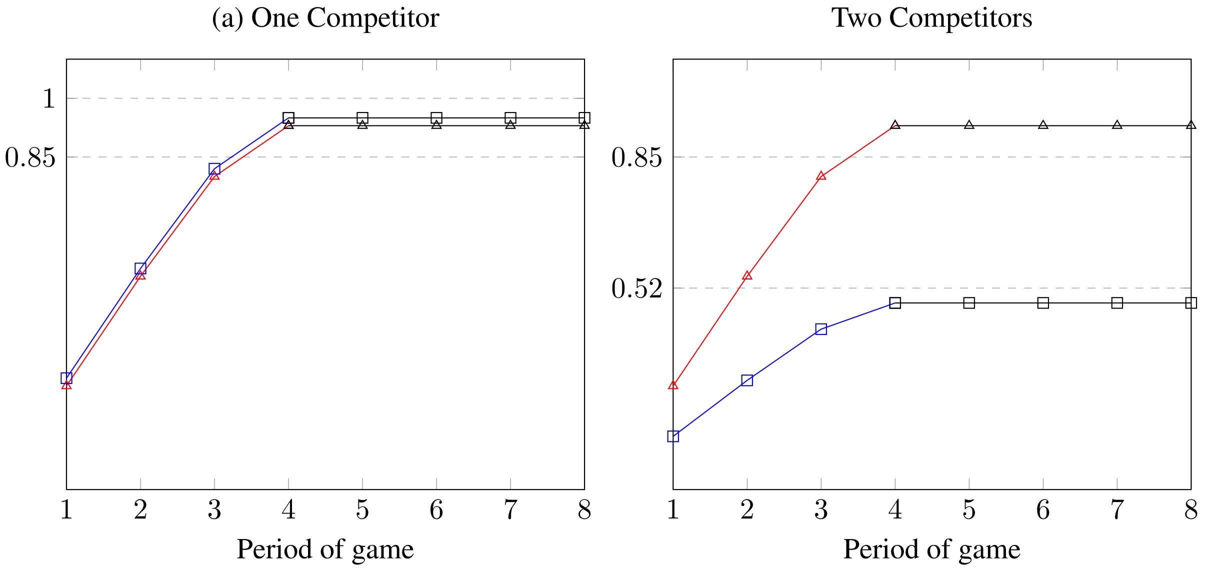
However, adding even one additional Competitor diminishes the Target's ability to attribute attacks to a *specific* Competitor. This is represented by the blue squares. In the 4 periods that the Target is selecting revision the Target updates about a specific Competitor at a slower rate than the one-period model. Also notice that even once we arrive at the coercion sub-game (period 4), the Target cannot attribute attacks to a specific Competitor with high confidence. Even though the Target is 85% sure that someone is attacking her, she cannot determine who is responsible for with more than 52% confidence—about as good as a coin flip.

For any number of Competitors, the most that the Target can attribute an attack to a specific Competitor is the lowest of either: slightly more than 1 divided by the total number of Competitors; or the prior belief that a specific Competitor is resolved.[17] This general result has an intuitive interpretation. Even if the Target is sure that someone is harming her, she does not know who it is. Thus, she distributes her expectation of blame to balance her prior belief that a specific Competitor is responsible and the number of Competitors that could harm her. The more Competitors, the harder it is to say that any one is responsible.

This logic helps alleviate the core of Schelling's paradox. But including even a second Competitor creates another strategic problem beyond what Schelling originally identified. Specifically, multiple Com-

[17] If the prior is large the results are trivial. The reason is that we see coercion in the first period because the Target believes that someone will attack her and never selects revision.

Figure 1: Shifting beliefs over time



Assumes risk of accident ($\lambda = .3$), initial expectation that a Competitor is resolved ($\psi = .05$), the harm the Target suffers from an attack ($c = 2$), and a discount rate ($\delta = .8$).

Colored Triangles represent Target's expectation some Competitor is punishing her given that the Target plays revision and suffers harm. Colored squares represent Target's ability to attribute that harm to a specific Competitor given the same history. After four revisions and harms, the Target's concern is sufficiently large that deterrence starts. Note beliefs in panel 1 are identical in equilibrium, I separated them to make the colors/shapes more legible.

petitors struggle to coordinate with each other. Coordination problems are exacerbated precisely because Competitors are attacking in secret. Since they are attacking in secret they cannot, by definition, reveal their intentions to each other and resolve this coordination problem.[18]

In appendix B.6 I study the model with an arbitrarily large number of Competitors. This variant of the model reveals a volunteer's dilemma. That is a problem where each highly resolved Competitor wants the Target to suffer harm, but wants to avoid the costs associated with it. In similar models with many potential Competitors, coercion unravels because each Competitor knows that other highly resolved Competitors exist who are also willing to punish. But since all apply this logic, no one is willing to volunteer to punish. Interestingly, my volunteer's dilemma is alleviated partly because each Competitor is uncertain if other Competitors are resolved. In standard coordination models, all Competitors know that there are others out there that want to punish. But when Competitors are uncertain about each other's resolve, they are not sure if the many other Competitors are the highly resolved types that are willing to inflict punishment. This added uncertainty drives highly resolved Competitors to punish even if there are many other Competitors out there.

In Appendix D, I extend the basic model to account for other coordination problems that arise from asymmetric learning. Specifically, if one Competitor does not attack in the first period, but observes harm come to the Target, he learns more than the Competitor who inflicts harm (and both learn more than the Target). This differential learning could cause the equilibrium to unravel. I show that the states can learn differently and sustain a coercion equilibrium.

A related coordination problem arises because when many Competitors launch attacks, the Target observes distinct episodes of harm. One might worry that the Target could exploit the number of harms to attribute attacks. With many Competitors this is not an issue precisely because the Target still cannot attribute attacks to a specific Competitor. Even with a small number of Competitors, this problem is alleviated if the number of accidents in any period occurs at random, and Competitors can launch multiple distinct attacks at the same time. Given that these coordination problems are surmountable, the results suggest that we should find cases of unattributable coercion when many potential Competitors exist.

[18] Condition 3 in and 8 governs a highly resolved Competitor's incentive to attack for the one-Competitor and two-Competitor models respectively. The coordination issue explains why perceptions of how resolved Competitors are appears in the latter but not the form.

2.1 Distinctive issues and preferences

In real life, foreign policy preferences are nuanced. This nuance provides both challenges and opportunities for the Target that may limit the real-life applicability of unattributable coercion. On the opportunities side, not all Competitors care about the same thing. As a result, the Target could exploit issues that one Competitor values but another does not to attribute attacks. On the challenges side, each policy choice activates different issue areas that a Competitor could care about. Even if the Target knows she is being attacked, she may not be able to figure out the logic behind the attacks because there could be several motivations. In this section, I loosely extend the logic of my theory to account for these real-life complications. This helps bring my rigorous but abstract formal results closer to the real world challenges that states face in the practice of coercion. As we shall see the features I identify somewhat depart from what we expect in coercion cases where attribution is not an issue. This is valuable because coercion and secrecy are both notoriously hard to evidence (Press 2007; Carnegie 2021). By providing more theoretical guidance, I can more precisely identify where I expect to find unattributable coercion, and what I expect it to look like.

2.1.1 Discriminating issues when Competitors hold divergent preferences

I assumed that all Competitors cared equally about all issues. In real life, different Competitor will object to some of the Target's foreign polices but not others, and the Target holds information about these discrepancies. For example, Russia and China both want to reduce US global influence. They likely share a common interest in limiting US influence in Latin America and Central Asia. But Russia may not care much about US influence in South East Asia, whereas China would.

This could create an opportunity for the Target to distinguish between two kinds of Competitors. Suppose the US had suffered a string of attacks for expanding its foreign policy agenda in South America. It may believe that both China and Russia could be responsible, but is not sure exactly who the perpetrator of violence is. The US, in theory, could discover the perpetrator by targeting its policies South East Asia (which for the purpose of this example, China is more likely to care about than Russia). If the Target can exploit discriminating issues, it could learn the attacker's true identity.

Appendix D shows that the theory is robust to this concern so long as there is a set of Competitors that hold enough common interests. The trick is that highly resolved Competitors launch unattributable attacks even when they don't hold a high value for the Target's revision opportunity. They are willing to incur the direct cost of harming the Target for issues they don't care much about in one period to (1) affect coercion; and (2) sustain anonymity in the long-run, so they can launch attacks over issues that they do care about. In fact, the fear of attribution enhances the Competitor's incentives to punish the Target for revision over issues that the Competitor does not care about. Thus, it can create resolve to attack absent a direct incentive to contest a disputed issue. Returning to our example, if the US did not suffer following an intervention in South East Asia, then the US would infer that China was not responsible for the long history of previous attacks. Therefore, Russia is likely to be.

Putting it altogether, the extension uncovers two more nuanced predictions about patterns of unattributable coercion. First,unattributable attacks will cover both issues of common interest to several potential Competitors, as well as issues that are of interest to only one (or a handful of) likely Competitors. As a result, we should be looking for patterns across a broad range of similar issues from the Target's perspective, and not a set of issues that are salient to a specific Competitor. Second, Targets will not rule out specific Competitors as the likely perpetrators of attack even when they observe attacks over issues that specific Competitor does not care about. The reason is that the Target understands that Competitors must launch attacks over a broader set of conditions to both generate a pattern and preserve anonymity. Thus, if we observe an attack in East Asia, the US should not rule out that Russia was responsible for it.

2.1.2 Multi-dimensional policies and calling cards

A different problem is that each policy choice a Target faces is multi-dimensional. Competitors may dislike some dimensions of the Target's choices but not others. For example, suppose the United States intervened against Syria for human rights abuses. Lots of potential Competitors could dislike the United States' choice, but for different reasons. Human rights abusers may dislike it because they fear that they will be next. Others may dislike that the US violated the sovereignty norm. Others may benefit from Syria's natural gas pipelines and worry that the US intervention will disrupt their gas supply. Syria's neighbors and terrorist groups that operate across the Syrian border (e.g. ISIS) may dislike that the US meddled in their operational area. Any of these actors could harm the US in response to the human rights intervention in Syria. But the US would not know whether the harm was inflicted because the Competitor disliked that the US intervened using force, intervened over human rights, whether the US intervened in Syria specifically, or a Middle Eastern country in general. These concerns are amplified when the Target can suffer harm by random chance.

The basic result, the Target eventually learns, survives this sort of complexity. No matter how many dimensions we add to the policy space, the Target will eventually infer which policy areas the Competitors who attack her care about. However, if we make no other changes to the theory, it takes the Target much longer to learn. And the time it takes for the Target to learn ultimately depends on the order of issues that arise.

For example, suppose a Target first intervenes in Syria over a human rights issue, and then is attacked. Then suppose the Target intervenes over a human rights issue in Lebanon and is attacked. After these two events, the Target infers that either Competitors only care about human rights, that Competitors care about both Syria and Lebanon but not human rights, or that Competitors care about some combination of human rights, and either Lebanon or Syria (or both). With more events, the Target can eventually figure it out. But what the Target learns is context dependent. For example, imagine a third event where the Target intervenes to protect human rights in Bulgaria, and does not suffer harm. In this case, the Target infers that human rights in general is not the reason behind harm. Looking back at the first two events, the Target also learns that Lebanon and Syria are both likely salient for Competitors and the source of attacks. Given enough diversity across events, the Target will eventually come to learn the logic behind attacks.

