The Psychology of Coercion Failure: How Reactance Explains Resistance to Threats*

Kathleen E. Powers† and Dan Altman‡

American Journal of Political Science, Forthcoming

Abstract: When confronted with coercive threats, targets often stand firm rather than back down. We identify one important yet unrecognized factor that causes actors to resist threats: psychological reactance. Reactance theory explains that when someone perceives a threat to their freedom to make choices, they attempt to restore their autonomy by refusing to capitulate. The result is unwillingness to concede to coercion that extends beyond rational incentives. We test for reactance as a cause of coercion failure with two novel experiments. Each experiment pairs a coercive threat treatment with a matched 'natural costs' counterpart that imposes the same choice on the target without intentional action by a coercer. Controlling for prominent alternative explanations including costs, benefits, power, credibility, and reputation, we find that the targets of threats capitulate less frequently and more often support aggression against their opponents.

Verification Materials: The data, code, and any additional materials required to replicate all analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/0UTWWV

Word count: 9714 (7716 in text + 1661 references + 337 Figure and Table captions/notes)

*We thank Steve Brooks, Jeffrey Friedman, Sophia Hatz, Phil Haun, Marcus Holmes, Danielle Lupton, Rose McDermott, Paul Orner, Andrew Payne, Jason Reifler, Jonathan Renshon, Patricia Sullivan, Michael Tomz, Ben Valentino and audiences at Columbia University, Dartmouth College, Georgia State University, George Washington University, the 2019 and 2020 Annual Meetings of the American Political Science Association, the 2020 Annual Meeting of the Peace Science Society, and the 2021 and 2022 Annual Meetings of the International Studies Association for helpful comments. Steven Li, Josie Pearce, and Ryan Waaland provided excellent research assistance.

†Corresponding author. Assistant Professor, Department of Government, Dartmouth College. kathleen.e.powers@dartmouth.edu. http://kepowers.com

‡Assistant Professor, Department of Political Science, Georgia State University. daltman@gsu.edu. http://www.danielwaltman.com/

This article investigates the possibility that a psychological aversion to capitulation contributes to explaining the prevalence of coercion failure in international politics. During the Cold War, for example, Soviet coercion failed to gain control of the precarious, indefensible enclave of West Berlin despite a blockade in 1948 and ultimatums in 1958 and 1961. Decades of U.S. sanctions failed to induce the Communist Party of Cuba to submit. In 1914, German ultimatums demanding that Russia cease mobilization and Belgium allow unobstructed passage both failed. Britain's ultimatum to Germany to vacate Belgian territory then followed suit amid a crisis that escalated after Serbia rebuffed a key Austro-Hungarian demand while relenting to others. Coercion has succeeded in some important cases — including when Serbia acceded to some of Austria-Hungary' demands — but leaders defy coercive threats so often that Art and Greenhill (2018, 78) describe the "track record" for compellent threats that demand changes to the status quo as "underwhelming."

We propose that an established concept from psychology — reactance (Brehm, 1966; Miron and Brehm, 2006; Rosenberg and Siegel, 2018) — provides an important yet unexamined explanation for why coercion often fails in international politics. The intuition is as follows: Imagine that someone tells you what to do. They threaten you with consequences if you do not comply. Do you dispassionately weigh the costs and benefits of acquiescence? Or do you immediately feel motivated to refuse — to react to their diktat by digging in your heels? Reactance theory tells us that when a person believes that another actor is trying to constrain her autonomy — such as issuing a coercive threat that attempts to dictate her policy choices — she will try to restore her freedom through defiance and aggression. Consequently, coercion operates at an inherent psychological disadvantage: Coercion itself induces a psychological motivation to resist threats.

We argue that leaders resist coercion — and sometimes retaliate against coercers (Dafoe, Hatz and Zhang, 2021) — partly because people have a psychological motivation to defy attempts to constrain their decisional autonomy. Given the prevalence of reactance in how people respond to coercion in their daily lives (Nisbet, Cooper and Garrett, 2015; Dillard and

Shen, 2005; Engs and Hanson, 1989; Laurin et al., 2013, 153), we doubt that leaders have insulated themselves from this psychological reality. Reactance theory plays an important role in American Politics research on political communication and persuasion (Peffley and Hurwitz, 2007; Nyhan and Reifler, 2010; Gadarian, 2014), but to our knowledge we are the first to employ it in coercion research. Although coercion scholars largely agree that coercion often fails, consensus about why remains elusive. Rational explanations account for relative power, the sizes of demands and threatened punishments, credibility of threats and assurances, and whether conceding endangers reputations with foreign and domestic audiences (e.g., Schelling, 1966; George and Simons, 1994; Slantchev, 2012; Mercer, 2010; Lupton, 2020; Fearon, 1994). Political psychologists add loss aversion, analogical reasoning, and emotions to the list, arguing that leaders take risks to maintain the status quo, avoid repeating history, or when coercers provoke anger (Berejikian, 2002; Khong, 1992; Hall, 2011; Markwica, 2018).

Reactance constitutes a potential missing piece that contributes more comprehensively to understanding coercion failure. It offers a compelling explanation that applies to a wide variety of threats, including circumstances where rational incentives favor capitulation. To be sure, demonstrating empirically that reactance provides a necessary addition to the catalog of existing explanations poses challenges, primarily because it is difficult to simultaneously rule out every alternative explanation. We meet these challenges with experiments. Experiments allow us to hold constant or manipulate existing explanations, thereby establishing that reactance still contributes to coercion failure. This research design offers a promising initial step in establishing reactance's potentially pervasive role in international politics.

Indeed, results from two novel experiments establish that reactance plays a causal role in coercion failure. Both experiments compare how participants respond to a coercive threat versus a "natural costs" counterpart that implicates the same choice and consequences without coercion. These experiments control for or manipulate other key factors, granting us inferential leverage over whether threats themselves promote resistance via reactance. Re-

sults from both studies strongly support a role for reactance. People exposed to coercive threats chose to defy their opponent more often than those facing the same choice from natural costs. Coercive threats also increased support for retaliating against coercers with sanctions and military strikes. Moreover, we find evidence that coercion affected policy preferences primarily by increasing anger and counterarguing — the emotional and cognitive mechanisms of reactance theory.

