Does plausible deniability work? Assessing the effectiveness of unclaimed coercive acts in the Ukraine war

(forthcoming in *Contemporary Security Policy*)

Costantino Pischedda
Department of Political Science, University of Miami, USA

Andrew Cheon
School of Advanced International Studies, Johns Hopkins University, Washington, DC, USA

Abstract

States conduct unclaimed coercive acts, imposing costs on adversaries to signal resolve but denying (or not claiming) responsibility. Some scholars posit that unclaimed acts have considerable potential to coerce targets, while containing escalation risks. Others suggest that unclaimed coercive efforts tend to fail and trigger escalation. We assess these competing perspectives about the effects of unclaimed attacks with a vignette experiment exposing US-based respondents to a scenario where, after Russia warns of unpredictable consequences if NATO continues providing weapons to Ukraine, an explosion occurs at a NATO base in Poland used to funnel weapons to Ukraine. Intelligence agencies and independent analysts identify Russia as the likely culprit, while not ruling out the possibility of an accident. We randomize whether Russia claimed or denied responsibility for the explosion and find that unclaimed acts have lower coercive leverage than claimed ones, but the two do not significantly differ in escalation risk.

Key words: plausible deniability; coercion; war in Ukraine; provocation; covert action; unclaimed attacks

Acknowledgements: The authors thank Brian Blankenship for helpful comments on an early draft of this article.

Replication materials available at:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KQPWRN

The research was approved by the Johns Hopkins University's Homewood Institutional Review Board on December 24, 2021 (protocol number: AM00014523). The experiment was preregistered with EGAP on May 17, 2022 (registration ID: 20220517AB).

Financial information: The authors gratefully acknowledge financial support from the University of Miami.

Disclosure statement: No potential conflict of interest has been reported by the authors.

Instances abound of states carrying out unclaimed coercive acts, that is, imposing costs on adversaries to signal resolve to prevail in a dispute while denying involvement or simply not making any claims about responsibility.[1] For example, Russia is believed to have launched a wave of cyber-attacks in 2007 to extract a series of concessions from Estonia, even though Moscow denied any connection to the events (Valeriano et al., 2018, pp. 124-127). The unclaimed 1988 bombing of Pan Am 103 over Lockerbie is thought to have been carried out by Muammar Gaddafi's regime in response to an earlier coercive bombing raid by the United States (Hoffman, 1997, p. 4). In 2010, the South Korean government claimed a North Korean torpedo was responsible for the sinking of a South Korean warship, though Pyongyang denied involvement (Hur 2017). Observers have suggested that the sinking may be part of a broader coercive campaign waged by North Korea, taking both covert and overt forms.

Unclaimed coercive acts are not limited to state covert action, but include activities carried out by nonstate actors.[2] For example, the Tamil Tigers resorted to both guerrilla and terrorist operations to compel the Sri Lankan government to concede Tamil self-determination but did not claim their attacks against civilians (Pluchinsky, 1997, p. 4). In fact, most acts of terrorism, often seen as a quintessential instrument of coercion, go unclaimed (Lieber & Press 2013; Bauer et al., 2017; Crenshaw & LaFree, 2017, pp. 131-164; Kearns, 2021).

Not all (or even most) unclaimed actions by states and nonstate actors have coercive purposes. Some aim at information collection (e.g., espionage), on-the-ground effects (e.g., sabotage of an opponent's military assets and election interference),[3] or strategic surprise—objectives that may be more likely to be achieved if the action itself (as distinct from the identity of the actor) remains secret.[4] Nonetheless, as the preceding examples illustrate, an empirically meaningful subset of unclaimed actions does appear to be coercive in nature.[5] This article seeks

to shed light on the effects of unclaimed coercive acts by deriving from the literature a series of hypotheses and then testing them with a vignette-based survey experiment.

We envision a coercer's denial of responsibility (or failure to make any claims about responsibility) as giving rise to a situation of plausible deniability, where various audiences may experience different degrees of uncertainty about who is behind the unclaimed act, depending on their prior beliefs, access to information about the event, and emotional reactions to it. Two key audiences in a situation of plausible deniability are the government and the population of the country being coerced (i.e., the target country). We do not expect a systematic divergence in the reactions of these two audiences, as long as they have access to comparable information.[6] Thus, throughout the article we refer to "targets" when a point is relevant to both the government and the population of the target country.

It should be noted that sometimes the government and the population will have fundamentally different information at their disposal. In some cases, the government may have access to intelligence eliminating virtually all uncertainty about the identity of the culprit (e.g., radar data tracing a missile to its launch site), but this information may be withheld from the public, which might therefore remain highly uncertain about who is behind an unclaimed act. In other cases, the public may even be unaware of the occurrence of an unclaimed act that the government can attribute with high confidence due to the latter's privileged access to intelligence, though the spread of technologies (e.g., commercial imagery satellites) eroding government intelligence advantages is likely to reduce the frequency of this type of cases (Lin-Greenberg & Milonopoulos, 2021). In situations where the government and the population of the target country diverge in their reactions to unclaimed coercive acts (due to asymmetric

information or other factors), the former will ultimately decide policy responses, but it may be constrained by the public's preferences (particularly in democratic settings).

A body of literature suggests that plausible deniability offers substantial coercive benefits: It may enable coercers to send intelligible, credible, and thus potentially effective messages to targets while containing escalation risks (Hoffman 1997, p. 5; Libicki, 2009, pp. 41-42; Byman & Kreps, 2010, p. 6; Carson & Yarhi-Milo, 2017; Carson, 2018), besides reducing coercers' costs of being seen as a norm violator (Downes & Lilley, 2010; Carey et al., 2015; O'Rourke, 2018; Poznansky, 2019).[7]

Other theoretical perspectives, however, imply that plausible deniability may backfire, with unclaimed actions failing to coerce their targets and leading instead to escalation. Coercers' efforts to conceal their hand may result in muddled communication, leaving targets unsure as to what is being demanded of them and whether making concessions will bring an end to the hostile actions, thus reducing the probability of compliance (Abrahms, 2008, pp. 89-90; Gartzke, 2013, p. 47; Borghard & Lonergan, 2017, pp. 453-459). Even if targets fully grasp the coercive message, they may infer from the unclaimed nature of the act that the coercer lacks resolve or that it is deceitful, which would reduce the probability of compliance. Furthermore, despite the fact that the absence of a claim of responsibility may make detached observers unsure about the identity of the perpetrator and thus about how to respond, anger may prompt targets to retaliate against the most plausible culprit (Lerner & Tiedens, 2006). Concerns about reputation and honor also may dispose targets to retaliate, rather than concede (Dafoe et al., 2021). Thus, unclaimed acts may not shield coercers from escalation risks.

Over the past years, unclaimed coercive acts have attracted a growing amount of policy and academic attention in the context of alleged Iranian attacks against oil tankers and Saudi oil

facilities (Carson, 2019; Cordesman, 2019), various high-profile cyber operations (Lindsay, 2015; Sharp, 2017; Valeriano et al., 2018), and Russia's intervention in the Donbas (Cormac & Aldrich, 2018; Bowen, 2019). The increasing interest in the topic notwithstanding, there has been little systematic empirical scrutiny of the effects of unclaimed coercion. An article by Carson and Yarhi-Milo (2017) examining the intelligibility and credibility of coercive covert action with case studies is an important exception. However, in the absence of counterfactuals where coercion is wielded overtly, Carson and Yarhi-Milo's analysis cannot assess the relative effectiveness of plausible deniability. As a result, we do not know whether claimed and unclaimed attacks differ in terms of their ability to coerce targets and prevent escalation.

This article helps fill this gap with a vignette experiment exposing US-based respondents to a fictional scenario depicting an explosion at a NATO base in Poland used to funnel heavy weapons to the Ukrainian armed forces during the ongoing war against Russia. All respondents were told that Russia's President Vladimir Putin had previously warned of "unpredictable consequences" if NATO continued providing heavy weapons to Ukraine and that both intelligence agencies and independent analysts identified Russia as the likely culprit without, however, ruling out the possibility of an accidental detonation. By randomizing whether Russia claimed or denied responsibility for the explosion, we can assess the effects of plausible deniability on targets' preferences for complying with Putin's demands and for taking escalatory actions in response.