Learning in a multi-dimensional space is slower. However, Competitors can partially speed the process along through distinctive attacks. By distinctive, I mean that the Competitor gets to choose the qualitative features of the attack he launches. At the broadest level, the Competitor can choose between different kinds of attacks: assassinations, car bombs, terrorist attacks, efforts to depose a leader via a revolution, etc.[19] At a more specific level, the Competitor can embed calling cards within an attack. For example, a Competitor could embed a specific line of code in their cyber-attacks in response to a specific transgression by the Target. The Competitor can also vary the asset that is destroyed. For example, in the case of assassinations, the Competitor can assassinate a foreign leader, the foreign leader's oldest living daughter, or a foreign leader's oldest brother. In the case of cyber-attacks, the Competitor can take down power plants, financial systems, electoral systems, etc. By using a specific, distinctive attacks against a specific policy dimension, the Competitor can clarify the logic behind the attack without explicitly stating it.

Broadly, many attack dimensions can off-set many policy dimensions. But the results are highly sensitive to equilibrium selection. With the strongest common conjectures, highly distinctive calling cards can

[19] This argument is similar to the ransomware argument where the Competitor reveals his exact demands anonymously Jun (2022). But ransomware assumes that the Competitor explicitly states the logic, which means the Target knows the exact logic behind an attack. The setting I describe is akin to sending a less precise message, that is refined over time via trial and error.

lead to rapid learning in a multi-dimensional policy environment. With the weakest common conjectures (Competitors deploy attacks at random when they see a policy they want to deter) distinctive attacks become meaningless and it takes a long time to achieve unattributable coercion. There are other equilibria where attacks take some, but not too much, time to generate coercion. For example, the Target believes the Competitor will assign calling cards that are substantively linked to attacks (e.g. attacks against a trade attache to oppose trade meetings).

Despite the multiplicity problems, it is clear that the welfare of all players increases in cases where the Target learns quickly (rather than slowly) because we reach coercion (and therefore minimize harm) faster. Thus, I informally expect that Competitors in real life settings will develop a calling card and repeatedly apply that calling card in response to the Challenger's transgressions. I also expect that Targets will search for calling cards in qualitative similarities in attacks. Thus, they will not only look for attacks that come shortly after their choices, they will derive inferences about what the Competitor could want based on qualitative features of how they suffer.

3 The future is (somewhat) grim: Havana Syndrome

My theory explains both why unattributable coercion was uncommon throughout the history of great power politics; and also why it could be commonly used against the US and other western democracies that hope to preserve the Liberal Order in the future. I found that when there is only one plausible Competitor, unattributable coercion is hard to maintain. As Nye (2017) argues, historical studies of coercion have focused mainly on Soviet-US interactions during the Cold War, and mainly considered the US perspective. Only the Soviet Union had the capability to plan and execute repeated attacks against the US that would inflict serious harm and leave no direct evidence.[20] It follows that if Western allies suffered a string of unlikely attacks that held similar features, they would have assumed that it was the Soviets. A similar story can be told for how the US exploited coercive actions against communist states.

I also find that as we increase the number of Competitors unattributable coercion becomes much easier. This is especially true if the harm caused by unattributable attacks is large relative to the Target's policy value, and the attacks hold distinctive features that Targets are likely to realize are part of a pattern. This result creates opportunities for unattributable coercion in the modern world. Changes in technology have

[20] While others had covert capabilities, they were either limited regionally, or were unlikely to work repeatedly and remain hidden. There are, potentially exceptions.

placed powerful, unattributable weapons in the hands of many violent political actors (Horowitz 2020). It is now common for middle powers to wield sophisticated covert operators. This gives them the ability to secretly fund and train insurgents and terrorists, or assassins. The US is also reliant on the internet. Both middle powers and non-state actors now hold the ability to launch cyber-attacks against the US government, and influence elections using social media (Levin 2021). With so many actors, who hold such diverse capabilities, there are many plausible Competitors that can harm us. My theory suggests these conditions are ripe for unattributable coercion.

As past scholars have argued, non-rationalist factors could limit the real-world application of unattributable coercion. For example, elites face psychological and bureaucratic barriers to drawing completely rationalist inferences (Schub 2023; Jervis, Lebow, and Stein 1989). This may prevent real-world Targets from inferring a pattern when a perfectly rational Target would have. These common arguments stack the deck against my theory. If they are dominant, I should not be able to recover evidence of unattributable coercion in the modern world.[21] However, my goal is to illustrate that unattributable coercion plausibly applies in the modern world.

In what follows, I illuminate a plausible case of unattributable coercion in the modern world: Havana Syndrome attacks against US embassies. As we shall see, my theory is especially useful for illuminating the features of coercion in this case, because they comply with my distinctive logic, and not necessarily the standard features of coercion cases (e.g. clear, unambiguous threats).

Havana Syndrome is a set of unexplained medical symptoms first experienced by U.S. State Department personnel and Canadian diplomats stationed in Cuba in 2016. Because the harm it causes is mysterious, there is no clinical definition for it. But the CDC adopts a working case definition that requires the onset of symptoms in two phases. The first phase includes at least one of head pressure, disorientation, nausea, headache, vestibular disturbances, auditory symptoms, or vision changes. The second phase includes vestibular disturbances or cognitive deficits, with no readily recognizable alternate.[22] As it stands, the State Department actively studies 200 cases (i.e. former Embassy staff and their families) as potential cases of Havana Syndrome.[23]

The initial details of the case match the prior parameters where my theory finds unattributable coercion

[21] Of course, the world is multi-causal. Just because I can support unattributable coercion in a case, does not mean that these explanations are not valuable in other cases.

[22] https://nsarchive.gwu.edu/briefing-book/cuba/2021-02-02/cdc-report-havana-syndrome-medical-mystery-remains-unresolved.

[23] https://www.state.gov/wp-content/uploads/2022/02/JASON-Study-Revised_10-February-2022-Redacted_V1.1.pdf

(Bates 1998, recommends selecting cases this way.). The attacks are repeated, no one has claimed responsibility for them, and the CIA has recovered no direct evidence that they are attacks, let alone that someone is behind them (validating x_t as distinct from the Competitor's choice to inflict harm).[24] As just noted, the harm these attacks cause is severe (c is large). The response to Havana syndrome illustrates that Congress places enormous value on the safety of diplomatic and intelligence personnel. In signing the Havana Act, Intelligence Committee Chair, Mark Warner states, "Every day, American diplomats and intelligence officers around the world put themselves at risk to keep our nation safe. In return, we have an obligation to provide ample support when these brave men and women are injured in the line of duty."[25] The Act provides extensive medical and financial support to victims, and calls on the US government to root out the cause of Havana Syndrome.[26] Finally, and as I will detail more below, the US believes that the harm suffered could be the result of an accident, or a deliberate attack (λ is non-zero). If it is an attack, the US estimates it could plausibly be launched by several different rivals including Iran, North Korea, Russia and China ($J > 1$, and possibly 4 or 5).

Havana Syndrome is also a good test of my theory because it supplies all the real-life complexity that Schelling conjectured would make it difficult to infer a pattern from history. My theory expects Targets to look across all of the harm that they suffer, identify episodes that could be part of an attack, and try to infer a connected pattern of harm. It is not clear that the CIA would notice a pattern in this case. One reason is that the symptoms are diverse, and many of these symptoms are consistent with illnesses that exist at Embassy posts.[27] The US government also does not have a clear medical theory about what could cause the symptoms that present in the unusual cases. For example, in 20 cases, it appeared that the brain chemistry of Embassy staff had changed in inexplicable ways. The US Government does not know what could cause these symptoms.[28] Because the Government cannot identify the mechanism that causes harm to staff, they cannot easily identify whether cases are connected. I theory, this should make it difficult to infer

And yet, consistent with my theory, the US government quickly inferred that the symptoms were strange, and sought to establish a pattern in Cuba, and beyond. A closer look at how estimates unfolded helps

[24] https://www.wsj.com/articles/havana-syndrome-symptoms-11626882951

[25] https://www.warner.senate.gov/public/index.cfm/2021/10/u-s-sen-mark-r-warner-s-bill-to-support-havana-syndrome-victims-signed-into-law#:~:text=The%20legislation%2C%20which%20passed%20Congress,%2C%20Cuba%2C%20beginning%20in%202016

[26] https://www.congress.gov/bill/117th-congress/senate-bill/1828/text?r=3&s=1.

[27] See https://www.state.gov/wp-content/uploads/2022/02/JASON-Study-Revised_10-February-2022-Redacted_V1.1.pdf

[28] After extensive study of each case, medical researchers propose electro-magnetic pulses, poison, pollution, and an unlucky combination of diseases as plausible explanations. See https://econpapers.repec.org/article/abfjournl/v_3a36_3ay_3a2021_3ai_3a3_3ap_3a28508-28510.htm.

illustrates how the logic of my theory applies. In 2016, six staff and family members at the US and Canadian Embassy in Cuba reported hearing unusual noises, headaches, nausea and vomiting. After examining the patients in Cuba, doctors could not reach a definitive finding. They noted that it was possible that these were attacks, but it was also possible that Embassy staff, who spend time together, suffered from an illness. One potential source of concern was that "the first incident occurred in Havana when U.S.-Cuba relations were rapidly changing amidst the Obama-Trump presidential transition (Power and Miner 2021)." Initially, the US State Department issued a warning about unexplained symptoms. It instructed Cuban staff to take health precautions including hand washing, and isolation if they displayed symptoms. It also instructed global Embassy staff to report similar symptoms. In early 2017, the State Department reported that, 'Embassy Havana employees have been targeted in specific attacks,' and reduced Embassy staff.[29]

Within a year, Embassy staff in Russia, Georgia, Taiwan, Australia, Columbia had reported similar symptoms. Again, the State Department did not instantly issue a warning of attacks.[30] But they did notice a pattern. The Office of the Director of National Intelligence (ODNI) commissioned a study into the chance the cases were connected. As part of this investigation, diplomatic security personnel in the afflicted countries examined the homes and routes of the harmed staff to search out potential explanations. Finally, the ODNI contracted outside scientists to develop theories as to the source of Havana Syndrome. A declassified summary of the original National Intelligence Estimate could not rule out the possibility of an attack, but held that local illnesses were a likely cause.

As the number of potential documented cases increased, estimates started to shift. In 2020, several Intelligence Agencies released estimates about the likely cause of Havan Syndrome. While estimates conflicted, they were all more confident that a rival state was responsible than the initial 2016 estimate. The most extreme estimate found 'a substantial likelihood of wrongdoing.'[31] But even the all-source ODNI estimate considered an attack a real possibility. This report was not the result of new medical evidence, or direct evidence that an attack had occurred. Rather it followed from the fact that US Government staff stationed overseas in 10 countries suffered unusual symptoms with different but common properties.

In terms of attribution there was also disagreement. Based on the location of reported symptoms, and beliefs about relative covert capabilities, a significant group believed Russia was responsible. But many senior analysts, including CIA Direct Haspel, believed that China, Iran or a combination of other actors

[29] https://cu.usembassy.gov/security-message-u-s-citizens-cuba-travel-warning/

[30] One reason is that they believed their global request to be on the look out sparked hyper-vigilance amongst Embassy staff.

[31] https://www.nytimes.com/2020/10/19/us/politics/diplomat-attacks-havana-syndrome.html

were plausible sponsors.[32]

This pattern of inferences is consistent with my theory in five ways, not all of which fit the common logic of explicit coercion. First, these attacks exhibit a calling card feature. They all present as medical symptoms in overseas Embassy staff. While the symptoms do vary, they share enough common features that the US Government can classify them. Furthermore, the combination of symptoms is unusual to some degree. This was useful because it alerted Diplomatic Security to them.