The Prevalence of Coercion Failure

The preponderance of coercion research concludes that compellence fails at a high rate, a track record that we review here before considering existing explanations for coercion failure. Following Sechser (2011, 379), we define 'compellence' as an explicit demand by one actor (the challenger) that another actor (the target) alter the status quo in some material way, backed by a threat to impose costs if the target does not comply.[1] Its counterpart, deterrence, differs only in that it tries to *sustain* the status quo (Schelling, 1966). These definitions place threat-making at the heart of coercion. Coercion failure occurs when the target does not comply with the demand, irrespective of whether the coercer then carries out the threat.

We focus on compellence failure for two reasons. First, we explain in the next section why we expect that psychological reactance bedevils nearly all compellent threats but only some deterrent threats. Second, although deterrence purportedly succeeds more frequently than compellence (Schelling, 1966), deterrence successes are famously difficult to gauge. Inaction by the target could signify that deterrence worked or merely that the target never intended aggression (Achen and Snidal, 1989, 161).

Across domains, methods, and types of actors, compellence elicits low rates of capitulation. Per the Militarized Compellent Threats dataset, 39 percent of 242 explicit threats from 1918 to 2001 obtained target compliance (Sechser, 2011). Research on specific compellence

[1]Following Schelling (1966), we amend Sechser's definition to include non-state actors and costs other than force such as sanctions.

types and issue areas often suggests lower success rates. For instance, Altman (2017) reports that states rarely acquire territory by making threats, instead seizing it by fait accompli.

Research investigating why the strong find it difficult to coerce the weak underscores that compellence often fails even with a favorable distribution of power (Mack, 1975; Arreguin-Toft, 2001; Sechser, 2010; Chamberlain, 2016). Sullivan (2012) concludes that the strong often fail to prevail against the weak because they pursue objectives that cannot be imposed and instead require coercion to achieve. Evidence also suggests that nuclear powers struggle to coerce non-nuclear states via "nuclear blackmail" (Betts, 2010; Sechser and Fuhrmann, 2017).

Aerial bombing, terrorism, economic sanctions, and cyber compellence also have poor track records. Aerial punishment rarely coerces regimes into granting major concessions (Pape, 1996; Horowitz and Reiter, 2001). Jones and Libicki (2008) find a 10% success rate for terrorist coercion — while doubting that terrorism caused many of those successes (see also Fortna, 2015). Scholars variously estimate economic sanctions' success rate at 27%, (Morgan, Bapat and Kobayashi, 2014), 34% (Hufbauer, Schott and Elliott, 1990), and 4% (Pape, 1997). Notably, the higher figures describe observed outcomes after sanctions, but sanctions could cause those outcomes or merely precede them (Nooruddin, 2002). Studies also suggest that cyber coercion seldom generates political concessions (Lindsay, 2013; Borghard and Lonergan, 2017).

Of course, coercion sometimes succeeds. In 1938, Czechoslovakia acceded to Hitler's ultimatum in Munich. Armed groups drove the United States from Lebanon in 1984. Economic sanctions contributed to Iran's 2015 nuclear concessions. Research suggests that suicide terrorism (Pape, 2003),[2] engineered migration (Greenhill, 2010), and demands for regime change (Downes, 2018) succeed more often than other threats.[3] Although these studies il-

[2]This finding is controversial. The data over-represent strong groups and count minor concessions like prisoner releases as successes (Krause, 2013).

[3]The regime change data include only 23 total cases and seven post-1945 cases; it may not be representative.

lustrate the absence of complete scholarly consensus, the preponderance of evidence affirms the prevalence of compellence failure.

Explaining coercion failure

Why do states find it so challenging to extract concessions via threats? Although most scholars agree that coercion often fails, explaining coercion failure remain an open and important question for understanding international politics. Compellence's tendency to fail across substantial variation in actors, domains, and punitive policies points to the importance of general explanations that apply across types of coercion. Traditionally, general explanations for why threats fail emerge from the core logic of coercion, including excessive demands, insufficiently costly punishments, too little power, a dearth of credibility, and the lack of credible assurances that compliance will avert punitive action (Schelling, 1966; George and Simons, 1994; Slantchev, 2011; Powell, 1990; Pauly, 2019). Other rational-actor explanations incorporate audiences. Leaders may refuse to submit because they fear that damaging their reputation would invite future aggression or that their political fortunes would suffer domestically (Schelling, 1966; Mercer, 2010; Press, 2005; Dafoe, Renshon and Huth, 2014; Lupton, 2020; Fearon, 1994; Snyder and Borghard, 2011).

However, the sweeping extent of coercion failure poses challenges for rational explanations and therefore provides grounds to investigate the possibility of a common psychological obstacle. For instance, coercion often fails under circumstances that favor capitulation per rationalist theories, such as when challengers are more powerful than targets (Sechser, 2010), make credible threats (Haun, 2015), and threaten severe costs (Pape, 1996). Such findings hint that a psychological factor contributes to coercion failure — offering a plausible alternative to rationalist explanations in some cases and, in other cases, explaining outcomes that contradict rationalist expectations. Still, these observations are suggestive, not conclusive. To simultaneously rule out these alternative explanations, we use experiments to control for

them, thereby isolating psychological reactance.[4]

Of course, IR scholars have developed psychological explanations for coercion failure. We aim to make three contributions to this research. First, our argument implies that threats themselves motivate resistance. Coercion operates at an inherent psychological disadvantage, diminishing the odds of success. In contrast, most existing psychological arguments emphasize explanatory variables that apply to a subset of coercion cases. For instance, scholars have explored how misperceptions, images, attribution errors, and flawed historical analogies undermine threats and cause wars (Jervis, 1976; Stein, 1992; Khong, 1992). The Munich analogy discourages concessions for fear of repeating the sin of appeasement. Prospect theory and loss aversion explain that targets stand firm because people take greater risks to avoid losses than to attain comparable gains (Berejikian, 2002). Individual traits might predispose certain leaders to stand firm or negotiate (e.g., Holmes and Yarhi-Milo, 2017). Reactance theory instead centers the common factor. Our study thus builds on Dafoe, Hatz and Zhang (2021)'s finding that violence provokes increased resolve and support for retaliation. Whereas Dafoe, Hatz and Zhang (2021) attribute this effect to combined reputational, honor, and psychological motives, our theory and experiments isolate psychological reactance from other causes of coercion failure.