In light of the accumulating evidence suggesting that foreign policy elites' reactions to real-world events and to experimental manipulation do not differ much from those of the general public (Mercer, 2013; Hall, 2017; Yarhi-Milo, 2018; Yarhi-Milo et al., 2018; Kertzer et al., 2021; Kertzer, 2022), we believe that our empirical approach offers indirect insight into how the

US government would respond in a similar scenario, in addition to more direct insight into the US public's response. Moreover, using a general population sample would be useful even if one expected, for some reason or another, divergent reactions for the US government and the US public, given that the latter's preferences may act as constraints on the leaders ultimately making policy decisions.

To preview our results, subjects are less likely to be in favor of interrupting the flow of weapons to Ukraine (i.e., to comply with Russian demands) when the attack goes unclaimed compared to when Russia claims it. On the other hand, the absence of a Russian claim of responsibility does not have a significant effect on respondents' preferences for escalatory responses. Thus, plausible deniability appears to reduce the coercer's leverage without countervailing benefits in terms escalation risk mitigation. Our mediation analysis indicates that the reduced coercive leverage of plausible deniability is not driven by the fact that respondents are less likely to confidently attribute the attack when it goes unclaimed. In fact, subjects who more confidently attribute the attack to Russia are more likely to favor defying, rather than complying with, Moscow's demands. We interpret this mediation result as indirectly suggesting that the coercive disadvantage of plausible deniability might be driven by the reduced credibility of coercers' signals of resolve and corresponding reassurances, a proposition that future studies should systematically examine.

This article makes three main contributions. First, it bridges the literature on unclaimed coercive acts (Carson & Yarhi-Milo, 2017; Cormac & Aldrich, 2018) and the literature on the provocative effects of coercion attempts (Hall, 2017; Dafoe et al., 2021; Powers & Altman, 2023), which have largely developed in parallel, drawing insights from both about the potential benefits and drawbacks of unclaimed coercive acts for coercers. Second, to our knowledge, the

article conducts the first systematic comparison of the effects of claimed and unclaimed coercion attempts, thus representing an initial step in addressing our limited empirical understanding of these issues. By shedding light on the effects of unclaimed coercive acts, the article also indirectly contributes to the literature examining the trade-offs and strategic considerations shaping states' decisions to act covertly rather than overtly (Carson, 2018; Cormac & Aldrich, 2018; Joseph & Poznansky, 2018; O'Rourke, 2018; Poznansky, 2021). Third, the article provides policy-relevant insight into possible reactions by the United States to a hypothetical but realistic development in the ongoing war in Ukraine.

Theoretical perspectives on plausible deniability and coercion

Sometimes states inflict costs on rivals to influence their behavior in a desired direction but deny, or simply not claim, responsibility for the action, that is, they engage in unclaimed coercive acts. The key intended message is one of resolve, such as: "We are prepared to do all it takes to prevent you from achieving your goal" or "we will keep hurting you until you come to the negotiating table."[8]

A number of observers posit that, by operating under the cloak of plausible deniability, coercers can minimize escalation risks while still shaping the behavior of targets, in addition to circumventing the costs of appearing as norm violators. Yet, some insights from the literature on signaling suggest that unclaimed acts may offer limited coercive leverage due to the reduced clarity or credibility of the underlying message in the eyes of targets. In addition, targets' psychological responses and concerns about reputation and honor may conspire to make unclaimed acts ineffective at shielding coercers from escalation risks.

Taking stock of existing scholarship, we present competing perspectives on the effects of unclaimed coercive acts, from which we draw testable hypotheses. We organize the following discussion in two subsections, collecting arguments on the benefits and drawbacks of plausible deniability, respectively. It should be noted that these subsections are not meant to paint integrated and mutually exclusive pictures of plausible deniability. In reality, unclaimed coercive acts may offer their perpetrators a mixed bag of advantages and disadvantages compared to overt coercive efforts. We are agnostic about each of the propositions we test and the corresponding mix of benefits and drawbacks of plausible deniability.

Plausible deniability works: Unclaimed acts as low-risk and effective tools of coercion

For coercion to succeed (i.e., to induce the targets' behavioral changes intended by the coercer), targets need to understand what is being demanded of them and to believe that they would be better off complying rather than resisting (Schelling, 1966). Thus, the coercive message's demands and threatened penalties should be clear to targets, the consequences of defiance should be sufficiently serious to warrant concessions, and both signals of resolve and reassurances that penalties will be withheld in case of compliance should be credible to targets.

Carson and Yarhi-Milo (2017) make the most articulate case in the literature that covert activities, such as unacknowledged aid programs and military strikes, can meet the requirement of message clarity and thus can coerce (other works suggesting that unclaimed actions have coercive potential include Kugler, 2009; Byman & Kreps, 2010; Bowen, 2019; Cormac & Aldrich, 2018; Blagden, 2020).[9] As Carson and Yarhi-Milo (2017, p. 132) observe, "the basic contours of covert behavior are often visible" to states' rivals due to their intelligence

capabilities, but not necessarily to other audiences, which enables private, nonverbal communication. Drawing on the work of Schelling (1966), these authors posit that rival governments can understand the messages of resolve underlying one another's covert but observable behavior, due to mutually meaningful focal points and salient thresholds (Carson & Yarhi-Milo, 2017, pp. 131-132).

Carson and Yarhi-Milo (2017, pp. 133-134) argue that the message of resolve embedded in unclaimed coercive acts is likely not only to be understood by the government of the target country, but also to be believed, that is, to be credible: the sinking of costs involved in carrying out covert actions (e.g., arming a proxy), the risk of escalation due to retaliation by the government of the target country or to the coercer's loss of control of proxies relied on for plausible deniability, and the possibility of exposure of covert action to disapproving domestic actors all contribute to the credibility of the message of resolve. In other words, sinking-costs and raising-risks mechanisms make unclaimed coercive acts different from cheap talk, imbuing them with credibility in the eyes of a target country government that is broadly aware of the coercer's covert activities (Schelling, 1966; Fearon, 1997).

Though Carson and Yarhi-Milo (2017) focus on the intelligibility and credibility of the coercive message embedded in covert action in the eyes of the government of the target country, their logic should apply to the population of the target country if it finds out about the occurrence of the unclaimed coercive acts due to media exposure (an actual occurrence in the cases studied by those authors). In that scenario, the public would then be a second audience, besides the government of the target country, involved in deciphering and assessing the credibility of the coercer's message.

While the arguments discussed thus far suggest that unclaimed coercive efforts may be as effective as overt ones, Kurizaki's (2007) theory about private threats during crises suggests the possibility that unclaimed coercive actions could even be more effective. According to this theory, private threats are more likely to lead to the desired behavioral change because their secret nature enables the governments of target countries to make concessions without developing a reputation for weak resolve.[10] If covert action entails the exchange of nonverbal messages that are intelligible to the governments directly involved in the dispute but not necessarily observable by other audiences, as in some of the scenarios envisioned by Carson and Yarhi-Milo (2017), unclaimed attacks may enjoy analogous coercive advantages as explicit, yet private threats in crisis diplomacy: Target governments could adopt a behavior in line with the coercers' wishes while avoiding the appearance of having been coerced, which should increase the probability of compliance.

We draw our first hypothesis from the foregoing discussion:

H1a (Coercive effectiveness): Unclaimed actions are at least as effective as claimed actions in coercing targets.

Unclaimed actions may offer coercers the benefit of containing escalation risks. As Carson (2018) argues, unclaimed acts may enable governments to confine coercive bargaining to the "backstage"—away from the scrutiny of other audiences—thus withholding from domestic hawks in the target country information that they would otherwise leverage in their advocacy of a tougher stance towards the coercer. According to Carson, the covert nature of an action could help keep the lid on escalatory dynamics even if the population of the target country found out

about the occurrence of the relevant events, because an unclaimed act that is not acknowledged by the government of the target country may be perceived as less provocative by the latter's public than an overt act, thus detracting from the persuasiveness of domestic hawks' demands for escalatory responses.[11]

Other arguments, too, suggest that, due to their less provocative nature in the eyes of targets, unclaimed coercive efforts should be less escalatory than overt ones. Dafoe et al. (2021) argue that threats that are public and explicit are particularly likely to engage the honor and reputation of the government and the population of the target country alike, thus prompting retaliation. Along similar lines, Powers and Altman (2023) find that using subtle, as opposed to direct and dogmatic, language in coercive communication reduces the risk of retaliation, as the former type of language is less likely to activate reactance, a bundle of emotional and cognitive processes motivating targets to push back against attempts at constraining their autonomy. Claimed acts may be more similar than unclaimed ones to the kind of public, explicit, direct, and dogmatic coercive messages that often trigger retaliation. Furthermore, Carson (2018) argues that the use of covert action may also reduce the risk of escalation because it signals the coercer's willingness to keep the conflict limited, which the government of the target country (and its population, as long as it is aware of the coercive act) may have incentives to reciprocate.