Second, the US government was able to connect the dots, and then retroactively research past cases to discern a logic behind them. That is, they noticed a pattern of unusual combinations (but not entirely identical) symptoms across many different diplomatic missions over several years. They did not instantly infer a connection, but thought these symptoms warranted further study.

Third, initial estimates expressed low confidence that an attack was likely. But as the number of events increased, estimates that an attack occurred also increased. Critical for my theory, the increased confidence did not rely on direct evidence that an attack had happened. Rather, the repeated presence of strange symptoms across different Embassies was enough to infer that some of the harm was deliberate with moderate confidence.

Fourth, while the ODNI sharply raised their beliefs that the harm was deliberate, they did not make as strong estimates about who was behind the attacks. Rather, every estimate leaves this possibility open. What is more, their estimates about the most likely sponsor (Russia) was based on prior beliefs of Russia's capabilities and interests (reflecting a larger ψ for Russia).[33] In fact, the focus on Russia helps validate the model extension described in section 2.1. I showed that to sustain plausible deniability the real attacker must continue attacks over issues that fall outside their interests, but that would interest other plausible attackers. Notably, estimates that Russia was responsible did not shift after attacks in rural China, or other areas that fell outside Russia's likely policy interest.

Of course, just because the US infers that attacks cause Havana Syndrome, it does not mean my theory applies. For unattributable coercion to hold the US must (a) infer a reason for attacks; and (b) alter their behavior. Again the complexity of this case makes unattributable coercion tough. The US has a multi-dimensional policy agenda. Embassy staff have reported symptoms consistent with Havana Syndrome in Columbia, Serbia, Russia, Poland, Georgia and Taiwan, China, Vietnam and France. These 200 staff fulfill

[32] https://www.nytimes.com/2021/03/04/us/politics/cia-havana-syndrome-mystery.html

[33] Although I cannot confirm it, interviews with intelligence elites suggest that the US Government did not retaliate against Russia. This fits my definition and conceptualization of the attribution threshold.

different Embassy roles (however a large number are likely CIA staff on non-official cover). There was also variation in that some staff were employed at the main Embassy, and others were employed at consulate offices in remote locations. It is also possible that multiple perpetrators are inflicting the attacks for different reasons. Indeed, this level of complexity makes it hard for the US to figure out why they are suffering attacks. Thus, we might expect that the US would not alter their behavior.

Despite these difficulties, US behavior supports my theory. As stated, the State Department demoblized some staff in Cuba in 2017. Following attacks against covert operators and other intelligence staff in Eastern Europe, the CIA demobilized intelligence officers in Serbia.[34] Although the CIA does not publicly disclose its Embassy deployments, and they have not publicly commented in detail on the matter,[35] they clearly appreciated that their staff (and not actual State Department employees) were at high risk and this caused them to alter their deployments.

In another example, Vice President Kamala Harris delayed a trade visit to Hanoi because of a report of a potential attack against a US Trade Attaché in Vietnam.[36] This episode illustrates how quickly common conjectures can form in a complex policy environment. Specifically, diplomatic security took it as a signal that the harm was suffered by the Trade Attaché that it related to an upcoming Trade visit. Consistent with my theory, this could have been a coincidence. However, the US was not willing to risk the health of the second most important politician in the United States in case the connection was meaningful.

In yet another example, the United States drew down staff in a rural Chinese consulate as the result of repeated harm suffered by staff that worked there.Different still, the US closed its Embassy in Havana for six years as the result of repeated attacks. Put in contrast with the Harris episode in Vietnam, these two cases further illustrate my theory. Since the future risk was towards low-level embassy staff and not the Vice President, it took the US longer to alter their strategic behavior in response to the risk.

3.1 Concerns

The US Government continues to study Havana Syndrome and update estimates. The most recent unclassified estimate found the collection Havana Syndrome reports are unlikely (although with dissent from two of seven agencies) the result of an attack.[37] It is worth noting that the estimate does not clear state

[34] https://www.dailymail.co.uk/news/article-10038159/CIA-evacuates-intelligence-officer-Serbia-result-Havana-Syndrome.html..

[35] Although DCI Burns has recently acknowledged that CIA staff have suffered acutely.

[36] https://www.cnn.com/2021/08/24/politics/kamala-harris-vietnam/index.html.

[37] https://www.dni.gov/files/ODNI/documents/assessments/Updated_Assessment_of_Anomalous_Health_Incidents.pdf

with high confidence that no reported case is the result of an attack. This estimate is not fully supported throughout the Government. The Department of Defense, for example, has launched a subsequent review to clarify their position. There is also evidence that senior intelligence elites believe that some cases fit an attack. For example, CIA Director Burns was upset when one of his staff reported Havana syndrome symptoms during the Director's visit to India.[38] Despite dissent and ambiguity, the most recent estimates is certainly less confident that the harm suffered followed from an attack than estimates from 2020 and 2021.

The downgraded estimates causes us to question whether Havana Syndrome is actually the result of an attack. But from the perspective of my theory it does not matter. What matters is that the US inferred a pattern from events and altered their behavior. In fact, these downgraded estimates provide additional support for my theory. After the revised estimate, the US re-opened the Consulate in China and Embassy functions in Cuba they closed in response to Havana Syndrome incidents.[39] Consistent with my theory, the US Government shut down embassies when it inferred a logic behind an attack, but when it down graded its estimate (i.e. lowered its expectation that it would strategically suffer harm), it deviated from being coerced.[40]

4 Conclusion

Against a 60-year-old conventional wisdom, I argue that rationalist, unattributable coercion is possible. I show it is most likely when a Target worries about more than one plausible Competitor, the harm the Target suffers from attacks is large, and the risk of harm from unrelated sources is not too large. I argued this result held grim policy implications because modern states face many capable adversaries that can inflict serious harm. Shifts in US diplomatic and intelligence practices in response to unattributed Havana Syndrome attacks suggest that unattributable coercion may already influence US policy.

My theory shows unattributable coercion is not only plausible, it arises in periods that coercion researchers would not intuitively look. As Nye (2017) has recently argued, our' ”minds [are] captured by Cold War images of massive retaliation to a nuclear attack by nuclear means.” Thus, we usually search for evidence of coercion in bilateral cases between great powers, and often focus narrowly on crises where

[38] https://www.cnn.com/2021/09/20/politics/cia-director-havana-syndrome-india-trip.

[39] https://www.scmp.com/news/world/americas/article/3169178/us-reopen-cuba-consulate-closed-after-mysterious-havana..

[40] In my model, estimates increase with events. But my theory does not include the kind of qualitative estimates that the CIA has conducted here. It is consistent with my overall argument that if beliefs harm would come decrease, the Target would no longer remain coerced, and revert to its original policies.

leaders make explicit and public threats. I showed that unattriutable coercion likely arises when there are many perpetrators of an attack (i.e. not bilateral cases), low-to-moderate risk that harm was caused by an accident, the attack inflicts severe harm on the Target relative to the policy goals, and the direct cost of launching the attack (i.e. the financial and manpower costs) for the Competitor are not too large. Similarly, I argue that the observable indicators of unattributable coercion are not the same as overt coercion. For example, unattributable coercion ought to include efforts made by Targets to search for patterns in the harm they suffer.

Finally, I reconcile a growing discrepancy between policy-makers and academics about the future of coercion with the proliferation of modern weapons (Horowitz 2020). Policy-makers are witnessing the coercive power of covert operations, like Havana Syndrome, and are sounding the alarm about future unattributable coercion. For example, the 2023 Annual Threat Assessment, which reports "the collective insights of the [US] intelligence community", noted that US rivals will use unattributed cyber-operations "to *deter* U.S. military action by impeding U.S. decision-making, inducing societal panic, and interfering with the deployment of U.S. forces." But the exact logic is vague, in part because the IC does not clearly understand how unattributable coercion works. Academics[41] have characterized this threat assessment as alarmist, based, in-part, on Schelling's conjecture that unattributable coercion is impossible. By overturning Schelling's conjecture, I validate the IC's intuition. By detailing the factors necessary to achieve unattributable coercion, I help the IC better understand how our rivals will wield it, and which unattributable weapons will pose the greatest coercive leverage. I expected that the damage from cyber-attacks may not be sufficiently large to off-set the value of major US policy initiatives. But this is a temporary feature. As these weapons become more sophisticated, and we increasingly rely on machines for banking, power, operating damns and other systems that could cause enormous harm if tampered with, the threat of these attacks will grow, and, consistent with IC estimates, unattributable coercion will become a problem.

[41] https://warontherocks.com/2023/05/are-we-asking-too-much-of-cyber/

References

Arena, Philip and Scott Wolford (2012, 6). Arms, intelligence, and war. *International Studies Quarterly 56*, 351–365.

Ashworth, Scott (2012, 6). Electoral accountability: Recent theoretical and empirical work. *Annual Review of Political Science 15*, 183–201.

Axelrod, Robert and Rumen Iliev (2014, 1). Timing of cyber conflict. *Proceedings of the National Academy of Sciences 111*, 1298–1303.

Baliga, Sandeep, Ethan Bueno De Mesquita, and Alexander Wolitzky (2020, 11). Deterrence with imperfect attribution. *American Political Science Review 114*, 1155–1178.

Baliga, Sandeep and Alexander Wolitzky (2018). Deterrence with imperfect attribution.

Bates, Robert H. (1998). *Analytic narratives*. Princeton University Press.

Borghard, Erica D. and Shawn W. Lonergan (2017, 7). The logic of coercion in cyberspace. *Security Studies 26*, 452–481.

Brutger, Ryan and Joshua D. Kertzer (2018, 5). A dispositional theory of reputation costs. *International Organization 72*, 693–724.

Canfil, Justin Key (2022, 1). The illogic of plausible deniability: why proxy conflict in cyberspace may no longer pay. *Journal of Cybersecurity 8*.

Carnegie, Allison (2021, 5). Secrecy in international relations and foreign policy. *Annual Review of Political Science 24*, 213–233.

Carson, Austin (2018). *Secret wars : covert conflict in international politics*.

Carson, Austin and Keren Yarhi-Milo (2017, 1). Covert communication: The intelligibility and credibility of signaling in secret. *Security Studies 26*, 124–156.

Colgan, Jeff D. (2021). *Partial Hegemony: Oil Politics and International Order*. Cambridge University Press.

Dafoe, Allan, Jonathan Renshon, and Paul Huth (2014, 5). Reputation and status as motives for war. *Annual Review of Political Science 17*, 371–393.

Danilovic, V. (2001, 6). The sources of threat credibility in extended deterrence. *Journal of Conflict Resolution 45*, 341–369.

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

Gartzke, Erik and Jon R. Lindsay (2015, 4). Weaving tangled webs: Offense, defense, and deception in cyberspace. *Security Studies 24*, 316–348.

Greenhill, Kelly and Peter Krause (2018). *Coercion: The Power to Hurt in International Politics* (1 ed.). Oxford University Press.

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

Horowitz, Michael C. (2020, 5). Do emerging military technologies matter for international politics? *Annual Review of Political Science 23*, 385–400.

Huth, Paul and Bruce Russett (1993, 3). General deterrence between enduring rivals: Testing three competing models. *The American Political Science Review 87*, 61–73.

Jervis, Robert, Richard Ned. Lebow, and Janice Gross. Stein (1989). *Psychology and deterrence*. Johns Hopkins University Press.