Second, reactance theory entails both cognitive and emotional processes. Reactance is not reducible to either — complementing political psychology's growing emphasis on emotions while emphasizing the link between "hot" and "cold" cognition. Prior studies report that provoking anger leads states to favor imposing punishment over bargaining (Hall, 2011), and that anger, fear, hope, pride, or humiliation incline targets toward noncompliance (Markwica, 2018). Markwica proposes that when coercers anger their targets, targets commit to defiance until the anger subsides or other emotions take over. Reactance theory concurs. But as our

[4]Importantly, reactance also implicates cognitive processes that could motivate biased perceptions about credibility, costs, or benefits (Petty and Wegener, 1999; Silvia, 2006, 56-57), further hindering efforts to distinguish reactance from rationalist explanations outside experiments.

results will show, cognitive processes also play an essential part in the psychology of coercion failure.

Third, we use experiments to test reactance as a cause of coercion failure. This method complements case studies in previous psychological work on coercion. Experiments hold constant the many alternative explanations for coercion failure in a given scenario, facilitating causal inference.

Reactance as a Cause of Coercion Failure

Psychological reactance theory explains that people resist threats to preserve their sense of autonomy. The theory contains four key components. First, people value controlling their own thoughts and behavior (Brehm, 1966; Miron and Brehm, 2006), reflecting a psychological need for autonomy that emerges early in life. Second, telling someone how to think or what to do often leads people to perceive a threat to that valued freedom. Coercive demands undermine personal autonomy by dictating choices — they tell an actor what policy concession they must make to avoid consequences. Third, people dislike feeling manipulated, backed into a corner, or like an external actor controls their behavior. Psychologists call this aversive response to perceived freedom threats "reactance." Fourth, resolving this psychological discomfort requires re-establishing control: Resisting or retaliating against challengers cements one's status as an independent actor, whereas compliance risks confirming that an external force holds the reins. We propose that this pervasive motivation constitutes a powerful cause of coercion failure in international politics.

Reactance produces two key behavioral consequences germane to IR. First, reactance causes people to resist demands — reducing the likelihood that they will adjust their behavior to comply with threats. Telling someone that she *must* take a specific action increases the chance that she will refuse, even if she might otherwise prefer that option. Behavioral resistance affirms a target's free will, thereby reducing the unpleasant feeling associated with

losing autonomy (Dillard and Shen, 2005, 146). American politics scholars invoke reactance to explain why censorship promotes information search (Behrouzian et al., 2016) or why political persuasion sometimes backfires (Peffley and Hurwitz, 2007; Nyhan and Reifler, 2010; Matland and Murray, 2013; Nisbet, Cooper and Garrett, 2015; Gadarian, 2014). In IR, for example, psychological reactance could plausibly explain why Chinese demands that the U.S. stay out of claimed waters motivate the U.S. to send warships on 'Freedom of Navigation Operations.'

Second, reactance promotes retaliation against the source of a threat. Reactance motivates people to regain their sense of self-control, which they can accomplish through derogation and hostility directed at the challenger. Subject to heavy-handed warnings about the dangers of smoking, for example, people sling insults at the source, dismissing messengers as untrustworthy or lacking credibility (Silvia, 2006) and warnings as pointless and stupid (Hall et al., 2016). And reactance sometimes induces aggression against the actor who attempted to infringe on someone's freedom (Nezlek and Brehm, 1975), a dynamic that may explain why coercion sometimes provokes violent backlashes (LaFree, Dugan and Korte, 2009; Dafoe, Hatz and Zhang, 2021).

These behavioral tendencies suggest that reactance dooms many coercive threats even when material incentives, power asymmetries, credibility, or other situational factors favor capitulation. We hypothesize the following:

> H1: Threats will reduce support for taking the action demanded by the coercer, all else equal.
>
> H2: Threats will increase support for aggressive policies (such as imposing sanctions and using force) against the coercer, all else equal.

These hypotheses entail comparing a coercion scenario to a hypothetical counterfactual where an actor faces the same choice and incentives with one difference: the absence of a threat. We develop this point when we discuss our research design and 'natural costs' comparison

group. Notably, these hypotheses do not specify a baseline for how often coercion should succeed or fail because many parameters vary across scenarios.

Reactance mechanisms: anger and counterarguing

Reactance sets off a cascade of mental processes characterized by two intertwined mechanisms (Quick and Stephenson, 2007; Rains, 2013): an emotional component marked by anger and a cognitive component marked by counterarguing. First, reactance triggers "hostile and aggressive" feelings (Quick and Considine, 2008) — anger. An "approach" emotion, anger increases risk tolerance and motivates people to take actions that change the situation (Lerner and Keltner, 2000). For example, anger drove some Americans to support war against Iraq after 9/11 (Huddy and Feldman, 2011) and prompts leaders to fight back in response to hostile threats or provocation (Markwica, 2018; Hall, 2011; Hall and Ross, 2015). Psychological reactance complements this research by explaining that feeling coerced or manipulated suffices to cause anger.