Even in cases in which the government of the target country openly points its finger at the suspect, the unclaimed nature of the act, in the absence of smoking gun evidence, may reduce the probability of retaliation and escalation, given that other audiences – in particular, the populations of the target country and third-party countries – might be unsure about the identity of the culprit and thus oppose retaliatory responses. For instance, Byman and Kreps (2010, pp. 4-6) argue that Tehran's covert reliance on the services of Hezbollah facilitates Iran's coercive

diplomacy towards Israel and the United States, because their retaliation would be politically problematic without unassailable evidence of the Iranian role.[12]

Our second hypothesis captures the key implication of the foregoing arguments, that reliance on unclaimed coercive acts may reduce the risk of retaliation and subsequent escalation:

H2a (Escalation risk): Unclaimed coercive actions are less likely than claimed ones to provoke targets to engage in escalatory responses.

Plausible deniability does not work: Unclaimed acts as risky and ineffective tools of coercion

Other theoretical perspectives point in the opposite direction, suggesting that unclaimed coercive acts may be less effective than overt efforts and fail to provide significant escalation containment benefits.

A potential problem with unclaimed coercion attempts is that plausible deniability may wind up being all too plausible, meaning that targets may be left clueless about the identity of coercers. Coercion typically requires attribution, as the identity of the perpetrator is often inextricably connected to demands. The likely concessions being tacitly demanded with an unclaimed act may be radically different depending on whether China, North Korea, or Russia, say, is behind it. Gartzke's (2013, p. 47) rhetorical question about coercive cyber attacks from unidentified sources applies more broadly to any situation in which targets genuinely do not know who is trying to coerce them: "How does one surrender to no one in particular?"[13] Thus, the unclaimed nature of the act risks undermining the clarity of the coercive message by rendering targets utterly unsure about what behavior would qualify as compliance.

Plausible deniability may reduce the clarity of the message, and thus undermine coercive effectiveness, even if targets suspect a specific actor, rather than being completely unsure about the identity of the perpetrator. In the absence of evidence unmistakably pointing to the culprit, the unclaimed nature of the act may still reduce by some amount targets' confidence in attribution. In some cases, the resulting margin of uncertainty may tilt targets' cost-benefit calculus towards defiance of the coercer: The chance that paying the costs of making concessions valuable to suspected culprit A may be for naught, as the actual culprit, B, would not be appeased, may make targets reluctant to offer any concessions at all. Put differently, even marginal uncertainty about the identity of the coercer produced by the absence of a claim of responsibility may decisively reduce targets' confidence in the reassurance that any coercive message entails—"if you do as I say, I will stop tormenting you."[14]

Thus far, we have discussed how plausible deniability could undermine coercive effectiveness by reducing confidence in attribution and thus message clarity. Plausible deniability could also hinder coercion by reducing the credibility of the perpetrator's message of resolve. If plausible deniability can shield perpetrators from the costs of being seen as norm violators by third parties and contain escalation risks, targets may perceive unclaimed coercive acts as cheap. A cheap action may signal unwillingness to engage in costlier and riskier overt acts to prevail in the dispute, that is, low resolve (Schelling, 1966; Fearon, 1994). Targets may thus be emboldened to defy coercers' demands.

Furthermore, an unclaimed coercive act could reduce the credibility of the perpetrator's reassurance implied in the coercive act. The target may interpret the perpetrator's unwillingness to acknowledge responsibility as diagnostic of its deceitful nature, suggesting that violations of any deal promising the end of hostile acts in exchange for concessions would be likely. The

resulting reduction in the expected utility of complying with the coercer's demands should lower the coercive effectiveness of unclaimed coercive acts.

The preceding arguments about the obstacles to clear and credible messaging in unclaimed coercive acts lead us to our next hypothesis:

H1b (Coercive effectiveness): Unclaimed actions should be less effective than claimed ones in coercing targets.

Psychological processes may prompt targets of unclaimed acts to retaliate, thus denying the purported advantages of unclaimed coercion in terms of reduction of escalation risks. The anger experienced by targets—whether government leaders or ordinary citizens—of an unclaimed coercive effort may prompt them to confidently attribute the attack to the perpetrator, despite objective gaps in the available evidence (Lerner & Tiedens, 2006). Given that the action tendency of anger is to punish those responsible for an offensive action against one's self or group, anger-driven attribution may lead to retaliation, thus potentially unleashing an escalatory process.[15] Moreover, as Dafoe et al. (2021, p. 381) observe, an "incident can be provocative even when culpability is uncertain." For example, after the sinking of the USS Maine in Cuba 1898, "Remember the Maine, to Hell with Spain!" became the popular rallying-cry for war against Spain, even though it was unclear whether Spain was behind the sinking or it was an accident. Thus, plausible deniability may well fail to defuse the escalatory potential of coercive bargaining.

Our last hypothesis captures the logic of these arguments about the limited escalation avoidance benefits of unclaimed coercive acts:

H2b (Escalation risks): Unclaimed coercive actions are as likely as claimed ones to provoke targets to engage in escalatory responses.

Research Design

Experimental setup

We test our hypotheses with a vignette-based experiment fielded between May 17 and May 29, 2022. The vignette reports an explosion at a NATO base in Poland used by the alliance to funnel heavy weapons to Ukraine's armed forces resisting Russian invasion, causing the death of at least 100 soldiers. Before reading the vignette, subjects were asked a battery of questions about themselves and were then informed that the vignette they were about to read depicted a fictional scenario, but that similar events had occurred in the past and may occur again. We implemented a speeding check and two quality check questions, dropping from the sample subjects that failed any of them.[16]

All respondents were informed that on multiple previous occasions Russia's President Vladimir Putin had warned of "unpredictable consequences" if NATO continued providing heavy weapons to Ukraine. Given these veiled threats, subjects were told, intelligence agencies and independent analysts believed that most likely the explosion resulted from a bomb placed at the base by a Russian operative, though some members of the intelligence community and some analysts also noted that they could not rule out the possibility of an accidental detonation due to unsafe handling of the large amounts of weapons being stored in the base. The individuals

randomly assigned to the treatment group were then told that Moscow denied responsibility (UNCLAIMED=1). Participants in the control group were instead informed that, after initial silence, Russia claimed responsibility for the explosion (UNCLAIMED=0). All respondents were then asked a series of questions, in particular about their beliefs concerning whether Russia was responsible for the explosion and their views on various possible US responses. Note that a second treatment, assigned in a factorial (and thus independent) way, was used to address a distinct research question; this treatment varies the nationality (American or Polish) of the soldiers that died in the explosion. As Table A7 in the online appendix shows, our results are not affected by the inclusion of this second treatment variable in the analysis. Figure 1 below reports the treatment vignette with US casualties as presented to respondents; the online appendix contains the other vignettes.

Figure 1: Treatment group vignette

Major explosion at NATO base in Poland near Ukraine Border

Among the casualties at least 100 American soldiers

The Associated Press

An explosion devasted a NATO military base in Poland. At least 100 American soldiers recently deployed to the base have died in the explosion.

Destruction is being televised around the world, with bodies still being recovered from the rubble by emergency response teams.

Putin's earlier threats of "unpredictable consequences"

The base, located near the border with Ukraine, has been used by NATO to funnel heavy weapons to the Ukrainian armed forces, which have been resisting Russian invasion since February.

On multiple occasions Russia's President Vladimir Putin had previously warned of "unpredictable consequences" if the NATO alliance (of which the United States is a member) continued providing heavy weapons to Ukraine.

Russia's agent believed to be behind attack

In light of Putin's veiled threats, intelligence agencies and independent analysts believe that likely the explosion resulted from a bomb placed at the base by a Russian operative.

Experts did not rule out possibility of accident. Russia denied responsibility

Nonetheless, some members of the intelligence community and analysts noted that the possibility of an accidental detonation due to unsafe handling of the large amounts of new weapons being stored in the base cannot be ruled out.