Joseph, Michael F and Michael Poznansky (2018, 5). Media technology, covert action, and the politics of exposure. *Journal of Peace Research 55*, 320–335.

Jun, Jenny (2022). Coercion in cyberspace: A model extortion via encryption.

Kertzer, Joshua D. (2017, 4). Resolve, time, and risk. *International Organization 71*, S109–S136.

Krcmaric, Daniel (2019, 6). Information, secrecy, and civilian targeting. *International Studies Quarterly 63*, 322–333.

Kurizaki, Shuhei (2007, 8). Efficient secrecy: Public versus private threats in crisis diplomacy. *American Political Science Review 101*, 543–558.

Kydd, Andrew H. and Roseanne W. McManus (2017, 2). Threats and assurances in crisis bargaining. *Journal of Conflict Resolution 61*, 325–348.

Kydd, Andrew H. and Barbara F. Walter (2006, 7). The strategies of terrorism. *International Security 31*, 49–80.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain (2018, 3). The science of fake news. *Science 359*, 1094–1096.

Levin, Dov H. (2021). *Meddling in the ballot box : the causes and effects of partisan electoral interventions*.

Levy, Jack S., Michael K. McKoy, Paul Poast, and Geoffrey P.R. Wallace (2015, 10). Backing out or backing in? commitment and consistency in audience costs theory. *American Journal of Political Science 59*, 988–1001.

Libicki, Martin (2009). *Cyberdeterrence and Cyberwar.* RAND Corporation.

Lin-Greenberg, Erik (2023, 1). Evaluating escalation: Conceptualizing escalation in an era of emerging military technologies. *The Journal of Politics*.

Lindsay, Jon R. (2015, 11). Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack. *Journal of Cybersecurity*, tyv003.

McManus, Roseanne W (2014, 11). Fighting words. *Journal of Peace Research 51*, 726–740.

Min, Eric (2022). Taking responsibility: When and why terrorists claim attacks.

Myrick, Rachel (2020, 7). Why so secretive? unpacking public attitudes toward secrecy and success in us foreign policy. *The Journal of Politics 82*, 828–843.

Nye, Joseph S. (2017, 1). Deterrence and dissuasion in cyberspace. *International Security 41*, 44–71.

Power, Sean and Michael Miner (2021). Report – havana syndrome: American officials under attack.

Poznansky, Michael (2019, 1). Feigning compliance: Covert action and international law. *International Studies Quarterly*.

Poznansky, Michael (2022, 6). Revisiting plausible deniability. *Journal of Strategic Studies 45*, 511–533.

Poznansky, Michael and Evan Perkoski (2018, 10). Rethinking secrecy in cyberspace: The politics of voluntary attribution. *Journal of Global Security Studies 3*, 402–416.

Press, Daryl G. (2007). *Calculating Credibility: How Leaders Assess Military Threats*. Cornell University Press.

Renshon, Jonathan, Allan Dafoe, and Paul Huth (2018, 1). Leader influence and reputation formation in world politics. *American Journal of Political Science*.

Schelling, Thomas C. (1960). *The strategy of conflict*. Harvard University Press.

Schram, Peter (2022, 3). When capabilities backfire: How improved hassling capabilities produce worse outcomes. *The Journal of Politics*.

Schub, Robert (2023). Informing the leader: Bureaucracies and international crises. *American Political Science Review*.

Schultz, Kenneth A. (2001, 2). Looking for audience costs. *Journal of Conflict Resolution 45*, 32–60.

Spaniel, William and Michael Poznansky (2018, 7). Credible commitment in covert affairs. *American Journal of Political Science 62*, 668–681.

Uzonyi, Gary, Mark Souva, and Sona N. Golder (2012, 12). Domestic institutions and credible signals. *International Studies Quarterly 56*, 765–776.

Weeks, Jessica L. (2008). Autocratic audience costs: Regime type and signaling resolve. *International Organization 62*, 35–64.

Weisiger, Alex and Keren Yarhi-Milo (2015, 5). Revisiting reputation: How past actions matter in international politics. *International Organization 69*, 473–495.

Wohlforth, W (2009, 12). Unipolarity, status competition, and great power war. *World Politics 61*, 28–57.

Wolford, S., D. Reiter, and C. J. Carrubba (2011, 8). Information, commitment, and war. *Journal of Conflict Resolution 55*, 556–579.

Wood, B. Dan. (2012). *Presidential saber rattling : causes and consequences*. Cambridge University Press.

A Core Coercion Model

Consider a game between a Target A and Competitor B. Denote an arbitrary period as t. Each period unfolds in the same way. A chooses revision or not. Regardless of what A does, B chooses to harm or not. Denote player A's choice in period t as $a_t \in r, nr$, and B's as $b_t \in h, nh$. A strategy for A is $s^A = a_1, a_2, a_3...$, and B is $s^B = b_1, b_2, b_3....$

Period payoffs are

$$U_t^A = 1 \times (a_t = r) - c \times (b_t = h)$$

$$U_t^B = \pi \times (a_t = nr) - k \times (b_t = h)$$

Here the bracketed conditions represent indicator functions equal to 1 if the condition in the bracket is satisfied, and 0 otherwise. Notice that the harm B inflicts is punitive. This fits the standard definition of coercion via punishment. The game repeats over an infinite horizon and players discount future payoffs by δ.

A.0.1 Analysis

I solve for sub-game perfect equilibria. In the manuscript, I described the on path features of the coercion equilibrium. Consistent with that informal definition, I define a pair of strategies as coercion starting in period t' if they include the following features. A's on-path strategy must include $a_t = nr$ for every $t \geq t'$, and that B's on-path strategy must include $b_t = nh|a_t = nr, b_t = h|a_t = r$ in every $t \geq t'$.

Before I characterize a complete the complete strategy profile in a coercion equilibrium, I need to solve for the off-path punishments. Thus, it is first necessary to identify the equilibrium that serves as the reversion point if players deviate from the coercion equilibrium.

Lemma A.1 ***The revision equilibrium:*** *The following pair of unconditional strategies is a subgame perfect equilibrium for all parameter values.* $s^A = r, r, r....r$, $s^B = nh, nh, nh...nh$. *In an arbitrary t, expected utilities are* $EU_t^A = \frac{1}{1-\delta}, EU_t^B = 0$

Consider an arbitrary period t. If B deviates, he does k worse than if he remains on path. If A deviates, she does 1 worse than if she remains on the path. Since the game is stationary, this completes the proof.

I now identify one coercion equilibrium.

Proposition A.2 ***Coercion (as an equilibrium):*** *If*

$$k < \frac{\delta\pi}{1-\delta} \tag{1}$$

$$c > 1 \tag{2}$$

then a coercion equilibrium exists. In it, on path strategies are as follows. $a_t = nr, b_t = nh|nr, h|r$. *Off path, players revert to the strategies described in Lemma A.1 in period* $t' + 1$ *if in period* t' *we observe* $a_{t'} = r, b_{t'} = nh$.

We've proven the the total reversion equilibrium holds. This solves for the off-path subgame starting at $t' + 1$. Now consider, B's incentive to punish in the off-path condition where he observes $a_{t'} = r$. If he harms, the game continues in the on-path coercion equilibrium. If he does not harm, the game reverts to the total revision sub-game. B prefers to harm if $-k + \frac{\delta\pi}{1-\delta} > 0$ as desired. Working backwards, A prefers to deviate to revision at t if $1 - c + \delta 0 + \delta^2 0... > 0 + \delta 0...$ as desired. This completes the proof.

A.0.2 Broader definition of coercion

More broadly, I define an equilibrium as a coercion equilibrium if for some history of the game, A's on-path strategy starting at t is $s^A = nr_t, nr_{t+1}, nr_{t+2}.....$ there exists an on-path sub-game starting

This definition is strict in one sense. It assumes total coercion in at least one on-path sub-game. That is, A never invests once the coercion subgame starts. But it is flexible in two respects. First, it does not assume coercion start in the first period. This allows states to learn and achieve coercion over time. Second, it only arises in some on-path subgames. This is important because we will assume variation in B's type. Weakly resolved types will rarely enter the coercion equilibrium.

B Model with attribution (Main model)

In what follows I report the results of the model presented in Table 1. Results 1,2 3a, 3b are derived from the analysis of this model.

B.1 Set-up

I now extend the core model to allow for an attribution problem. First, assume J Competitors, and allow each Competitor's level of resolve to vary. We write $B_j = H$ if Competitor j's resolve is high, and $B_j = L$ if B_j's resolve is low. B_j's type determines his value for the issue in dispute. That is, $B_j = H \implies v_j(\pi) = \pi_H$, and $B_j = L \implies v_j(\pi) = \pi_L$

The model proceeds as follows. Before the first period commences, Nature draws each Competitor's type $pr(B_j = H) = \psi, pr(B = L) = 1 - \psi$, i.i.d.

Then, the game repeats over infinitely many periods. An arbitrary period is t. Each period unfolds as follows.

- A selects revision or not. (public) $a_t \in r, nr$
- Each B_j simultaneously decides to inflict harm on A or not (private) $j_t \in h, nh$
- Regardless of what has happened before, Nature harms A with probability λ, and does not harm A with probability $1 - \lambda$ (private)
- If any Competitor or Nature has harmed A in t, Nature reveals to all that A has suffered harm ($x_t = h$) and does not otherwise ($x_t = nh$).
- Per period payoffs are realized.

To be clear about information, each Competitor observes his own type, A's action and D's own action, and whether A suffers harm or not. But does not observe the choices of any other Competitor or their types. A observes whether she suffers harm or not (x_t). But A does not observe B's type (ψ) at the beginning of the game, and does not know if she suffered harm as the result of an accident λ or a Competitor's choice, or which specific Competitor inflicted harm.

Define $\beta_{tj} = E(B_j = H|s^j, s^A, \lambda, x_t)$ as A's belief in period t that B_j is highly resolved. Define a plausible deniability parameter $\hat{\beta} \in (0, 1)$ as the plausible denaibility threshold.

Period t payoffs are

$$U_t^A = 1 \times (a_t = r) - c \times (x_t = h)$$

$$U_t^j = v(\pi) \times (a_t = nr) - k_d \times (j_t = h) - k_i \times (\beta_{jt} > \bar{\beta})$$

The bracketed terms are indicator functions equal to 1 if the condition is true. $(a_t = r)$ is A selects revision. $(x_t = h)$ is A suffers harm. $(j_t = h)$ is a specific Competitor inflicts harm. $(\beta_{jt} > \bar{\beta})$ says that A's belief that $B_j = H$ exceeds the plausible deniability threshold.

We dis-aggregate the direct k_d (i.e. financial, attention, casualty) cost of launching an attack, from the indirect k_i cost. The indirect cost are things that could only follow if the attack was attributed.

Total expected utilities are time discounted such that: $EU_t^A = U_t^A + \sum_{\tau=1}^{\infty} \delta^{\tau} U_{t+\tau}^A, EU_t^j = U_t^A + \sum_{\tau=1}^{\infty} \delta^{\tau} U_{t+\tau}^j$.

Definition of attribution I say A attributes attacks to B_j if $(\beta_{jt} > \bar{\beta})$ in any t. This is a very tough definition because it means if ever A realizes that B is the one responsible behind any attack, attribution has happened. I get stronger results if I set the definition so that A has to be confident that B is responsible for a specific attack, or some amount of attacks.