Second, reactance prompts people to counterargue threats. That is, people actively generate thoughts that dispute aspects of the threat that favor compliance. Mustering these thoughts "[reduces] the persuasiveness and credibility of both message and source" (Nisbet, Cooper and Garrett, 2015, 42), providing a cognitive rationale to resist. Consequently, people doubt the threat's credibility or underestimate the costs of defiance (Silvia, 2006; Gadarian, 2014). For instance, smokers respond to graphic warnings by insisting that authorities exaggerate the health risks (Hall et al., 2016). People can further counterargue threats by overstating the benefits from resistance: Threatening to take something away increases its subjective value (Miron and Brehm, 2006). Just as banning a book can amplify its popularity, leaders might place greater value peripheral territory after a rival demands it. Motivated to pursue a valued goal (Petty and Wegener, 1999, 56-57) — retaining autonomy — people produce reasons to resist.

Psychological reactance complements other IR research that theorizes roles for both coun-

terarguing and anger: For example, counterarguing partly explains why people reject adversaries' peace proposals — they undervalue the negotiated outcome that an adversary thrust upon them ("reactive devaluation") (Maoz et al., 2002). Similarly, Gadarian (2014) finds that Democrats dismissed heavy-handed messages about terrorist threats from the Bush administration. Per Markwica (2018), angry leaders downplay the risks associated with non-compliance — especially when coercers invalidate their identity. Reactance explains how and why threats create these cognitive and emotional tendencies in the first place, linking psychology to the prevalence of coercion failure.

These emotional and cognitive pathways interweave — reactance is an "amalgam of anger and counterarguing" and neither has causal primacy (Rains, 2013, 49; Dillard and Shen, 2005). We therefore expect that coercive threats will a) increase anger against adversaries and b) increase counterarguing by leading participants to diminish the perceived consequences of resistance and increasingly value the prize. Both dynamics will mediate the relationship between threats and resistance/retaliation.

Reactance conditions

We expect that reactance impedes coercion under three conditions. First, a decision-maker must believe that they hold decisional autonomy within a specific domain (Brehm, 1966). Notably, even low stakes issues provoke reactance in psychology research (Dillard and Shen, 2005) — the key antecedent is whether the leader values their policy-making autonomy. Reactance lessens when the decision-maker believes that another actor holds legitimate authority over the decision. Although we doubt that leaders of sovereign states perceive coercion by even high-status peers as legitimate, government officers acting domestically or international organizations may hold sufficient perceived authority over some issues to weaken reactance.

Second, psychological reactance requires a stimulus. It arises when another actor directly acts to impinge on the decision-maker's autonomy (Miron and Brehm, 2006; Hall et al., 2016). For example, we expect that threatening sanctions while demanding that a state end

its nuclear program will induce reactance. But if the mere prospect of economic isolation deters proliferation — without any actual threat of sanctions — then reactance should not manifest. Miller (2014) reports the "secret success" of nonproliferation sanctions in just this way. Generally, reactance should apply in cases of overt coercion, perhaps suggesting better prospects for latent coercion than overt threats.

Consequently, we expect that reactance impedes virtually all compellent threats but only some deterrent threats.[5] Compellence demands changes to the status quo that infringe on targets' decisional autonomy (Schelling, 1966; Sechser, 2011). Furthermore, compellence generally requires coercers to take overt actions likely to arouse reactance, such as issuing explicit threats and aggressively signaling resolve. Deterrence sometimes requires similar policies, but is often latent. For instance, we expect reactance against a demand not to invade an ally, but a mutual defense pact signed years previously could deter without generating present-day reactance.

Third, forceful, dogmatic language intensifies reactance (Quick and Considine, 2008; Gadarian, 2014). Message features amplify or mitigate reactance by altering the degree to which a target perceives it as threatening their autonomy (Dillard and Shen, 2005, 148). Dogmatic language entails explicit demands that a target *must* abide an appeal (Quick and Stephenson, 2007). People perceive demands as more autonomy-threatening than requests that they *consider* behavioral change, which prompt less reactance and resistance. Importantly, however, formulating a threat that preserves the target's sense of autonomy is likely difficult in interstate coercion. Mitigating reactance requires that decision-makers view complying as them choosing to abide a request rather than conceding to pressure cloaked in diplomatic language. Study 2 explores empirically whether policymakers might reframe threats as requests to mitigate reactance, but taken together these three conditions suggest that reactance will afflict most overt coercive threats in international politics.

[5]This could help explain Schelling's (1966) observation that deterrence succeeds more often than compellence.

Research Design: A Natural Costs Approach

We conducted two survey experiments to test whether reactance provides an important psychological cause of resisting threats. We designed these experiments to overcome a significant inferential challenge: coercive threats lack clear counterfactuals. We address this obstacle by randomly assigning participants to receive either a threat or a *natural costs* alternative that presents the same situation and choice without coercion.

Consider the following: Suppose that we want to know how people respond when threatened with coercive bombing akin to NATO's 1999 intervention against Serbia. A straightforward coercion treatment informs participants that another actor threatened to bomb them unless they relent. But what constitutes the equivalent to a bombing threat that allows participants to *choose* whether to capitulate or resist, with the same costs, benefits, and credibility — but *without coercion*? "Nature" could drop bombs or otherwise inflict comparable damage after a natural disaster, but that strains credulity for most coercion scenarios. Worse still, for the decision environment to remain the same, actors in Serbia's position must believe that the natural disaster would only occur if they continue operations in Kosovo but not otherwise. We do not know of any natural disasters that are so discerning.

To overcome this challenge, we design vignettes with two characteristics. First, the costs (threatened punishments) could occur either due to deliberate coercion or a realistic but impersonal *natural costs* mechanism. Crucially for our theory, the same choice set has different implications for reactance. Someone must perceive a threat as a deliberate attempt to restrict their options *by another actor* for it to arouse reactance. Early reactance research used a similar strategy, comparing "personal" events perceived as attempts to constrain freedom to "impersonal" events that incidentally limited choices.[6] Second, the actor (coercive target) decides whether to accept a loss (capitulation) to avoid those costs. This differentiates

[6]For example, participants expecting to select a prize among five options were more hostile when a researcher removed one option than when told it was lost in shipment (Cherulnik and Citrin, 1974).