Russia has denied responsibility for the explosion.

Following Sagan and Valentino (2017), the layout of our vignettes mimics typical newspaper articles, enabling us to emphasize and repeat key elements of the story, and in particular the treatment, in the headline and in pull quotes. Our experiment embeds within an ongoing international conflict a hypothetical but realistic event, given that in early March 2022 US officials publicly reported indications that Russia might attack supply lines used to transport weapons from Poland to Ukraine (Raddatz, 2022) and on March 13 Russia hit Ukrainian targets just a few miles from the border with Poland (Cathey, 2022).[17] The key objective driving our decision to adopt a highly realistic experimental setup was maximizing the odds that respondents' reactions would approximate those of the US public in case of the actual occurrence of events similar to those described in the vignette, given the policy importance of the war in Ukraine.

Our approach, however, has the potential drawback of limiting the generalizability of our findings. Subjects' responses may be influenced by idiosyncratic features of the war in Ukraine, such as its unusually high political salience for a foreign policy issue as well as strong beliefs and attitudes about Russia and Putin held by the US public. In particular, the facts that the US public appears to be strongly committed to providing military aid to Ukraine and that Putin's threats may lack credibility due to their frequency and vagueness might make the war in Ukraine a difficult testing ground for arguments positing the coercive effectiveness of unclaimed acts. That being said, Brutger et al. (2022, p. 14) find that varying the "hypotheticality" of the situation described in experimental vignettes and the degree of abstraction with which actors are identified generally does not matter, while adding contextual detail "leads to more conservative estimates of treatment effects, dampening treatment effects by hindering respondents' ability to successfully recall the main treatment." This finding suggests the possibility that re-running our experiment with a more abstract design might lead to broadly similar, or stronger, effects. Thus, further research would be helpful in clarifying the extent to which our findings travel beyond the Ukraine war case.

Given that our hypotheses apply to both the government and the population of the target country, it would be ideal to conduct survey experiments on both a general population sample and a sample of policymakers working on foreign policy (foreign policy elites). However, in light of the significant obstacles to conducting survey experiments with the second type of subjects (Renshon et al., 2018, p. 336), we opted to use a general population sample as a way to garner insights at both levels of analysis. Our sample consists of 854 adults residing in the United States, recruited via the Qualtrics online survey platform using sampling quotas to match US Census statistics for gender, age, and education.[18]

Using a general population sample to test theoretical expectations about the responses of both the government and the population of the target country is a sensible approach for two reasons. First, using a mass sample may yield indirect insight into the likely responses of policymakers. In fact, several studies suggest that, notwithstanding their distinct expertise and backgrounds, foreign policy elites' responses to experimental manipulation and real-world events may not radically differ from those of the broader public (Mercer 2013; Hall 2017; Yarhi-Milo 2018; Yarhi-Milo et al., 2018; Kertzer et al., 2021). In a meta-analysis of paired experimental studies on political elite and mass samples, Kertzer (2022) found that 88 percent of the treatment effects did not significantly differ in magnitude across the two sets of samples and that in 98 percent of the cases there was no significant difference in sign. Second, the public's reactions to international events may constrain policymakers' options. A number of studies show that policymakers risk significant public disapproval, or the opportunity cost of forgoing enhanced approval, if they fail to take tough action in response to foreign provocation (Kurizaki, 2007; Debs & Weiss, 2016; Narang & Staniland, 2018; Dafoe et al., 2022; Tomz et al., 2020). Emphasizing the dangers of conflict (Quek & Johnston, 2018; Clary et al., 2021) and resorting to rhetorical tactics such as bluster (Weiss & Dafoe, 2019) may enable policymakers to reduce these costs, but not necessarily to eliminate them.[19]

Dependent variables

We test hypotheses H1a and H1b about the coercive effectiveness of unclaimed acts by examining subjects' views about the continuation of heavy weapon provision by the United States to Ukraine in the aftermath of the explosion, the behavior that the vignette notes Putin is

trying to influence. The corresponding dependent variable is WEAPONS VIEW, a five-point indicator of respondents' views on the United States continuing to give heavy weapons to Ukraine and ranges from "strongly against" (1) to "strongly in favor" (5).

Our tests of hypotheses H2a and H2b about escalation risks focus on two possible US responses to the explosion: a single retaliatory air strike against a Russian military base on Ukrainian soil (a form of tit-for-tat escalation) and waging war against Russian forces in Ukraine (a form of vertical escalation). Respondents' views on these are captured by two five-point dependent variables, ranging from "strongly against" to "strongly in favor" – STRIKE VIEW and WAR VIEW.

Control variables

In addition to a bivariate analysis with our treatment as the sole explanatory variable, we ran a multivariate analysis with a series of pretreatment controls to ensure the robustness of our findings and to increase power. We use as controls the answers to questions devised by Kertzer and Brutger (2016) to capture respondents' militant assertiveness (i.e., hawkishness), national chauvinism, and international trust. The five-point scale hawkishness items are: "The best way to ensure world peace is through American military strength" ("strongly disagree"=1; "strongly agree"=5); "Going to war is unfortunate, but sometimes the only solution to international problems" ("strongly disagree"=1; "strongly agree"=5); and "The use of military force only makes problems worse" ("strongly agree"=1; "strongly disagree"=5). The four-point scale national chauvinism variables are the answers to two questions: "How superior is the United States compared to other nations?" ("not at all superior"=1; "vastly superior"=4) and "How

many things about America make you ashamed?" ("very many"=1; "none"=4). The dichotomous international trust variable is the answer to the question "Generally speaking, would you say that the United States can trust other nations, or that the United States can't be too careful in dealing with other nations?" (0=low trust; 1=high trust). Unlike in previous studies, in our sample the hawkishness variables do not load highly onto the same factor, thus we include all of them in the analysis. The same holds for the national chauvinism variables.

We also include controls for political ideology as well as feelings towards Vladimir Putin and Donald Trump (with feeling thermometers) to capture, respectively, respondents' priors about Russia's key decisionmaker and aspects of respondents' ideology and worldview that are not picked up by other variables.[20]

Empirical Analysis

Main findings

Table 1 reports bivariate and multivariate linear regression analyses. Our results indicate that unclaimed actions offer less coercive leverage than claimed ones, which is consistent with H1b but at odds with H1a. The significant, positive effect of UNCLAIMED shows that respondents are more supportive of continued heavy weapon provision to Ukraine when the explosion goes unclaimed. The effect of UNCLAIMED is substantively meaningful, as it amounts to a decrease by one-third (from 18 percent to 12 percent) in the share of respondents who are against or strongly against continuing to give weapons to Ukraine, compared to the scenario where Russia claims the attack.[21] This finding is robust to using a dichotomous version of the dependent

variable and an alternative dependent variable that captures respondents' preferences for continuing over interrupting the flow of heavy weapons to Ukraine, as distinct from their views (i.e., degree of approval) on the policy of giving weapons (see Table A1 in the online appendix).[22]

Table 1: Assessing the coercive and escalatory effects of plausible deniability

DV	H1a/b Weapons view	H1a/b Weapons view	H2a/b Strike view	H2a/b Strike view	H2a/b War view	H2a/b War view
Unclaimed	0.183**	0.185**	0.092	0.063	0.129	0.098
	(2.28)	(2.50)	(1.11)	(0.80)	(1.52)	(1.22)
Ideology		-0.057		-0.053		-0.040
		(-1.41)		(-1.25)		(-0.92)
International trust		0.199**		0.065		0.241***
		(2.40)		(0.74)		(2.68)
Shame		0.001		-0.059		-0.102**
		(0.03)		(-1.18)		(-2.00)
US superior		0.073		0.138**		0.038
		(1.43)		(2.56)		(0.69)
Use force		0.100**		0.086**		0.038
		(2.50)		(2.04)		(0.87)
Military strength		0.133***		0.165***		0.210***
		(3.52)		(4.12)		(5.13)
Going to war		0.230***		0.271***		0.233***
		(5.21)		(5.81)		(4.87)
Trump		-0.006***		-0.003***		-0.003**
		(-5.06)		(-2.71)		(-2.19)
Putin		-0.009***		0.004**		0.006***
		(-4.87)		(2.21)		(2.97)
Constant	3.657***	2.398***	2.761***	1.045***	2.582***	1.088***
	(64.94)	(12.22)	(47.00)	(5.02)	(43.14)	(5.11)
N	854	854	854	854	854	854

Notes: Coefficient estimates from linear regression models. Inference: *p < .1; **p < .05; ***p < .01 (t values in parentheses).