B.2 Analysis roadmap

Our main goal is to prove that unattributable coercion is rationalizable. To do it, I solve for Pure Bayesian Equilibria (PBE) that meet my definition of unattributed coercion. I proceed in four parts. First, I identify some preliminary results that will simplify the equilibrium analysis. Second, I study the one Competitor model and solve for a pure strategy equilibrium. This simplest model illustrates the core mechanism, and explains why unattributable coercion is unrealistic with one Competitor. Second, I study a 2 Competitor model. I show that Schelling's paradox is now plausibly solvable. But the result also hints at coordination problems that the 1-Competitor model cannot consider. As I explain, these coordination problems are minor with only two Competitors (we can still solve for pure strategy equilibria). But the result suggests they grow severe if we add in many Competitors. Finally, I solve for arbitrarily large number of Competitors. I show that a volunteer's dilemma amongst Competitors prevents pure strategy equilibria from arising. However, I show that mixed strategy equilibria are reasonably easy to solve for.

Throughout the analysis, I make one assumption for simplicity,

$$\mathcal{A}_1 = \pi_L \delta < k_d < \pi_H \delta$$

This effectively means that the weakly resolved Challenger could not credibly promise to harm. While not strictly necessary to generate the result, it vastly simplifies the analysis because weakly resolved types strictly select no revision.

B.3 Preliminary analysis

As stated in the manuscript, unattributable coercion relies on a race between two components of the Challenger's beliefs. The first belief is the Challenger's perception that a specific Competitor is the source of harm. The second belief is the Challenger's expectation that harm will come to the Challenger if she seeks revision. Exactly where this belief falls depends on the underlying parameters of the model and equilibrium strategies. To ease discussion, we utilize the Markov Perfect equilibrium as our reversion point.

Lemma B.1 ***Reversion Equilibrium.*** *The following strategies are subgame perfect for all parameter values and types.* $s^A = r, r, r....,\ s^j = nh, nh, nh...nh.$ *for all* $j \in J$.

The proof is identical to the equivalent baseline Lemma.

This result vastly simplifies the analysis for two reasons. First, we can support these pure strategies for any parameters and any history. Second, even though attack choices are private, we always observe $a_t = r, x_t = nh$. It follows, that we can use this strategy as the trigger for reversion to Lemma B.1.

In what follows, we conjecture a set of strategies that meet our definition of coercion. Then we (a) see if we can support them in equilibrium; (b) make sure that we can support them given that A is not certain that a specific Competitor is attacking her ($\beta_{tj} < 1 \forall j$), and if we can, identify the smallest thresholds ($\hat{\beta}$) for which the equilibrium holds.

Define $T \geq 1$ as a critical period where the coercion sub-game starts.

Definition: Conjectured Pure Strategies: For periods $t < T$ if we ever observe $x_t = nh$, players revert to strategies described in Lemma B.1 in all subsequent periods. Else, $a_t = r \forall t < T$, $j_t = h | B_j = H \forall t < T$, $j_t = nh | B_j = L \forall t < T$.

For periods $t \geq T$, suppose players observe $a_t = r, x_t = nh$, and players revert to strategies described in Lemma B.1 in all subsequent periods. Otherwise, for periods $t \geq T$, $j_t(B_j = L) = nh \forall t \geq T$, $j_t | (B_j = H) = nh | a_t = nr, j_t = h | a_t = r$. $a_t = nr \forall t \geq T$.

Informally, starting in period T (possibly 1), coercion has happened. Before T, coercion has not happened. A is taking revision opportunities and suffering harm. If A stops taking revision opportunities before T, B still inflicts harm. But if B ever stops, we revert to the reversion subgame.

We can solve for the beliefs these strategies generate even before we solve for equilibria. Since many of our claims rest on these beliefs, it is useful to solve them first. We distinguish between two kinds of beliefs. We have already defined β_{tj} as A's belief that $B_j = H$ in period t. Let α_t be A's belief that at least one Competitor will harm her if she seeks revision at t.

A's prior beliefs are $\beta_{1j} = \psi$, and $\alpha_1 = 1 - (1-\psi)^J$. Then, if in any period $t - 1 < T$, A observes $x_{t-1} = nh$, A's posterior belief is $\beta_{t,j} = \alpha_t = 0$. However, so long as A has observed harm for every $t < t < T$, then, we can characterize A's beliefs at the beginning of period t as:

$$\alpha_t = \frac{1 - (1-\psi)^J}{1 - (1-\psi)^J + (1-\psi)^J \lambda^t}$$

Here $(1 - \beta_t)^J$ is the probability that no one attacks.

We can characterize A's beliefs that a specific j is highly resolved to attack for $t \leq T$ and given a history of harms in every prior period, as:

$$\beta_{t,j} | j_t = h = \frac{\psi}{\psi + (1-\psi)(1 - (1-\psi)^{J-1}) + (1 - \alpha_{t-1})\lambda^t}$$

$$\equiv \frac{\psi}{1 - (1-\psi)^J + (1 - \alpha_{t-1})\lambda^t}$$

We now proceed to equilibrium analysis, linking it to the specific results in the manuscript.

B.4 Establishing the Mechanism with one competitor (result 1, 2) and demonstrating its plausibility in the one-competitor case (result 3A).

B.4.1 Explaining how the informal results map onto the equilibrium analysis.

We now assume that $J = 1$ and detail an unattributable coercion equilibrium in Proposition B.2. The analysis of this case supports results 1, 2 and 3A. Before the technical analysis, I further discuss how the informal results relate to the technical results. Result 1 is an existence claim. Since its existence is the most interesting feature to a large group of informal readers who study coercion, the description below it in the manuscript details the causal mechanism. Proposition B.2 provide the clearest expression of the causal mechanism that establishes the existence claim in result 1 and supports the discussion below it. Result 2 is a statement about how long it takes to reach the coercion sub-game following one set of on-path actions.

Thus, result 2 describes a substantively interesting sub game within the equilibrium characterized by result 1. To be clear, results 1 and 2 also detail several scope conditions. With the exception of the claim about the number of Competitors, the scope conditions are consistent with re-arranging inequalities 3, 4 and are analyzed in more detail as comparative statics in section B.4.3. Later on we will study the multi-Competitor models. We will see that the technical conditions that support equilibrium differ somewhat from conditions 3, 4. However, the informal description of them does not. Specifically, it is still the case that the c must be high, k_d must be low, λ is not too high, and $\hat{\beta}$ must be high to support unattributable coercion.

Result 3a is a description of the attribution thresholds necessary to support Proposition B.2. This condition on its own terms is described as equilibrium condition 5. In the manuscript, I contrast that with the belief threshold necessary to support coercion (α_T). This contrast is effectively a comparative static. It is discussed in the third remark of section B.4.3.

B.4.2 Equilibrium analysis

Assume $J = 1$. This implies: $\beta_{1j} = \psi$, then if A has observed harm for all $t - 1 < T$

$$\beta_{tj} = \frac{\beta_{t-1,j}}{\beta_{t-1,j} + (1 - \beta_{t-1,j})\lambda^t}$$

Proposition B.2 *If*

$$k_d < \frac{(1-\lambda)\delta^T \pi_H}{1 - \delta + \delta(1 - \delta^{T-1})(1 - \lambda)} \tag{3}$$

$$1 > \frac{1 - \lambda\delta(1 - c(1 - \lambda))}{(1 - \lambda)[\delta + c(1 - \delta + \delta\lambda)} \tag{4}$$

and

$$\hat{\beta} > \beta_T \tag{5}$$

then the conjectured strategies are a PBE that fits my definition of unattributable coercion. We enter the coercion sub-game (period T arises) in the first period that $\beta_T = \alpha_T > \frac{1-\lambda\delta(1-c(1-\lambda))}{(1-\lambda)[\delta+c(1-\delta+\delta\lambda)}$

The conjectured equlibrium is stationary starting in period T (not yet identified). So we conjecture we can support it, and start our analysis in period T. We've shown we can support reversion to the reversion subgame if A deviates, and B does not punish her. For now, conjecture B could avoid attribution if A deviated to revision, and B played harm on the path. If true, then B is willing to punish A if $-k_d + \frac{\delta\pi_H}{1-\delta} > \frac{\delta\pi_H\lambda}{1-\delta}$. This is always true if 3 is true as desired.

Working backwards, I argue A prefers to play $a_T = nr$ over $a_T = r$. True if

$$-\frac{\lambda c}{1-\delta} > 1 - (\beta_T + \lambda - \beta_T\lambda)(c + \frac{\lambda\delta}{1-\delta}) + \delta\frac{1 - \beta_t - \lambda + \beta_T\lambda}{1-\delta}$$

This rearranges to

$$\beta_T > \frac{1 - \lambda\delta(1 - c(1 - \lambda))}{(1 - \lambda)[\delta + c(1 - \delta + \delta\lambda)} \tag{6}$$

We define T as the first period that satisfies this inequality. Condition 4 guarantees it can be satisfied given some history $x = h_1, h_2, ...h_{T-1}$.

We now consider $B_j = H$'s strategy in an arbitrary period $\tau < T$. If $x_\tau = nh$ players revert to Lemma B.1. Suppose A plays on the path at τ, then B cannot profit from deviating to $j_\tau = nh$ if:

$$\delta\lambda EU^j_{\tau+1} < -k_d + \delta EU^j_{\tau+1} \tag{7}$$

where $EU^j_{\tau+1}$ represents B_j's expected utility in the next period, given that A observes $x_\tau = h$.

Generally, we can write B's expected utility at τ for remaining on the path, given a history $x = h, h, h, , h...$ as

$$EU^j_t = \frac{\delta^{T-t+1}\pi - k_d(1-\delta^{T-t+1})}{1-\delta}$$

B's expected utility is clearly increasing in t. Thus, we can sub in $\tau = 1$ into inequality 7. Doing so, solves for the equilibrium condition 3 as desired.

The only thing left to check is that A's beliefs remain below the threshold. Notice that all Competitors pool their on-path behaviors for $t \geq T$. It follows, that if we arrive at the coercion sub-game, that A's beliefs about B's type are stable there.

The function that determines A's beliefs β_T is strictly increasing and bound by 1. It follows that we can always find a $\hat{\beta} > \beta_T$.

B.4.3 Discussion of parameter ranges

We now analyze the coercion and attribution thresholds necessary to support this result.

Remark As $c \to 0$ the RHS of condition 6 $\to \frac{\delta(1-\lambda)+1}{(1-\lambda)\delta} > 1$. As $c \to \infty$ the RHS of condition 6 $\to \frac{\delta\lambda}{(1-\delta)(1-\lambda)}$. This is less than 1 if $\lambda < \delta$ the risk of an accident is sufficiently small.

As $\lambda \to 1$ the RHS of condition 6 $\to \infty$. As $\lambda \to 0$ the RHS of condition 6 $\to \frac{1+\delta}{\delta+c(1-\delta)}$

As $\frac{1-\delta\lambda}{\delta(1-\lambda)} \to 0$ $T \to 1$ and the minimum $\hat{\beta} \to \psi$.

The reason that A's deterrence threshold depends on c is obvious. But it is valuable to note that B's utilities do not depend on c. The reason coercion hinges on λ is that A worries that she will suffer harm no matter what choice she makes. This makes clear exactly the attribute of an accident that matters. Specifically, to affect coercion, A must believe that she would not have otherwise suffered that harm. Thus, if you want to affect coercion, you cannot hide your attacks in a type of harm that A would incur anyway. Rather, you need to hide in a kind of harm that is rare.