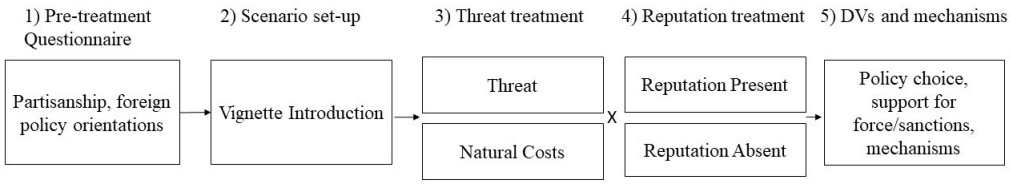
our approach from recent work that assesses responses to aggressive acts (Dafoe, Hatz and Zhang, 2021) by allowing us to directly incorporate threats and concessions.[7] Few coercion scenarios allow both, but we developed and implemented two experiments meeting those criteria.

Study 1: A Disputed Island in a Storm

We fielded a pre-registered survey experiment to a sample of 1,442 U.S. participants via Prolific Academic in February 2021.[8] Our primary interest lies in comparing the effects of coercive threats to natural costs alternatives. We control for alternative explanations such as costs, benefits, and power. Disentangling psychological reactance from reputation and audience costs poses a challenge because it is plausible that threats also automatically arouse concerns about costs from visibly backing down. We therefore include a second manipulation that directly raises or minimizes reputation and audience considerations together, isolating reactance from reputational dynamics and allowing us to compare their effects. These manipulations produce a 2 (*natural costs/coercion*) x 2 (*reputation absent/reputation present*) between-subjects experiment.

Methodology

Figure 1: Experiment 1 Design



[7]Our study further differs from Dafoe, Hatz and Zhang (2021) by isolating psychological reactance from reputation and honor motives, especially where audience perceptions matter.

[8]Pre-registration available on OSF: https://osf.io/q6jgd. Analyses follow our pre-analysis plan unless explicitly noted. See Appendix §A.1 for sample characteristics.

The experiment proceeded in 5 steps, summarized in Figure 1. After completing a pre-treatment questionnaire, participants read about a scenario that might take place in the future but does not represent any particular countries or conflicts (Tomz and Weeks, 2020; Kertzer, Renshon and Yarhi-Milo, 2019). All participants read the introductory vignette in Table 1. Importantly, using fictional countries increases control over factors like relative power without compromising inferences about treatment effects (Brutger et al., 2021). Although hypothetical, this scenario resembles a potential crisis in the South China Sea and has historical precedents; states seized islands in peacetime 28 times since 1918 (Altman, 2020).

Participants randomly assigned the *coercion* treatment read that a Navalian submarine captain threatened to sink their ship if it continued toward the island.[9] In the *natural costs* groups, participants read that returning to the island through the typhoon risked sinking the ship.[10] All participants received the same information about the risks from continuing, thereby accounting for the consequences of defiance, before random assignment to the *reputation absent* or *reputation present* treatments.

Several manipulation checks validate the treatments and vignette. Analyses in the Appendix §A.2 show that the threat treatments raised reactance (but not concerns about reputation costs) and the reputation treatments raised reputation concerns (but not reactance). We also present important evidence for information equivalence, showing that the treatments do not systematically alter perceptions of relative power nor the stakes of the crisis.

Participants then indicated whether they would order the ship to turn back (capitulate) or continue to the island (resist), and reported their confidence in that choice. We combined these branched responses to create a continuous scale where higher values indicate confidence in the decision to continue. We measured support for retaliation via a) economic sanctions or b) going to war against Navalia on 5-point scales. We measure anger via closed ended

[9]Specifying a submarine forecloses options like boarding, shooting at the rudder, or having soldiers on the supply ship shoot first.

[10]This storm draws inspiration from Dafoe, Hatz and Zhang (2021).

Table 1: Vignette and Experimental Manipulations

All: Introduction	• Your country is involved in a dispute with the country of Navalia over a small island. Both your country and Navalia claim to own the island, which has valuable resources. • Both countries are equally powerful. • Your country is at peace with Navalia and has never fought a war with Navalia. Your country has no other disagreements with Navalia. • Your country has stationed soldiers on the island for many years. They were withdrawn temporarily due to an incoming typhoon. • You recently ordered a supply ship to return your country's troops to the island. • However, a spy for your country just learned that Navalia is sending a military force to seize the island before your troops can return. • Their orders are to seize the island if they arrive first but to turn back if your troops have already returned to defend it. • The island is easy to defend, so it would be very difficult for you to take it back after Navalian troops arrive.
Randomized: Threat treatment	• To reach the island before the Navalian forces, your supply ship must travel **through the typhoon.** (*natural costs*)
	• To reach the island before the Navalian forces, your supply ship must travel **past a Navalian submarine**. The captain of that submarine just got on the radio and **demanded that your ship turn back.** • He threatened you, **"You must turn back now, or we will fire and you will be destroyed!"** (*coercion*)
All: Risks	• If you order your ship to continue on, your best guess is that it has a 50% chance of surviving to reach the island and a 50% chance of being sunk. • If the ship sinks, **all 100 soldiers on board will die**.
Randomized: Reputation treatment	• Neither your public nor other countries will know what has happened, so your reputation for standing firm is not at stake. (*reputation absent*)
	• Your public and other countries will know what has happened, so your reputation for standing firm is at stake. (*reputation present*)

questions about feelings toward Navalia (anger scale $\alpha = 0.91$) (Quick and Stephenson, 2007; Gadarian and Albertson, 2014). To assess counterarguing, separate items ask participants whether "the threat to my soldiers from continuing to the island is exaggerated" (Hall et al., 2016), whether the island is valuable, and whether to prioritize protecting soldiers' lives. Finally, alongside questions about relative power and costs, participants estimated the chances that turning back would cause their country's reputation to suffer or the public to disapprove of their leadership (Tomz and Weeks, 2020).