In our analysis of the two other possible responses to the explosion—an air strike against a Russian military base on Ukrainian soil and going to war against Russian forces in Ukraine—the estimate of the effect of UNCLAIMED is positive but fails to reach statistical significance. Thus, the evidence is consistent with hypothesis H2b and leads us to reject hypothesis H2a:

Unclaimed coercive acts do not offer escalation avoidance benefits compared to claimed acts. This null finding is robust to using dichotomous versions of the dependent variables and alternative dependent variables capturing respondents' preferences for launching an air strike rather than just continuing to provide weapons to Ukraine and for going to war with Russia rather than simply launching an air strike, though the sign of the coefficient for UNCLAIMED flips in some specifications (see Tables A2-A3 in the appendix).[23]

Plausible deniability could offer coercers benefits of reduced escalation risk resulting from political dynamics within the target country not captured by our experiment. For example, if unclaimed attacks make the target country's population less angry, less concerned about reputation, or less suspicious about the intentions of the alleged culprit than claimed attacks, target country's policymakers that prefer to avoid escalatory responses may not face a significant popular backlash, while policymakers bent on escalation would lack popular backing. The evidence at our disposal, however, does not offer grounds to expect this type of benefit, either. In fact, the unclaimed nature of the act does not affect respondents' self-reported anger levels, concerns about the United States' reputation, and perception of the aggressiveness of Russia's intentions (see Table A4).

Given that the three outcome variables we examine are correlated (in particular respondents' views about the air strike and going to war responses) residuals across regression models may be correlated, too. Thus, we use seemingly unrelated regression, which can yield more precise estimates, as an additional robustness check. As Table A5 shows, our results are substantively unchanged.

In sum, our analysis indicates that not claiming an act puts coercers at a disadvantage in terms of influence on target behavior and does not offer countervailing benefits of reduced

escalation risk. The fact that the unclaimed nature of the coercive act reduces Russia's coercive leverage on respondents but also does not prompt them to support escalatory measures may reflect perceived differences in risk of loss of American lives across alternative courses of actions, with retaliatory responses being seen as entailing more serious risks than continued military aid to Ukraine. Our data lends some support to this interpretation as there is a negative, significant association between support for the retaliatory strike and going to war with Russia, on the one hand, and self-reported importance of "avoiding the risk of a war with Russia that could cost many lives" as a factor affecting subjects' views of possible responses, on the other. By contrast, the negative association between concern for loss of life and support for continuing to provide weapons to Ukraine does not reach statistical significance (see Table A6).

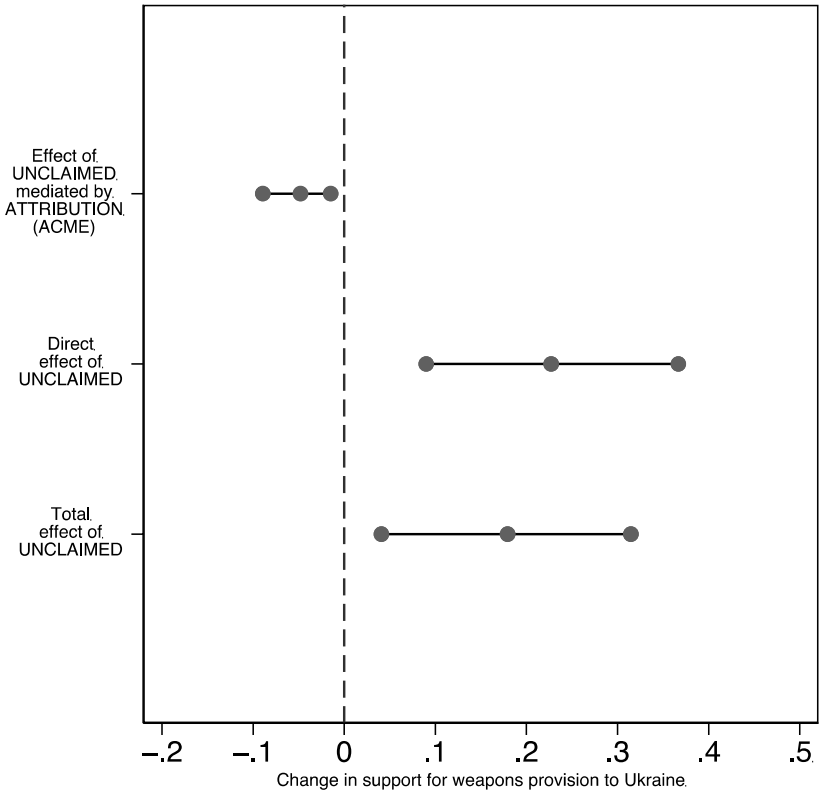
Why does plausible deniability reduce coercive leverage?

Our theoretical discussion identified two general reasons why unclaimed acts may entail reduced coercive leverage: uncertain attribution and limited credibility of signals of resolve and corresponding reassurances. We designed our experiment for the primary purpose of assessing the existence of effects of plausible deniability on targets' behavior, rather than examining underlying mechanisms, a task that we hope future studies will take up. Nonetheless, the data at our disposal allows us to conduct a preliminary test of the uncertain attribution mechanism.

Figure 2 reports results of mediation analysis with the procedure developed by Imai et al. (2019), using as mediator ATTRIBUTION, a five-point variable corresponding to subjects' answers to a question about the likelihood that Russia conducted the attack, ranging from "very unlikely" to "very likely." The average causal mediation effect (ACME) is the expected

difference in the outcome (WEAPONS VIEW) when ATTRIBUTION takes the value it would realize with UNCLAIMED equal to 1, as opposed to 0, while UNCLAIMED is held fixed. The direct effect is the expected difference in the outcome when UNCLAIMED goes from 0 to 1 while the mediator is held constant, and the total effect the sum of the direct effect and ACME.[24]

Figure 2: Mediation analysis



Note: ACME stands for average causal meditation effect. The horizontal bars correspond to 95 percent confidence intervals.

The negative, significant ACME indicates that the absence of a Russian claim of responsibility reduces respondents' confidence in attributing the explosion to Russia, which in

turn makes them *less* likely to support the continued transfer of weapons to Ukraine. The fact that the mediating effect of ATTRIBUTION goes in the opposite direction as the direct effect of our treatment allows us to rule out uncertain attribution as an explanation for the reduced coercive effectiveness of unclaimed attacks: unclaimed attacks do not lack coercive leverage because targets are unsure about who is responsible. Though future research should explore other mechanisms, we interpret this finding as suggesting that limited credibility of signals of resolve and reassurances may be behind the ineffectiveness of unclaimed coercive acts.[25]

The finding of a negative mediation effect for ATTRIBUTION on WEAPONS VIEW is surprising, as our earlier discussion of competing theoretical perspectives about plausible deniability emphasized that the absence of a claim of responsibility should make targets unsure about what behavior would constitute compliance and thus reduce the odds of coercive success. Instead, we observe the opposite, as respondents who confidently attribute an attack due to Russia's claim of responsibility are less likely to support compliance with its demands. The concept of reactance (Powers & Altman, 2023) suggests a possible explanation for this finding: by making respondents more likely to see themselves/their country as targets of coercion, confident attribution of the attack to Russia due to Moscow's claim activates a preference for defiance over compliance in the face of coercion ("let's stick it to Russia by providing more weapons to Ukraine!"). This preference is not comparably activated in subjects that do not see themselves/their country as targets of coercion, which would explain why respondents who do not confidently attribute the attack to Russia are less supportive of continuation of military aid to Ukraine.

Conclusions

When it comes to shaping the behavior of targets in the direction intended by a coercer and containing escalation risks, our analysis offers a negative answer to the question the article's title poses: Does plausible deniability work? US-based participants in our survey experiment were more supportive of continuing the provision of weapons to Ukraine—the policy that Putin had demanded to be discontinued—when Russia denied involvement in an explosion at a NATO base in Poland compared to when it claimed responsibility. Furthermore, the absence of a Russian claim of responsibility did not have a significant effect on respondents' support for escalatory US responses to the explosion. Thus, our study points to reduced coercive leverage for unclaimed coercive acts, compared to overt ones, and to the absence of countervailing benefits in terms of a lower risk of escalation.