Another factor in this result is that as this value increases towards 1, holding ψ constant, it increases the number of periods it takes to arrive at coercion. Thus, we want to know how this value varies to understand when we can affect coercion or not (omitting any attribution concerns). Fortunately, B's incentives are straight forward and this gives us some confidence that coercion is achievable.

Remark For any T, there is a k_d sufficiently low, that B is willing to launch attacks that will eventually generate coercion.

Putting these together, we think we can find attacks that can support the coercion aspect of the equilibrium. Specifically, these attacks must inflict massive harm, could have happened as the result of an unrelated strategic process (an accident), but that unrelated strategic process occurs rarely. We think many attacks fit this. For example, assassinations (in most countries), kidnapping and torture of elites. However, we do not believe that things like online election meddling will coercion because the Target often faces local actors that provide this sort of information. Therefore, they believe they will suffer this kind of harm either way.

So far, we have only considered whether we can support coercion. We have not yet considered whether B can avoid attribution.

Remark We can only support unattributed coercion if $\hat{\beta} > \beta_T > \frac{1+\delta\lambda c+\delta(1-\lambda)}{(1-\lambda)(\delta+c(1-\delta))}$. Informally, it must be possible that the audience is not confident enough that B caused an attack to impose a punishment on A, but A is confident enough that A is deterred from future revision.

We see this as a major limit on unattributable coercion with a single Competitor.

Putting these two parts together, we think unattributable coercion with a single Competitor is rare. The kinds of attacks that inflict severe harm are also the kinds of attacks that engender outrage quickly. Furthermore, the attacks must be rare enough to generate A's incentives for coercion, but no so rare that A instantly infers that B was responsible for them.

B.5 Two Competitors

We now introduce a second Competitor ($J = 2$). The equilibrium analysis supports result 3(b) and is used to derive panel (b) of Figure

In this case, $\beta_{1j} = \psi$, and $\alpha_1 = (2-\psi)\psi$. Further, $\frac{\psi}{1-(1-\psi)^J}$ simplifies to $\frac{1}{2-\psi}$.

$$\alpha_t = \frac{(2-\psi)\psi}{(2-\psi)\psi + (1-\psi)^2\lambda^t}$$

$$\beta_{t,j} = \frac{\psi}{(2-\psi)\psi + (1-\alpha_{t-1})\lambda^t}$$

Proposition B.3 *If condition 4 holds and*

$$k_d < \frac{(1-\lambda-\psi+\lambda\psi)\delta^T\pi_H}{1-\delta+\delta(1-\delta^{T-1})(1-\lambda-\psi+\lambda\psi)} \tag{8}$$

$$\hat{\beta} > \frac{1}{2-\psi} \tag{9}$$

then the conjectured strategies are a PBE that fits my definition of unattributable coercion. We enter the coercion sub-game (period T arises) in the first period that $\alpha_T > \frac{1-\lambda\delta(1-c(1-\lambda))}{(1-\lambda)[\delta+c(1-\delta+\delta\lambda)} > \beta_T$

The proof is very similar to the one Competitor model. There are but three differences that require additional analysis. First, $B_1 = H$'s decision to harm A must also factor in the possibility that B_2 is also highly resolved. If that was the case, then B_1 could shirk on his responsibility to punish. At $t = T$, B_1's prefers to punish A, then not if:

$$-k_d + \frac{\delta\pi}{1-\delta} > (\psi+\lambda-\lambda\psi)\frac{\delta\pi}{1-\delta}$$

Always satisfied if 8 is. In fact, 8 represents B's first-period incentive to harm A for T periods, factoring in both A's strategy, and the possibility that B_2 is highly resolved.[42] In this way, condition 8 represents what is necessary to support pure strategy coercion given that we do not observe whether/who any specific Competitor attacks. If it is violated, then B_1 is not willing to harm for certain because, in equilibrium, B_2 harms for certain.

The second change is that $\alpha_T > \beta_T$. The condition we use to define T is the same as the 1-Competitor model wrt α_T. But it allows for the possibility that β_T is much lower. Furthermore, the attribution threshold

[42] In this variant of the model, B_1 never learns if B_2 is also inflicting harm because B_1 always harms and only learns whether A accrues harm or not. We complicate this in an extension.

necessary to support attributable coercion is a fixed constant, depending only on B's priors (condition 9). This condition represents the limit of $\beta_{j,t} < 1$.

All other aspects of the proof are the same.

B.5.1 Substantive discussion: Contrasting 1 and 2 Competitors

Adding a second Competitor generates several interesting results. First, in the one Competitor model in periods $t \in 2, T$, $\alpha_t = \beta_t$. While we never achieved $\alpha_t = 1$ because A was coerced for lower levels of α_T, it was also the case that as $T \to \infty, \beta_T \to 1$. As we argued, this limited the supportable attribution thresholds to substantively undesirable areas. Adding in a second Competitor guaranteed that in periods $t \in 2, T$, $\alpha_t > \beta_t$. What is more, as $T \to \infty, \beta_T \to \frac{1}{2-\psi} < 1$. This is substantively appealing. After all, absent any history of harm, states usually believe that others are unlikely to sponsor secret and strictly punitive assassinations or coups, or otherwise launch punitive and secret military strikes (the prior belief ψ is low). It is only after they observe harm come to them that they begin to suspect a rival is involved (this is also captured in that the jump from the prior belief, to the posterior β_2 is the largest jump).

Second, adding a second Competitor introduces a coordination problem among the Competitors. This is represented in the ψ that appears in 8. This arises because one highly resolved Competitor's value for shirking to no-punishment, hinges on what he believes other Competitors will do. We solved for the condition where one Competitor is sufficiently confident that the other is weakly resolved. This suppresses the volunteer's dilemma. But an important concern, which we address next, is that a volunteer's dilemma will arise.

B.6 Many Competitors (Addressing the Volunteer's Dilemma)

Let's now consider a model with $J + 1$ defenders. Each Competitor is given a private level of resolve drawn i.id from ψ. For simplicity, we analyze the case where $\lambda = 0$, meaning an accident cannot happen. All other features of the model are the same.

When there are many Competitors the volunteer's dilemma takes hold. Thus, our goal is to identify a mixed strategy coercion equilibrium. To keep things simple, we focus on the case where coercion arises in the first period ($T = 1$).

The conjectured mixed strategies: On the path, A plays $a_t = nr$ for all t. $B^j = L$ plays $j_t = nh$ for all t. $B^j = H$ plays $j_t = nh | a_t = nr$, and $pr(j_t = h) = \gamma^* | a_t = nr$. If in period t' we observe $a_t = r, x_t = nh$, or $a_t = nr, x_t = h$ then players revert to $a_t = r, j_t = nh \forall t > t'1$.

In what follows, we characterize the γ^*, and $\hat{\beta}$ values necessary to support these strategies in PBEE. First, we characterize the pure strategy equilibrium where $\gamma^* = 1$. Second, we characterize a mixing equilibrium where $\gamma^* \in (0, 1)$. Finally, we show that when J is large, it is very easy to support these strategies in equilibrium and with very low attribution thresholds. Therefore, we conclude that unattributable coercion is very easy with many potential Competitors.

To get to this final result, we will solve for the pure strategy and mixing coercion equilibria. In both cases, all of the analysis hinges on players' expectation that A will suffer harm at two key moments. The first moment determines B's credible threat of retaliation. Consider the highly resolved Competitor's choice to harm A following the off-path play $a_t = r$. Let's say that all other highly resolved Competitors attack with probability γ. Then any Competitor's prior expectation expectation that A will suffer harm if he deviates to $j_t = nh$ is:

$$1 - \sum_{j=0}^{J} (1 - \psi)^{J-j} \psi^j (1 - \gamma)^j \binom{J}{j}$$

$$1 - (1 - \gamma\psi)^J$$

The second key expectation, is A's expectation that she will suffer harm if she deviates to revision, given that she has never suffered harm before. A's expectation of suffering harm is:

$$\alpha_1 = 1 - (1 - \gamma\psi)^{J+1}$$

This is almost identical it differs only in that A considers $J+1$ Competitors whereas any highly resolved Competitors need only consider J other Competitors. The difference between these two means that A's expectation that she will suffer harm is strictly larger than B's expectation harm will come to A if B deviates to no harm.

We are now ready to classify existence of pure strategy unattributed coercion equilibria.

Proposition B.4 *If*

$$\frac{k_d(1-\delta)}{\delta\pi} < (1-\psi)^J \tag{10}$$

and

$$(1-\psi)^{J+1} < \frac{(1-\delta)(C-1)}{\delta + c(1-\delta)} \tag{11}$$

and

$$\hat{\beta} > \frac{\psi}{1-(1-\psi)^{J+1}} \tag{12}$$

then the conjectured strategies form a pure strategy equilibrium with $\gamma^* = 1$.

Since the game is stationary, we need only check one shot deviations.

Starting with B's incentives to play $\gamma^* = 1$. Conjecture γ^* holds for all Competitors. If any highly resolved Competitor observes $a_t = r$, they prefer to punish this with certainty if they avoid attribution and:

$$-k_d + \frac{\delta\pi}{(1-\delta)} > \frac{\delta\pi}{(1-\delta)}(1-(1-\gamma\psi)^J)$$

Plugging in $\gamma^* = 1$ gives us the first condition.

Now consider A's incentive to never deviate away from $a_t = nr$.

A never attacks if:

$$0 > 1 - c(1-(1-\gamma\psi)^{J+1}) + \frac{\delta(1-\gamma\psi)^{J+1}}{1-\delta}$$

$$\delta(1-\gamma\psi)^{J+1} < \frac{(1-\delta)(C-1)}{\delta + c(1-\delta)}$$

This gives us the second equilibrium condition as desired.

The final thing to check is that B can avoid attribution in the off-path case that A sets $a_1 = r$ and B_j punishes A with $j_1 = h$.

The belief that a specific B is resolved after a first-period punishment is

$$\beta_2 | a_t = r, x_t = h = \frac{\psi}{(1-\psi)(1-(1-\gamma\psi)^J) + \psi} \tag{13}$$

With $\gamma^* = 1$ this gives us the final equilibrium condition.

This completes the proof.

We now use these results to construct the mixing equilibrium.

Proposition B.5 *If*

$$(1-\psi)^J < \frac{k_d(1-\delta)}{\delta\pi} < 1 \tag{14}$$

and

$$(\frac{k_d(1-\delta)}{\delta\pi})^{\frac{J+1}{J}} < \frac{(1-\delta)(c-1)}{\delta + c(1-\delta)} \tag{15}$$

and

$$\hat{\beta} > \frac{\psi - (1 - \sqrt[J]{\frac{k_d(1-\delta)}{\delta\pi}})(\frac{k_d(1-\delta)}{\delta\pi})^{\frac{J+1}{J}}}{1 - (\frac{k_d(1-\delta)}{\delta\pi})^{\frac{J+1}{J}}} \tag{16}$$

then the conjectured strategies form a mixed strategy with $\gamma^* \in (0,1)$.

First, we solve for the γ^* that leaves any highly resolved Competitor indifferent between punishment and not, given their prior beliefs about who else is highly resolved (and assuming that they avoid attribution):

$$\frac{k_d(1-\delta)}{\delta\pi} = (1-\gamma\psi)^J$$

$$\gamma^* = \frac{1 - \sqrt[J]{\frac{k_d(1-\delta)}{\delta\pi}}}{\psi} \tag{17}$$

The first equilibrium condition tells us when the value of γ^* in equation 17 lies between 0 and 1. By construction we can support all $B_j = H$ playing it in reply to A's off path deviation to $a_t = r$.