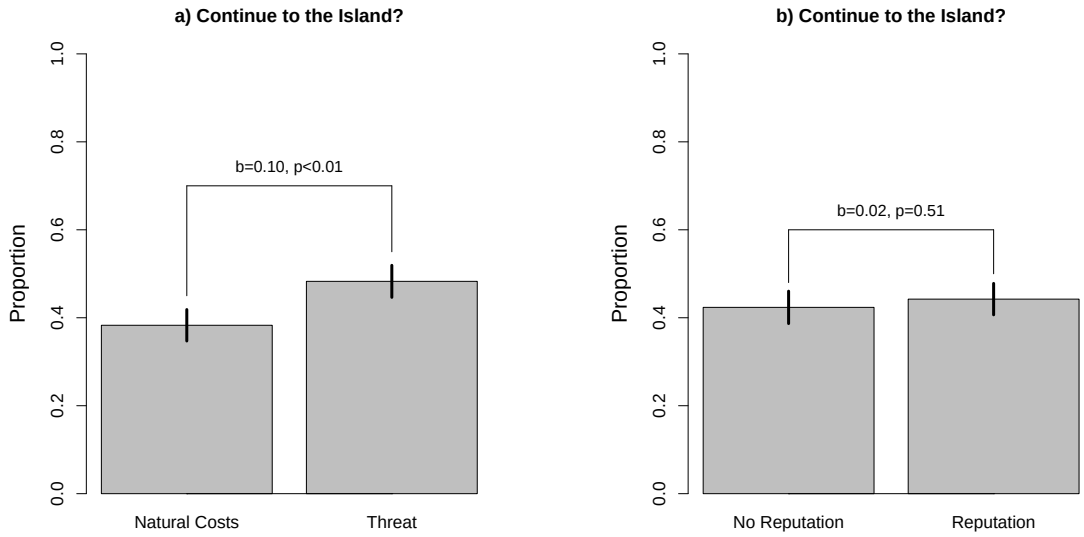
Threats increased resistance and support for conflict with Navalia

Evaluating hypotheses 1 and 2 requires comparing policy preferences between participants in the *coercion* group and those in the *natural costs* group. Figure 2 displays the proportion of participants in each group who chose to continue to the island — defying the threat — versus those who turned back. The results in panel (a) provide striking evidence for reactance as a cause of coercion failure: When participants received a direct threat, 48.3% chose to continue and attempt to retain their island. Random assignment to the natural costs group decreased the rate of resistance by 10 percentage points ($p < 0.01$). Faced with a typhoon that might kill their soldiers, most participants retreated, relinquishing the territory. But after a direct threat from Navalia's submarine captain, more participants risked their soldiers' lives to retain their territory.

To our surprise, panel (b) shows that raising reputation and audience costs had no statistically significant effect on capitulation rates ($p = 0.51$). Indeed, the threat's effect is over five times the size of the reputation effect — a striking difference given reputation and audience costs' centrality in explanations for coercion failure. Crucially, reputation costs did little to dampen the threat's effect — the threat increased defiance by 8.4 percentage points in the reputation absent group, and 11.6 percentage points in the reputation present group (interaction $b = 0.03$, $p = 0.52$).

Reactance also implies that people will retaliate against actors who threaten their au-

Figure 2: Coercion Increased the Proportion Continuing to the Island



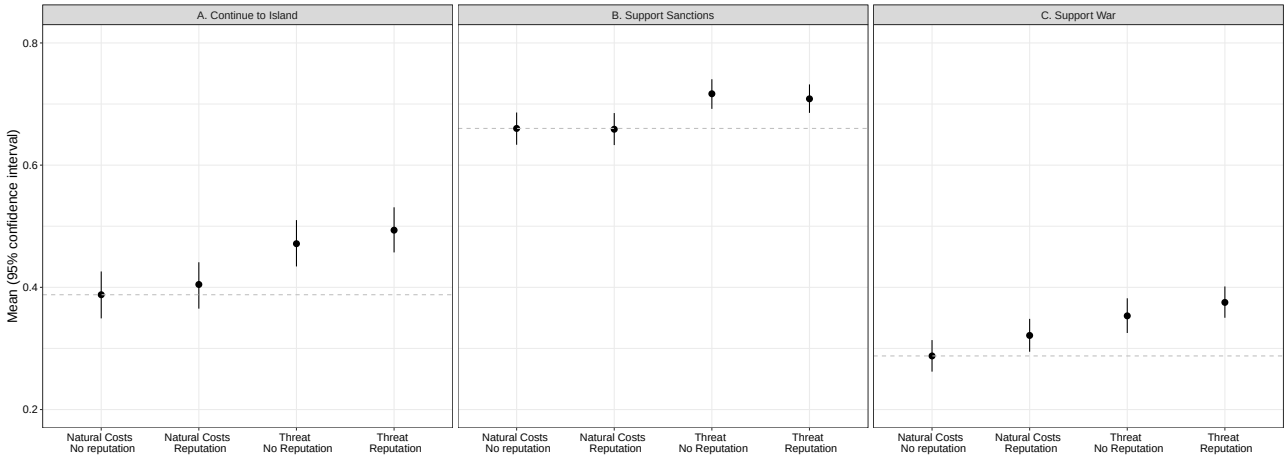
tonomy (H2). Panels **B** and **C** in Figure 3 support this expectation. In comparison to natural costs, the coercion treatment increased support for sanctions by 5.3 percentage points ($p < 0.01$) and support for going to war by 5.8 percentage points ($p < 0.01$). We did find evidence that raising reputation costs increases support for war ($b = 0.03$, $p < 0.05$). However, the difference within the natural costs group drives this effect. Combining reputation concerns with a threat did not significantly increase support for war compared to a threat with minimized reputation considerations ($p = 0.25$).

Do threats affect anger and counter-arguing?

The main effects provide powerful support for our argument: the threat increased resistance and support for retaliation. If reactance causes coercion failure, then we should also find evidence that the threat treatment increases anger and counterarguing. In turn, anger and counterarguing should mediate the effect of the threat treatment on the dependent variables.