None of this should be interpreted as implying that plausible deniability does not offer coercers *any* advantage. In particular, by denying responsibility coercing states might induce sufficient uncertainty in the minds of their domestic publics and third parties (such as other countries and international organizations) about who carried out the unclaimed act to substantially reduce corresponding image costs.

Future studies should systematically explore this type of potential benefit of plausible deniability by examining the responses of respondents from third-party countries and from the coercing state. What our analysis does permit us to say is that, within the confines of the relationship between the coercer and its targets, we find no support for plausible deniability's much touted benefit of low escalation risks, while we find evidence of coercive ineffectiveness. Besides assessing the robustness of our findings to different types of coercive acts, demands,

actors, and disputes, future research should examine whether "non-denial denials" (i.e., situations when the suspect neither confirms nor denies responsibility) have similar effects as the explicit denials that we study (on non-denial denials, see Brown & Fazal, 2021). Empirically parsing out underlying causal mechanisms is another important task for further research.

Future studies may also incorporate our findings in theorizing about why states and non-state actors sometimes opt to openly attempt coercion, while other times they deny or refrain from claiming responsibility for coercive acts. The absence of evidence of coercive and escalation-containment benefits emerging from our study suggests the possibility that image preservation in the eyes of domestic and/or third-party audiences is a more important driver of states' decisions to resort to plausible deniability, in line with several existing studies (e.g., Downes & Lilley, 2010; O'Rourke, 2018; Poznansky, 2019). Alternatively, it may be that states embrace plausible deniability in the misguided anticipation of significant coercive and escalation-containment benefits, which in turn raises intriguing questions about the sources of this misperception.

Though we cannot rule out the possibility that US policymakers would react differently from the public, the key policy implication is that our study provides no grounds for believing that by engaging in unclaimed coercive attacks Russia would be able to reap substantial strategic benefits—plausible deniability is not a silver bullet that would magically solve Putin's predicament in the war in Ukraine. More generally, while keeping in mind the caveat that generalizability to other contexts remains to be examined, our findings suggest that policymakers considering resorting to plausibly deniable actions or fretting about adversaries doing so should be aware of the fact that the coercive and escalation-containment payoffs of unclaimed acts are likely to be limited.

Reference list

Abrahms, M. (2008). What terrorists really want: Terrorist motives and counterterrorism strategy. *International Security*, *32*(4), 78–105. https://doi.org/10.1162/isec.2008.32.4.78

Associated Press-NORC Center for Public Affairs Research. (2022). *The May 2022 AP-NORC Center poll*. https://apnorc.org/wp-content/uploads/2022/05/AP-NORC-May-2022-Topline-Biden.pdf

Bauer, V., Ruby, K., & Pape, R. (2017). Solving the problem of unattributed political violence. *Journal of Conflict Resolution*, *61*(7), 1437–1564. https://doi.org/10.1177/0022002715612575

Berrier, S. (2022). *Worldwide threat assessment*. Armed Services Committee. Intelligence and Special Operations Subcommittee. United States House of Representatives. http://docs.house.gov/meetings/AS/AS26/20220317/114527/HHRG-117-AS26-Wstate-BerrierS-20220317.pdf

Blagden, D. (2020). Deterring cyber coercion: The exaggerated problem of attribution. *Survival*, *62*(1), 131–148. https://doi.org/10.1080/00396338.2020.1715072

Borghard, E., & Lonergan, S. (2017). The logic of coercion in cyberspace. *Security Studies*, *26*(3): 452–481. https://doi.org/10.1080/09636412.2017.1306396

Bowen, A. (2019). Coercive diplomacy and the Donbas: Explaining Russian strategy in Eastern Ukraine. *Journal of Strategic Studies*, *42*(3–4), 312–343. https://doi.org/10.1080/01402390.2017.1413550

Brown, J., & Fazal, T. (2021). #SorryNotSorry: Why states neither confirm nor deny responsibility for cyber operations. *European Journal of International Security*, *6*(4), 401–417. https://doi.org/10.1017/eis.2021.18

Brutger, R., Kertzer, J. D., Renshon, J., Tingley, D., & Weiss, C. M. (2022). Abstraction and detail in experimental design. *American Journal of Political Science*, *0*(0), 1–16. https://doi.org/10.1111/ajps.12710

Byman, D., & Kreps, S. (2010). Agents of destruction? Applying principal-agent analysis to state-sponsored terrorism. *International Studies Perspective*, *11*(1), 1–18. https://doi.org/10.1111/j.1528-3585.2009.00389.x

Carnegie, A., & Carson, A. (2020). *Secrets in global governance: Disclosure dilemmas and the challenge of international cooperation*. Cambridge University Press.

Carson, A. (2019, September 17). After the Saudi oil attack, will the U.S. and Saudis start a war with Iran? Here are 3 things to know. *The Washington Post.*

https://www.washingtonpost.com/politics/2019/09/17/after-saudi-oil-attack-will-us-saudis-start-war-with-iran-here-are-things-know/

Carson, A. (2018). *Secret wars: Covert conflict in international politics*. Princeton University Press.

Carson, A., & Yarhi-Milo, K. (2017). Covert communication: The intelligibility and credibility of signaling in secret. *Security Studies*, *26*(1), 124–156. https://doi.org/10.1080/09636412.2017.1243921

Cathey, L. (2022, March 14). Russian missile strike near Poland raises tough questions for Biden. It remains unclear what consequences Russia could still face from the U.S. *ABC News*. https://abcnews.go.com/Politics/russian-missile-strike-poland-raises-tough-questions-biden/story?id=83436540

Clary, C., Lalwani, S., & Siddiqui, N. (2021). Public opinion and crisis behavior in a nuclearized South Asia. *International Studies Quarterly*, *65*(4), 1064–1076. https://doi.org/10.1093/isq/sqab042

Cohen, E. T., & Huggard, K. (2019, December 6). *What can we learn from the escalating Israeli raids in Syria?* Order From Chaos. Brookings. https://www.brookings.edu/blog/order-from-chaos/2019/12/06/what-can-we-learn-from-the-escalating-israeli-raids-in-syria/

Cordesman, A. H. (2019). *The strategic threat from Iranian hybrid warfare in the Gulf*. Center for Strategic and International Studies https://www.csis.org/analysis/strategic-threat-iranian-hybrid-warfare-gulf#:~:text=The%20Threat%20of%20Hybrid%20Warfare,near%20the%20Strait%20of%20Hormuz

Cormac, R., & Aldrich, R. (2018). Grey is the new black: Covert action and implausible deniability. *International Affairs*, *94*(3), 477–494. https://doi.org/10.1093/ia/iiy067

Crenshaw, M., & LaFree, G. (2017). *Countering terrorism*. Brookings. https://www.brookings.edu/book/countering-terrorism-no-simple-solutions/

Dafoe, A., Liu, S., O'Keefe, B., & Weiss, J.C. (2022). Provocation, public opinion, and international disputes: evidence from China. *International Studies Quarterly*, *66*(2), 1–14. https://doi.org/10.1093/isq/sqac006

Dafoe, A., Hatz, S., & Zhang, B. (2021). Coercion and provocation. *Journal of Conflict Resolution*, *65*(2–3), 372–402. https://doi.org/10.1177/0022002720957078

Debs, A., & Weiss, J. C. (2016). Circumstances, domestic audiences, and reputational incentives in international crisis bargaining. *Journal of Conflict Resolution*, *60*(3), 403–433. https://doi.org/10.1177/0022002714542874

Downes, A. B., & Lilley, M. L. (2010). Overt peace, covert war? Covert intervention and the democratic peace. *Security Studies*, *19*(2), 266–306. https://doi.org/10.1080/09636411003795756

Economist/YouGov. (2022). *Russia and Ukraine poll* https://today.yougov.com/topics/international/articles-reports/2022/05/18/russia-and-ukraine-economistyougov-poll-may-15-17

Fearon, J. D. (1997). Signaling foreign policy interests: Tying hands versus sinking costs. *Journal of Conflict Resolution*, *41*(1), 68–90. https://doi.org/10.1177/0022002797041001004

Fearon, J. D. (1994). Domestic political audiences and the escalation of international disputes. *American Political Science Review*, *88*(3), 577–592. https://doi.org/10.2307/2944796

Gartzke, E. (2013). The myth of cyberwar: Bringing war in cyberspace back down to earth. *International Security*, *38*(2), 41–73. https://doi.org/10.1162/ISEC_a_00136

Hall, T. H. (2017). On provocation: outrage, international relations, and the Franco–Prussian War. *Security Studies*, *26*(1), 1–29. https://doi.org/10.1080/09636412.2017.1243897

Hicks, R. & Tingley, D. (2012). Causal mediation analysis. *Stata Journal*, *11*(4), 605–619. https://doi.org/10.1177/1536867X1201100407

Hoffman, A. M. (2010). Voice and silence: Why groups take credit for acts of terror. *Journal of Peace Research*, *47*(5), 615–26. https://doi.org/10.1177/0022343310376439

Hoffman, B. (1997). Why terrorists don't claim credit. *Terrorism and Political Violence*, *9*(1), 1–6. https://doi.org/10.1080/09546559708427381

Hur, M. (2017). Revisiting the Cheonan sinking in the Yellow Sea. *The Pacific Review*, *30*(3), 348–364. https://doi.org/10.1080/09512748.2016.1249905.