The rest of the proof follows the same pattern as the pure strategy equilibria using γ^* in 17 rather than $\gamma^* = 1$. The second condition tells us when A is deterred by by the highly resolved Competitors threat to punish with probability γ^*.

The final condition tells us the $\hat{\beta}$ that allows B to avoid attribution.

Proposition B.6 *If*

$$\frac{k_d(1-\delta)}{\delta\pi} < \frac{(1-\delta)(c-1)}{\delta + c(1-\delta)} \tag{18}$$

and

$$\hat{\beta} > \psi \tag{19}$$

then there exists a J sufficiently large where the conjectured strategies form an equilibrium for $\gamma^* \in (0,1)$. Thus, under these conditions, there is always an unattributable coercion equilibrium in which A is deterred in the first period.

Remark This equilibrium is unattributable if the attribution threshold barely exceeds A's prior belief that any type is greedy.

The results follow from taking $J \to \infty$ for the three conditions in proposition B.5. The remark is important. It implies that even if A was to attack, she would learn basically nothing beyond her prior about which Competitor attacked her. The reason is that there are so many to choose from.

C Extension: Asymmetric information conditional on history

The basic model limits what the Competitor knows. In real life, Competitors often know more. Specifically, because each Competitor knows what harm they inflicted, they also know whether the Challenger suffers harm that they did not cause. To account for this complication, we now introduce a state variable $z_t^j \in h, nh$ that Nature reveals privately to j in period t. $z_t^j = h$ if A suffers harm that Competitor j did not cause, and nh otherwise.

We adjust the timing of each period as follows. First, A selects revision or not. Second, The Competitors and Nature (with probability λ) simultaneously decide to harm A or not. Then nature reveals x_t, z_t^j.

In terms of information, all players observe A's action, their own and x_t. Then each Competitor privately observes z_t^j. We can see some things instantly. First, if Competitor B^1 sets $b_t^1 = h$ they know that $z_t^2 = h$. But if Competitor B^1 sets $b_t^1 = nh$, then $pr(z_t^2 = h) = \lambda$.

C.0.1 Analysis

This sort of set-up is often intractable because each player's actions hinge on beliefs about other player's beliefs. Fortunately, our model has some nice features that allow for a simpler result. The basic insight is that all actors observe whether $x_t = nh$. Therefore, irrespective of beliefs, we can always support reversion to the reversion equilibrium described in Lemma B.1 given $a_t = r, x_t = nh$.

We are going to solve for an equilibrium in which we arrive at coercion in the second period, and Competitors punish probabilistic in the first period (and off path in second period). The critical factor are the Competitor's mixing probabilities for attack in the first period, and subsequent periods off the path. Define, γ as $pr(b_1 = h|a_1 = r, B^j = H)$. Because of symmetries in both Competitor's payoffs and beliefs in the first period, this value will wind up being the same for all players. Define γ_{zb} as $pr(b_t = h|a_t = r)$ for $t > 1$. The subscript $z \in 0, 1$ represents whether B^j observed harm that he did not cause. Subscript b represents whether B^j inflicted harm in the first period. The super-script indicates that this value will vary for each player because the history will vary. Thus, γ_{10}^1 means that B^1 did not inflict harm in the first period, but observed harm inflicted on A that B^1 did not cause. Therefore, if in any future period A seeks revision, B attacks A with probability γ_{10}^1.

We conjecture the following strategies. A's on path strategy is $a_1 = r$. Then $a_t = nr|x_1 = h$ for $t > 1$, and $a_t = r|x_1 = nh$. B's on-path strategy is as follows. If $B^j = L$, then $j_t = nh$ for all t. If $B^j = H$, then in the first period, $pr(j_1 = h) = \gamma^*$. For all $t > 1$, $j_t = nh|a_t = nr$. Off path in any period $\tau \geq 2$ if $a_\tau = r$, $pr(j_\tau = h) = \gamma_{zb}^j$ where $\gamma_{00}^* = \gamma_{10}^* = 0 < \gamma_{11}^* \leq \gamma_{01}^* = 1$. Off path, if in the first period if $a_1 = nr, x_1 = nh$ or in any subsequent period if $a_t = r, x_t = nh$ then players revert to the strategies described in Lemma B.1.

Proposition C.1 *If any of the three following conditions are met, then we can find a c for which we can support the conjectured strategies as a PBE with the following γ^*, γ_{11}^* values.*

First, if

$$\frac{\delta\pi(1-\lambda)}{1-\delta}(1 - \frac{\psi^2[\psi(1-\lambda^2)+\lambda]}{[\psi(1-\lambda)+\lambda]^3}) \geq k_d$$

holds, then $\gamma^* = \gamma_{11}^* = 1$.

Second, if

$$\frac{\delta\pi(1-\psi}{1-\delta} \geq k_d \geq \frac{\delta\pi(1-\lambda)}{1-\delta}(1-\frac{\psi^2[\psi(1-\lambda^2)+\lambda]}{[\psi(1-\lambda)+\lambda]^3})$$

and

$$4\psi[(1-\psi)(1-\lambda)+\frac{(\delta\pi(1-\lambda)-k_d(1-\delta))[\psi-\psi\lambda+\lambda]^3}{\delta\pi(1-\lambda)\psi^2[2\psi(1-\lambda)+\lambda]}]+(1-\psi)^2>1$$

hold, then $\gamma^* = 1$ *and*

$$\gamma^*_{11} = \frac{(\delta\pi(1-\lambda)-k_d(1-\delta))[\psi-\psi\lambda+\lambda]^3}{\delta\pi(1-\lambda)\psi^2[2\psi(1-\lambda)+\lambda]} < 1$$

Third, if $k_d > \frac{\delta\pi(1-\psi}{1-\delta}$ *and some other very intractable condition that I define later on, then*

$$\gamma^* = \frac{\delta\pi - k_d(1-\delta)}{\psi\delta\pi}$$

$$\gamma^*_{11} = \frac{(\delta\pi(1-\lambda)-k_d(1-\delta))[\psi\gamma^*(1-\lambda)+\lambda]^3}{\delta\pi\gamma^*(1-\lambda)[\psi\gamma^*(1-\lambda)+\lambda\psi]^2[(\psi\gamma^*(1-\lambda)+\lambda\psi)\gamma^*(1-\lambda)+\lambda(\psi\gamma^*(1-\lambda)+\lambda)]}$$

There are three basic insights that drive the proof. First, each type of B's mixing probabilities over punishments are chosen based on beliefs about what the other Competitor will do and not what A will do. Thus, we first solve for B's on path mixing probabilities and show we can support them.

Second, because there is the risk of an accident, and on the path no Competitor harms A after the first period, then A and B's on-path beliefs are static starting in the second period. By the one-shot deviation principle, we need only consider incentives to deviate in the second period.

Third, each period A's preference for revision or not can be summarized as a comparison based on A's expectation that A will suffer harm if A chooses to seek revision. We can always choose a value of c that supports or violates this condition, but B's choices do not depend on c at all. It follows, that so long as A's expectation for suffering harm increase from the first to the second period if A observes harm, then we can find a c for which A wants to seek revision in the first period, but does not want to seek revision in the second if A observes harm in the first.

We start by solving for B's second period mixing probabilities given A's off-path deviation to $a_2 = r$. Define α_{zb} as B^1's belief that A will suffer harm if B does not act given the beliefs that follow from on path strategies. Then, B^1 strictly prefers to attack if:

$$\frac{\delta\pi - k_d(1-\delta)}{1-\delta} > \frac{\delta\pi\alpha_{zb}}{1-\delta}$$

$$1 - \frac{k_d(1-\delta)}{\delta\pi} > \alpha_{zb} \tag{20}$$

For B's equilibrium mixing probabilities to hold together, we must be able to solve for:

$$\alpha_{01} < \alpha_{11} = \min(\frac{\delta\pi - k_d(1-\delta)}{\delta\pi}, 1) \leq \alpha_{10}$$

These α expectations rely on three kinds of posterior beliefs. First, they rely on B^1's expectation that B^2 is highly resolved given what B^1 observed in the first period. There are two outcomes to consider.

If $z^1_1 = h$, then B^1's second period belief that B^2 is resolved is:

$$\psi_{21} = \frac{\psi\gamma + (1-\gamma)\lambda\psi}{\psi\gamma + (1-\gamma)\lambda\psi + (1-\psi)\lambda} = \frac{\psi\gamma(1-\lambda) + \lambda\psi}{\psi\gamma(1-\lambda) + \lambda}$$

If instead $z_1^1 = nh$, then B^1's second period belief that B^2 is resolved is:

$$\psi_{20} = \frac{\psi(1-\gamma)(1-\lambda)}{\psi(1-\gamma)(1-\lambda) + (1-\psi)(1-\lambda)} = \frac{\psi(1-\gamma)(1-\lambda)}{(1-\gamma)(1-\psi\lambda)}$$

Second, we must solve for B^1's belief that B^2 observed harm. If $b_1^1 = h$ then this is equal to 1. Otherwise, it is the probability of an accident λ.

Finally, we must compute B^1's equilibrium belief that B^2 inflicted harm on B^1 given z_1^1. If $z_1^1 = nh$, then this is 0. But if $z_1^1 = h$, then

$$\begin{aligned} \omega := pr(b_1^2 = h | z_1^1 = h) &= \frac{\psi_{21}\gamma}{\lambda\psi_{21}\gamma + \lambda(1-\psi_{21}) + \psi_{21}\gamma(1-\lambda) + \lambda\psi_{21}(1-\gamma)} \\ &= \frac{\psi_{21}\gamma}{\psi_{21}\gamma(1-\lambda) + \lambda} \end{aligned}$$

Thus, given the conjectured equilibrium strategies,

$$\alpha_{01} = (\psi_{20}\gamma_{10})(1-\lambda) + \lambda = \lambda$$

$$\alpha_{10} = \psi_{21}(\lambda\omega\gamma_{11} + (1-\lambda)\gamma_{01}\omega + \lambda(1-\omega)\gamma_{10})(1-\lambda) + \lambda = (1-\lambda)\psi_{21}\omega(1-\lambda+\gamma_{11}) + \lambda$$

$$\alpha_{11} = \psi_{21}(\omega\gamma_{11} + (1-\omega)\gamma_{10})(1-\lambda) + \lambda = \psi_{21}\omega\gamma_{11}(1-\lambda) + \lambda$$

Notice, $\alpha_{10} > \alpha_{11} > \alpha_{01}$, as desired.

Plugging ω and ψ_{21} into α_{11} and then α_{11} into 20 means that B^1, then we can solve for the equilibrium value of γ_{11}^* as:

$$\gamma_{11}^* = \min(\frac{(\delta\pi(1-\lambda) - k_d(1-\delta))[\psi\gamma(1-\lambda)+\lambda]^3}{\delta\pi\gamma(1-\lambda)[\psi\gamma(1-\lambda)+\lambda\psi]^2[(\psi\gamma(1-\lambda)+\lambda\psi)\gamma(1-\lambda)+\lambda(\psi\gamma(1-\lambda)+\lambda)]}, 1)$$

In the first period, A seeks revision on the path. Then highly resolved B is indifferent between attacking and not if:

$$-k_d + \frac{\delta\pi}{1-\delta} = \frac{\psi\gamma\delta\pi}{1-\delta}$$

so

$$\gamma^* = \min(\frac{\delta\pi - k_d(1-\delta)}{\psi\delta\pi}, 1)$$

If we ignore the probability boundary at 1, then it is clear that $\gamma^* > \gamma_{11}^*$. However, it is also easy to show that both can be larger than 1 so long as $\psi < 1, \lambda \in (0,1)$. That gives us three conditions to check.