Figure 4 plots the direct effects of our treatments on six mechanism questions that capture

Figure 3: Response to Threat by Treatment



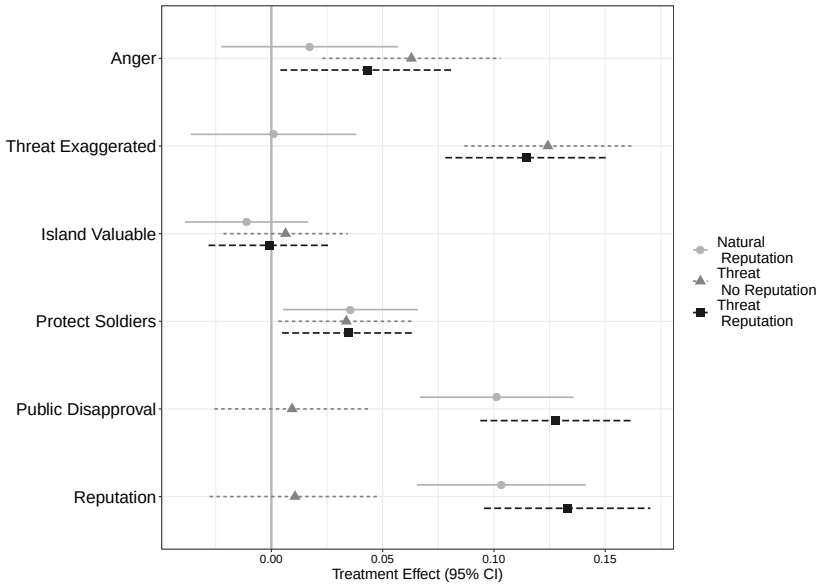
Note: Dashed horizontal lines display mean in the natural costs/no reputation group.

reactance (anger, threat exaggerated, island valuable, protect soldiers) alongside reputation and audience cost concerns, relative to the *natural costs, no reputation* group. The results support three important conclusions. First, coercive threats increase anger. Both scenarios depicted an adversary moving to capture the participant's territory. But the personal and identifiable threat aroused more anger, consistent with reactance.

Second, coercion increased counterarguing via beliefs that the threat was exaggerated. Although both the natural costs and threat treatments contained identical risk information — a 50% chance that the ship would sink — people exposed to the threat minimized the dangers ($b = 0.12$ and $b = 0.11$ when reputation absent/present; both $p < 0.01$). We might expect the opposite effect if people perceived Navalia's threat as a costly signal of resolve. Although counterarguing often leads people to overvalue prohibited actions, we find no evidence that any treatments affected estimates about the island's value. Contrary to our expectations, all treatments increased the value that participants placed on protecting soldiers' lives ($p < 0.05$).[11]

[11]Contrary to our pre-registered interpretation, we speculate that placing more value on soldiers' lives is consistent with reactance because the threat placed both the island and troops at risk.

Figure 4: Direct Effects on Mechanisms



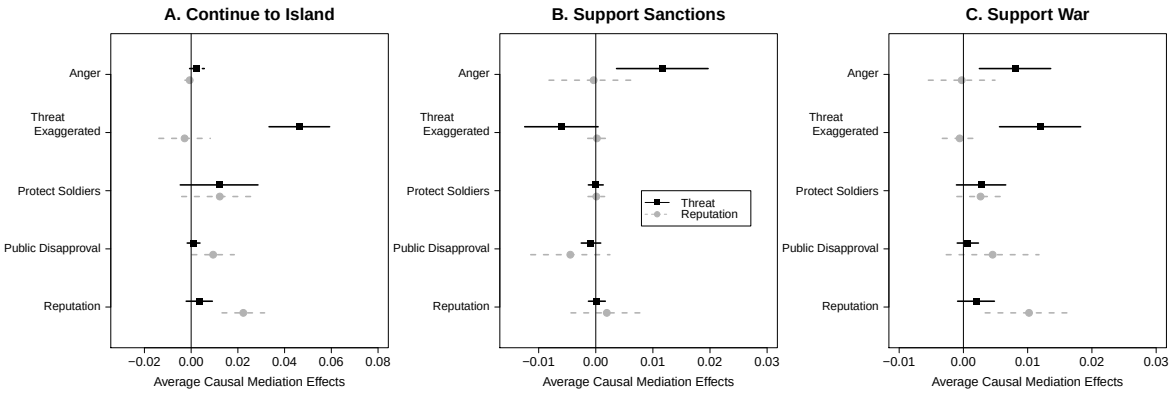
Note: Average effect relative to natural costs/no reputation (95% confidence intervals). Models include pre-treatment controls.

Third, comparing the *threat/no reputation* treatment to the *natural costs/no reputation* presents a crucial test for our theory. If the threat increases anger and counterarguing when we minimize reputation — but does not increase reputation or public disapproval concerns — we have strong evidence for reactance as a cause of coercion failure. And indeed, results depicted by triangles in Figure 4 clearly confirm that the threat raised reactance without implicitly raising reputation's salience. The results allow us to rule out reputation as the primary mechanism behind coercion failure in our experiment, since the reputation treatment raised reputation costs without augmenting the main effect (Bullock and Green, 2021).

Next, we estimate the causal pathways that connect threats to participants' policy choices via causal mediation analysis. Mediation models estimate the proportion of the main effect that flows through each mechanism. Standard mediation models assume causal independence between multiple mechanisms, but diagnostic tests suggest significant dependence among the 5 mechanisms implicated in our analyses. Panels **A-C** in Figure 5 estimates for average causal mediation effects (ACMEs) using non-parametric models that account for causal dependence

between multiple mediators (Imai and Yamamoto, 2013).[12]

Figure 5: Mediation Estimates



Note: Lines depict 95% confidence intervals. Models include pre-treatment controls.

The results support our contention that reactance plays a role in coercion failure, with the interesting wrinkle that the emotional and cognitive mechanisms unevenly contributed to defiance and retaliation. Anger does not mediate the threat's effect on continuing to the island but accounts for 23.11% of its effect on support for sanctions and 12.92% of its effect on support for war — anger plays a substantial role in support for retaliation.