Imai, K., Keele, L., Tingley, D. & Yamamoto, T. (2019). *Causal mediation analysis using R* (Working Paper). https://cran.ism.ac.jp/web/packages/mediation/vignettes/mediation-old.pdf

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334. https://doi.org/10.1037/a0020761

Joseph, M., & Poznansky, M. (2018). Media technology, covert action, and the politics of exposure. *Journal of Peace Research*, *55*(3), 320–335. https://doi.org/10.1177/0022343317731508

Kearns, E. M. (2021). When to take credit for terrorism? A cross-national examination of claims and attributions. *Terrorism and Political Violence*, *33*(1), 164–193. https://doi.org/10.1080/09546553.2018.1540982

Kearns, E. M., Conlon, B., & Young, J.K. (2014). Lying about terrorism. *Studies in Conflict & Terrorism*, *37*(5), 422–439. https://doi.org/10.1080/1057610X.2014.893480

Kertzer, J. D. & Brutger, R. (2016). Decomposing audience costs: Bringing the audience back into audience cost theory. *American Journal of Political Science*, *60*(1), 234–249. https://doi.org/10.1111/ajps.12201

Kertzer, J. D., Rathbun, B. C., & Rathbun, N. S. (2020). The price of peace: Motivated reasoning and costly signaling in international relations. *International Organization*, *74*(1), 95–118. https://doi.org/10.1017/S0020818319000328

Kertzer, J. D., Renshon, J., & Yarhi-Milo, K. (2021). How do observers assess resolve? *British Journal of Political Science, 5*(1), 308–330.

Kertzer, J. D., & Zeitzoff, T. (2017). A bottom-up theory of public opinion about foreign policy. *American Journal of Political Science*, *61*(3), 543–558. https://doi.org/10.1111/ajps.12314

Kertzer, J. D. (2022). Re-assessing elite-public gaps in political behavior. *American Journal of Political Science*, *66*(3), 539–553. https://doi.org/10.1111/ajps.12583

Kugler, R. L. (2009). Deterrence of cyber attacks. In F. D. Kramer, S. H. Starr, & L. Wentz (Eds.), *Cyberpower and national security* (pp. 309–340). Potomac Books.

Kurizaki, S. (2007). Efficient secrecy: Public versus private threats in crisis diplomacy. *American Political Science Review*, *101*(3), 543–558. https://doi.org/10.1017/S0003055407070396

Kydd, A. H. & Walter, B. F. (2006). The strategies of terrorism. *International Security*, *31*(1), 49–80. https://doi.org/10.1162/isec.2006.31.1.49

Lerner, J. S. & Tiedens, L. Z. (2006). Portrait of the angry decision maker: How appraisal tendencies shape anger's influence on cognition. *Journal of Behavioral Decision Making*, *19*(2), 115–137. https://doi.org/10.1002/bdm.515

Libicki, M. C. (2009). *Cyberdeterrence and cyberwar*. RAND Corporation. https://www.rand.org/pubs/monographs/MG877.html

Lieber, K. & Press, D. (2013). Why states won't give nuclear weapons to terrorists. *International Security*, *38*(1), 80–104. https://doi.org/10.1162/ISEC_a_00127

Lindsay, J. R. (2015). Tipping the scales: The attribution problem and the feasibility of deterrence against cyberattack. *Journal of Cybersecurity*, *1*(1), 53–67. https://doi.org/10.1093/cybsec/tyv003

Markwica, R. (2018). *Emotional choices: How the logic of affect shapes coercive diplomacy.* Oxford University Press.

McDermott, R., Lopez, A., & Hatemi, P. (2017). Blunt not the heart, enrage it: The psychology of revenge and deterrence. *Texas National Security Review, 1*(1), 68–88. http://hdl.handle.net/2152/63934

Narang, V., & Staniland, P. (2018). Democratic accountability and foreign security policy: Theory and evidence from India. *Security Studies*, *27*(3), 410–447. https://doi.org/10.1080/09636412.2017.1416818

O'Rourke, L. (2018). *Covert regime change: America's secret Cold War*. Cornell University Press.

Quek, K., & Johnston, A. I. (2018). Can China back down? Crisis de-escalation in the shadow of popular opposition. *International Security*, *42*(3), 7–36. https://doi.org/10.1162/ISEC_a_00303

Pape, R. A. (1996). *Bombing to win: Airpower and coercion in war*. Cornell University Press.

Pauly, R. B. C. (2019). *Stop or I'll shoot, comply and I won't: Coercive assurance in international politics* [Unpublished doctoral dissertation]. Massachusetts Institute of Technology.

Petersen, R. D. (2011). *Western intervention in the Balkans: The strategic use of emotion in conflict.* Cambridge University Press.

Pluchinsky, D. (1997). The terrorism puzzle: Missing pieces and no boxcover. *Terrorism and Political Violence*, *9*(1), 7–10. https://doi.org/10.1080/09546559708427382

Powers, K. E., & Altman, D. (2023). The psychology of coercion failure: How reactance explains resistance to threats. *American Journal of Political Science*, *67*(1), 221–238. https://doi.org/10.1111/ajps.12711

Poznansky, M. (2021). *In the shadow of international law: Secrecy and regime change in the postwar world*. Oxford University Press.

Poznansky, M. (2019). Feigning compliance: Covert action and international law. *International Studies Quarterly*, *63*(1), 72–84. https://doi.org/10.1093/isq/sqy054

Raddatz, M. (2022, March 7). Officials concerned Russia may strike supply line to Ukraine. *ABC News*. https://abcnews.go.com/International/live-updates/russia-ukraine/?id=83184729#83308582

Renshon, J., Dafoe, A., & Huth, P. (2018). Leader influence and reputation formation in world politics. *American Journal of Political Science, 62*(2), 325–339. https://doi.org/10.1111/ajps.12335

Rid, T., & Buchanan, B. (2015). Attributing cyber attacks. *Journal of Strategic Studies*, *38*(1–2), 4–37. https://doi.org/10.1080/01402390.2014.977382

Sagan, S. D., & Valentino, B. A. (2017). Revisiting Hiroshima in Iran: American attitudes about nuclear weapons and non-combatant immunity. *International Security*, *42*(1), 41–79. https://doi.org/10.1162/ISEC_a_00284

Schelling, T. (1966). *Arms and influence*. Yale University Press.

Sharp, T. (2017). Theorizing cyber coercion: The 2014 North Korean operation against Sony. *Journal of Strategic Studies*, *40*(7), 898–926. https://doi.org/10.1080/01402390.2017.1307741

Snyder, J. (2020). Backlash against human rights shaming: Emotions in groups. *International Theory*, *12*(1), 109–132. https://doi.org/10.1017/S1752971919000216

Snyder, G. H. (1959). *Deterrence by denial and punishment*. Woodrow Wilson School of Public and International Affairs, Center of International Studies, Princeton University.

Tomz, M., Weeks J. L. P., & Yarhi-Milo, K. (2020). Public opinion and decisions about military force in democracies. *International Organization*, *74*(1), 119–143. https://doi.org/10.1017/S0020818319000341

Valeriano, B., Jensen, B., & Maness, R. (2018). *Cyber strategy: The evolving character of cyber power and coercion*. Oxford University Press.