Condition 1. Subbing in $\gamma_1^* = 1$,

$$\frac{\delta\pi(1-\lambda)}{1-\delta}(1-\frac{\psi^2[\psi(1-\lambda^2)+\lambda]}{[\psi(1-\lambda)+\lambda]^3}) \geq k_d$$

Then $\gamma^* = \gamma^*_{11} = 1$.

Condition 2. Subbing in $\gamma^*_1 = 1$,

$$\frac{\delta\pi(1-\psi}{1-\delta} \geq k_d \geq \frac{\delta\pi(1-\lambda)}{1-\delta}(1-\frac{\psi^2[\psi(1-\lambda^2)+\lambda]}{[\psi(1-\lambda)+\lambda]^3})$$

Then $\gamma^* = 1$ and

$$\gamma^*_{11} = \frac{(\delta\pi(1-\lambda)-k_d(1-\delta))[\psi-\psi\lambda+\lambda]^3}{\delta\pi(1-\lambda)\psi^2[2\psi(1-\lambda)+\lambda]} < 1$$

Condition 3.

$$k_d > \frac{\delta\pi(1-\psi}{1-\delta}$$

Then $1 > \gamma^* > \gamma^*_{11}$ where the min values above give the equilibrium mixing values.

We now turn to A's incentives and beliefs. On the path, A revises in period 1. If A observes harm, A selections $a_t = nh$ for $t > 1$. Let α_{At} represent A's belief that harm will come to her if she selects revision in period t. Then, A selects revision if: $\frac{-c\lambda}{1-\delta} > \frac{1-\alpha}{1-\delta} - \alpha c$.

Given A's on path strategy, if $x_1 = h$, then $\alpha_{A2} = \alpha_{A3}$.... Thus, we focus on a contrast between A's first period belief, and A's second period belief conditional on $a_1 = r, x_1 = h$.

For the equilibrium to hold together, it must be that:

$$\alpha_{A2}|x_1 = h > \frac{1-c\lambda}{1+c(1-\delta)} > \alpha_{A1} \tag{21}$$

Notice that c is independent of B's strategy. Clearly, if $\alpha_{A2}|x_1 = h > \alpha_{A1}$ then we can find a c that satisfies this inequality, and therefore can support A's on path strategy.

We start by solving for α_{A2}. A's belief that harm will come depends on A's posterior belief that no (ρ_0), one (ρ_1) or 2 (ρ_2 Competitors are highly resolved given that $x_1 = 1$. Critically, A's expectations are different from B^j because A only observes x_t.

$$\rho_0 = \frac{(1-\psi)^2\lambda}{1-(1-\lambda)[(1-\gamma)\psi(2-\psi\gamma-\psi)]}$$

$$\rho_1 = \frac{2\psi(1-\psi)(1-\gamma(1-\lambda))}{1-(1-\lambda)[(1-\gamma)\psi(2-\psi\gamma-\psi)]}$$

$$\rho_2 = \frac{\psi^2(\lambda+(1-\lambda)\gamma(2-\gamma))}{1-(1-\lambda)[(1-\gamma)\psi(2-\psi\gamma-\psi)]}$$

A knows the equilibrium mixing probabilities $\gamma_{01} = 0 < \gamma_{11} \leq \gamma \leq 1 = \gamma_{10}$. A can use γ, ρ, to compute the probability that each Competitor will deploy a specific second period mixing probabilities. That gives us:

$$\alpha_{A2} = \lambda + (1-\lambda)[\rho_0 \times 0 + 2\rho_1(\gamma(1-\lambda)+\gamma\lambda\gamma_{11}) + 2\rho_2(\gamma^2\gamma_{11} + \gamma(1-\gamma)(1-\lambda) + \gamma(1-\gamma)\lambda\gamma_{11})]$$

$$\alpha_{A2} = \lambda + 2\gamma(1-\lambda)[\rho_1(1-\lambda(1-\gamma_{11})) + \rho_2(\gamma\gamma_{11} + (1-\gamma)(1-\lambda(1-\gamma_{11})))]$$

In the first period, A's expectation of harm is based on the priors:

$$\alpha_{A0} = 1 - (1-\lambda)(\psi^2(1-\gamma)^2 + 2\psi(1-\psi)(1-\gamma) + (1-\psi)^2)$$

Thus, $\alpha_{A2} > \alpha_{A1}$ if:

$$2\gamma[\rho_1(1-\lambda(1-\gamma_{11}))+\rho_2(\gamma\gamma_{11}+(1-\gamma)(1-\lambda(1-\gamma_{11})))]+\psi^2(1-\gamma)^2+2\psi(1-\psi)(1-\gamma)+(1-\psi)^2 > 1$$

This is always true if $\lambda^*_{11} = 1$. It follows, that we can always support this equilibrium under condition 1.

Given condition 2, where $\lambda^* = 1 > \lambda^*_{11}$. Then, the condition solves for:

$$2[2\psi(1-\psi)(1-\lambda(1-\gamma_{11})) + \psi^2\gamma_{11}] + (1-\psi)^2 > 1$$

$$4\psi[(1-\psi)(1-\lambda) + \gamma_{11}] + (1-\psi)^2 > 1$$

Subbing in

$$\gamma_{11} = \frac{(\delta\pi(1-\lambda) - k_d(1-\delta))[\psi - \psi\lambda + \lambda]^3}{\delta\pi(1-\lambda)\psi^2[2\psi(1-\lambda) + \lambda]}$$

$$4\psi[(1-\psi)(1-\lambda) + \frac{(\delta\pi(1-\lambda) - k_d(1-\delta))[\psi - \psi\lambda + \lambda]^3}{\delta\pi(1-\lambda)\psi^2[2\psi(1-\lambda) + \lambda]}] + (1-\psi)^2 > 1$$

D Extension: Distinctive issues.

Consider the following example. Assume $J = 2$, then make the following adjustment to the set-up. At the beginning of every period nature draws a random variable z_t, such that $pr(z_t = 1) = \omega$, $pr(z_t = 0) = 1 - \omega$. The revelation z_t is public and ω is known. If $z_t = 0$, then there is no change to the payoffs of the two-player model. If $z_t = 1$, then B_1's value for revision is π_L irrespective of B_1's type.

We conjecture the strategies that form the PBE described in proposition B.3. That means that both Competitors still play identical strategies. We have effectively proven that A and B^2's strategies hold, all we need to show is that $B^1 = H$ could not profitably deviate given that $B^1 = H$'s preferences are now different.

If we replace condition 8 with

$$(1-\psi-(1-\psi)\lambda)\delta^T\frac{(1-\omega)\pi_H + \omega\pi_L}{1-\delta^T} > k_d \tag{22}$$

The conditions described in proposition B.3 sustain an equilibrium in the extension wherein B^1 does not value all issues high.

Since the proof is very similar, we only highlight the differences. Starting with the coercion phase, we conjectured all Competitors play $b_t = nh|a_t = nr$ and Challengers play $a_t = nr$ on path. However, if ever the Challenger was to deviate to $a_t = r$ highly resolved Competitors play $b_t = h|a_t = r$. Finally, if ever we observe $a_t = r, x_t = nh$, we revert to the total revision subgame and this does not depend on A's beliefs about B's type.

In the coercion sub-game Competitor's get their maximum possible total expected utility from on-path play. Thus, neither Competitor can profitably deviate if A plays on-path.

However, the equilibrium requires that if A deviates to $a_t = r$, that all $B^j = H$ punish A's decision with $j_t = h$. It must be the case that no B can profitably deviate to $j_t = nh|a_t = r$. Proposition B.3 defines the conditions wherein B^2 cannot profitably deviate in this off-path sub-game. But B^1's incentives could be different in the extension because $B^1 = H$'s expected utility from on path play in the coercion phase is strictly less than $B^2 = H$'s. In any $t > T$, $EU_{t,1} = z_t\pi_L + (1-z_t)\pi_H + \delta\frac{(1-\omega)\pi_H+\omega\pi_L}{1-\delta} < \pi_H + \frac{\delta\pi_H}{1-\delta} = EU_{t,2}$. Thus, B^1 has a stronger incentive to deviate to no punishment.

We now consider B^1's incentive to deviate from this punishment for any set of off-path beliefs. Since players enter the reversion sub-game for any set of beliefs, the beliefs only influence whether B suffers the additional cost of attribution. First, assume the case, that $\beta_{t+1,1}|(x_t = 0, a_t = r) < \hat{\beta}$. Then, B^1 punishes A's deviation if:

$$-k_d + \delta\frac{(1-\omega)\pi_H + \omega\pi_L}{1-\delta} > (\psi + (1-\psi)\lambda)\delta\frac{(1-\omega)\pi_H + \omega\pi_L}{1-\delta}.$$

$$k_d < (1-\psi-(1-\psi)\lambda)\delta\frac{(1-\omega)\pi_H + \omega\pi_L}{1-\delta} \tag{23}$$

Notice this is very similar to the equilibrium condition we solved for in the baseline model. It differs only in that we replace π_H with $\delta\frac{(1-\omega)\pi_H+\omega\pi_L}{1-\delta}$. Like in the baseline model, it is solvable for k_d sufficiently small.

Notice that B^1's decision to punish A's deviation at t does not depend on the draw z_t, but only the expectation of future value. This is because if A decides to deviate, B^1 choice to punish ex-post does not influence A's prior policy choice (a_t).

Second, assume the case that $\beta_{t+1,1}|(x_t = 0, a_t = r) > \hat{\beta}$. Then, B^1 punishes A's deviation if:

$$-k_d + \delta\frac{(1-\omega)\pi_H + \omega\pi_L}{1-\delta} > (\psi + (1-\psi)\lambda)\delta\frac{(1-\omega)\pi_H + \omega\pi_L}{1-\delta} + (1-\psi-(1-\psi)\lambda)\frac{-k_i}{1-\delta}$$

$$k_d < (1-\psi-(1-\psi)\lambda)\delta\frac{(1-\omega)\pi_H + \omega\pi_L + k_i}{1-\delta} \tag{24}$$

Contrasting, 23, 24, the latter is easier to satisfy by $\frac{k_i}{1-\delta}$. This illuminates how the fear of attribution actually keeps B^1 on the path even though B^1's benefit from following through with punishment is smaller than B^2's.

All other components of the proof for the coercion sub-game are identical.

We now turn to the periods $t < T$. The proof structure for B^1 is identical, so we proceed to B^1's first-period expected utility for remaining on the path.

$B^1 = H$'s first period expected utility from remaining on path is:

$$EU_1^1 = \delta^T\frac{(1-\omega)\pi_H + \omega\pi_L}{1-\delta} - \frac{k_d(1-\delta^{T-t+1})}{1-\delta}$$

Assume the case, that $\beta_{2,1}|(x_1 = nh) < \hat{\beta}$. Then, B^1's value for deviation is:

$$(\psi + (1-\psi)\lambda)\delta^T\frac{(1-\omega)\pi_H + \omega\pi_L}{1-\delta}.$$

B^1 remain on the path if:

$$(1 - \psi - (1 - \psi)\lambda)\delta^T \frac{(1 - \omega)\pi_H + \omega\pi_L}{1 - \delta^T} > k_d \tag{25}$$

as stated.