Weiss, J. C., & Dafoe, A. (2019). Authoritarian audiences, rhetoric, and propaganda in international crises: Evidence from China. *International Studies Quarterly*, *63*(4), 963–973. https://doi.org/10.1093/isq/sqz059

Yarhi-Milo, K. (2018). *Who fights for reputation: The psychology of leaders in international conflict*. Princeton University Press.

Yarhi-Milo, K., Kertzer, J. D., & Renshon, J. (2018). Tying hands, sinking costs, and leader attributes. *Journal of Conflict Resolution*, *62*(10), 2150–2179. https://doi.org/10.1177/0022002718785693

[1] We use the term coercion broadly, to include both compellence and deterrence.

[2] Following Carson and Yarhi-Milo (2017, p. 128), we conceive of covert action as "a variety of secret foreign policy actions that may be administered by military or intelligence bureaucracies … in a way that conceals and renders deniable the role of the sponsoring state for most audiences."

[3] Thus, in a Venn diagram covert action and unclaimed coercive acts would overlap only partially. Some unclaimed coercive acts would not qualify as covert action because of the identity of the perpetrator (e.g., unclaimed terrorist attacks by a non-state actor following a coercive strategy), while some instances of covert action would not amount to unclaimed coercive acts because they only aim at on-the-ground effects. However, as Carson and Yarhi-Milo (2017, p. 133) note, coercive and on-the-ground goals may both be present at the same time in a given situation. For example, an unclaimed attack on a military facility may aim at both crippling a rival's capabilities and signaling the perpetrator's resolve.

[4] In other cases of unclaimed actions that are not primarily coercive in nature, publicity, though not attribution to a specific actor, is necessary, as these actions are not carried out to extract concessions from a target. For example, terrorist attacks may follow a spoiling strategy, seeking to undermine trust in a peace process, or a destabilization strategy, aiming to create a general climate of chaos (Kearns et al., 2014; Kydd & Walter, 2006).
[5] Recent works on secrecy in international politics include Carnegie & Carson, 2020, Carnegie, 2021, and Poznansky, 2021.
[6] Our approach to this levels-of-analysis issue draws inspiration from Renshon et al., 2018 (p. 326) and Dafoe et al., 2021 (p. 383).
[7] We use the terms "coercer" and "target" to refer to the perpetrator of a specific coercive action and to the actor on the receiving end, respectively. In the broader process of coercive bargaining, the target of a given coercive action may be the perpetrator of some other act of coercion, that is, all parties involved may be trying to coerce others.
[8] In line with much of the literature, we conceptualize resolve as determination to pay high costs and run high risks to advance one's interests.
[9] Note that unclaimed coercive acts, like their claimed counterparts, can follow both punishment and denial logics (Snyder, 1959; Pape, 1996). Punishment succeeds by raising targets' expected costs of defiance above the value of targets' interests in the dispute. Denial succeeds by lowering targets' expected probability of ultimately being able to protect their interests in the dispute to the point that continued defiance becomes futile.
[10] Though Kurizaki (2007) focuses on the vulnerability of governments to domestic accusations of capitulation, the logic of the argument applies more broadly to any case in which giving in to threats before the eyes of third-party audiences, whether domestic or international, can lead an actor to acquire a reputation for weak resolve with such audiences. Pauly (2019) introduces the related concept of "visibility reduction," the idea that coercive success is made more likely by coercers' credible reassurance to targets that concessions will be kept secret.
[11] In fact, Carson (2018) posits that rival governments often collude to keep one another's coercive actions secret, or unacknowledged when exposed by the media, specifically for the purpose of reducing escalation risks.
[12] Similarly, in discussing the September 2019 attack on Saudi oil facilities, allegedly sponsored by Iran, Carson (2019) observes that "[n]either the United States nor the Saudis would be able to justify a harsh military response" if the "attack can't be definitively pinned on Iran. … Ambiguity would make any response look illegitimate while jeopardizing allies' support." This perspective on the escalation containment benefits of plausible deniability is shared by policymakers, too. See for example, the statement to the House by Scott Berrier (2022, p. 23), the Defense Intelligence Agency's director.
[13] For similar observations, see Abrahms, 2008, pp. 89-90 and Borghard & Lonergan, 2017, p. 459. In some cases, targets may not even be able to tell apart coercive acts from human-caused and natural accidents.
[14] Even in instances where the government of the target country has access to information enabling it to confidently attribute an unclaimed act, if this information cannot be shared with the public (to protect sources and methods or to keep a lid on escalatory pressures), the public of the target country may remain unsure about the identity of the perpetrator, thus constraining the government's ability to make the concessions demanded by the coercer.
[15] Studies on the effects of anger in coercive settings include Petersen, 2011; Hall, 2017; McDermott et al., 2017; Markwica, 2018; and Snyder, 2020.
[16] All subjects were asked two manipulation check questions, too. Results discussed below are robust to dropping subjects that failed both manipulation checks.
[17] Our approach is similar to that adopted by Quek & Johnston (2018), whose experimental treatment is a hypothetical event within the ongoing Sino-Japanese dispute over ownership of the Diaoyu/Senkaku Islands. See also the experiment presented by Kertzer et al. (2020), which manipulates Iranian signals of reassurance to a US audience at the time when the details of the "Iran nuclear deal" were being negotiated.
[18] We determined sample size with the goal in mind of being able to detect with 80 percent power a 0.2 difference in means between two samples in 5-point scale Likert variables (ranging from 1 to 5), assuming a standard deviation of 1. With these parameters, a minimum of 394 respondents are need for the treatment group and another 394 for the control group. We thus rounded up our sample to 800 (the additional subjects were already taking the survey when the 800 limit was met).
[19] On the limits of the ability of elites to shape public narratives and the importance of bottom-up understandings, see Kertzer & Zeitzoff, 2017.
[20] Our findings are robust to the inclusion of a standard measure of partisanship (see Table A8 in the appendix).
[21] Given that our experimental design does not envision a condition without an explosion, we cannot directly assess whether our finding indicates that, compared to the no-explosion baseline, the unclaimed coercive act is

emboldening respondents to support weapon provision to Ukraine, is having no effect on their behavioral intentions, or is producing some coercive leverage, albeit less than a claimed act. However, two polls conducted at around the same time as our survey by the Economist/YouGov (May 15-17, 2022) and the Associated Press-NORC Center for Public Affairs Research (May 12-16, 2022) help shed some light on this issue. The Economist/YouGov poll asked whether sending weapons to Ukraine was a good or a bad idea for the United States to pursue (possible answers: good idea, bad idea, don't know); the Associated Press-NORC poll asked respondents whether they favored, opposed or neither favored nor opposed providing weapons to Ukraine (possible answers: strongly favor, somewhat favor, neither favor nor oppose, somewhat oppose, strongly oppose). The two polls found that 17 percent and 19 percent of US adults, respectively, opposed providing weapons to Ukraine. If we take the 17-19 percent figure from these polls as the baseline level of opposition to weapon provision in the absence of any explosion at a NATO base in Poland, the 12 percent opposition that we observe in our sample when the explosion goes unclaimed would suggest that the absence of a Russian claim of responsibility has an emboldening effect on US public opinion. On the other hand, the fact that when the attack is claimed the share of our respondents opposing weapon provision (18 percent) is about the same as in the two polls is suggestive of the absence of a coercive effect. Thus, it would appear that unclaimed actions embolden targets to defy coercers' demands, rather than simply having less coercive power than claimed ones.

[22] The five-point alternative dependent variable ranges from "strongly prefer no longer giving weapons" to "strongly prefer continuing to give weapons."

[23] The two five-point alternative dependent variables range, respectively, from "strongly prefer just continuing to give weapons" (1) to "strongly prefer also launching single air strike" (5), and from "strongly prefer single air strike" (1) to "strongly prefer going to war" (5).

[24] To address concerns about a possible violation of the "sequential ignorability assumption" on which mediation analysis relies, it is important to control for possible confounders (Imai et al., 2010). The analysis underlying Figure 2 uses the same list of controls as in Model 2 in Table 1. Figure A1 in the appendix is based on an analysis with additional controls.

[25] We also conducted mediation analysis using as mediators self-reported subjects' levels of anger experienced in response to the scenario, their concerns about US reputation, and their assessment of the aggressiveness of Russia's intentions. None of the mediators has a significant ACME (see Figures A1-A3).