By contrast, counterarguing by insisting that the threat has been exaggerated explains a greater (52.16%) portion of the threat's effect on continuing to the island. It also accounts for 19.19% of the threat's effect on support for war, complementing anger. Contrary to our expectations, we find a negative indirect effect for underestimating the threat on support for sanctions, though the 95% confidence interval contains 0.

Figure 5 provides some evidence that concerns about reputation and audience costs played an indirect role in coercion failure when the vignette explicitly raised these considerations. The findings suggest that — to the extent that the reputation treatment raises reputation

[12]Plots depict the average ACME between treated and control group. We excluded perceptions of the island's value due to the absence of direct effects. Appendix §A.4 reveals similar results when we include this variable. These models deviate from our pre-registered product-of-coefficients approach (see Appendix §A.4) but present more conservative tests.

costs — rates of capitulation decline and support for war increases. The reputation treatment's effect on resistance also flowed partly through concerns that turning back would cause public disapproval, though this does not mediate the effects on support for retaliation.

These analyses show that the threat treatment affected psychological reactance mechanisms without implicitly raising reputation and audience cost considerations. Estimates suggest that anger and counterarguing mediate substantial portions of the threat's causal effect. Notably, our research design randomly assigns the threat treatment but measures anger and counterarguing, and confounding from a potential treatment-mediator interaction could bias the ACME estimates (Imai and Yamamoto, 2013).[13] We interpret the results as consistent with our theory and important suggestive evidence supporting reactance, but precise causal evidence for mechanisms requires additional research (Bullock and Green, 2021).

Study 2: Arms Sales and Terrorist Bases

We fielded study 2 over two weeks in August 2020 to an effective sample of 3,526 U.S. adults recruited via Lucid Theorem.[14] Study 2 builds on study 1 in three ways. First, external validity requires replication, so we shift from hypothetical to real states and from a symmetric power distribution to extreme asymmetry. Second, drawing from research showing that dogmatic language increases reactance (Gadarian, 2014), we evaluate whether subtler, less bellicose threats weaken reactance and mitigate coercion's propensity to inspire defiance. Third, we designed study 2 to improve (modestly) on study 1 by further minimizing participant concerns about reputation and domestic audience costs.

[13] See the Appendix §A.4 for sensitivity analyses.

[14] Effective sample excludes participants who spent <2.1 seconds reading the pre-treatment vignette ($< 4\%$) and terminates people who failed an instructional manipulation check on page 2 of the survey. See Appendix §B.1 for details and sample characteristics.

Methodology

We adopt the same experimental structure in study 2. We first instruct participants to imagine being the U.S. Ambassador to Eritrea and responsible for making policy decisions. The vignette describes U.S. interests in removing terrorist bases in Eritrea while simultaneously inhibiting its dictator from massacring civilians via an arms embargo.[15] Participants must choose whether to end the embargo (capitulate) and gain cooperation from the Eritrean government to destroy the bases — reducing the risk of a homeland terrorist attack — or continue to block arms sales (resist) thereby leaving the U.S. vulnerable. Participants receive a *strong threat*, *weak threat*, *natural costs control*, or *baseline control* via random assignment, again controlling for other factors. Although we include a *baseline control* group to assess preferences about blocking arms sales, our primary interest lies in comparing threats to natural costs.

[15]Foreign sanctuaries provide common sources of resiliency for violent nonstate actors (Salehyan, 2009).

Table 2: Vignette and Experimental Manipulations

All: Intro	• A terrorist group affiliated with ISIS has sworn to attack the United States. • This terrorist group is setting up bases in the country of Eritrea. • Because U.S. intelligence agencies have been unable discover the exact locations of these bases, you need the cooperation of the Eritrean government to remove them. • Eritrea is ruled by an oppressive dictator whose secret police have killed hundreds of innocent civilians in towns seen as less loyal to him. • Consequently, you (the United States) have been blocking all sales of weapons to Eritrea. • If the dictator had these weapons, he would use them to massacre thousands of civilians.
Randomized: Treatment	• In a private meeting earlier today, **the dictator threatened you, "You must immediately stop meddling with Eritrea! If you do not allow us to buy the weapons, I will let the terrorist group operate freely."** (*strong threat*)
	• In a private meeting earlier today, **the dictator appealed to you: "I hope that you consider allowing us to buy the weapons. It's your choice. But if you do not, I will have to let this terrorist group operate freely."** (*weak threat*)
	• Although the dictator hasn't said anything to you, you know that **his government needs the weapons to be able to destroy the terrorist group's bases.** (*natural costs*)
	• [no additional information] (*baseline control*)
All: Risks	• If you change U.S. policy to allow Eritrea to buy weapons, the CIA is confident that the Eritrean government would then remove the bases. • Allowing weapons sales to Eritrea would happen secretly. It would not receive attention from the media or other countries. • The CIA assesses that future conflicts with Eritrea are unlikely. • If the bases are not destroyed, the CIA assesses that the terrorist group will attack the United States in the next year, **likely killing dozens of Americans.**

All participants receive information minimizing reputation, including that the American public and foreign third parties would not know of concessions. We believe that secrecy is modestly more plausible in this scenario. Unable to eliminate the Eritrean government's knowledge of concessions, we minimized its importance. First, we described future conflicts with Eritrea as unlikely using stronger language than was possible for study 1 (because territorial disputes tend to persist). Second, we minimized the importance of maintaining a reputation with the coercer by selecting a small, distant country without a history of confronting the United States. We account for residual reputation concerns empirically.

After the vignette, we asked participants whether the United States should "continue to block weapons sales to Eritrea" (resist) or "allow weapons sales to Eritrea" (capitulate) and how strongly they supported that policy. Using a 5-point scale, we assessed support for economic sanctions, military strikes, and invading Eritrea. We again probed anger and counterarguing post-treatment. We measure the latter with participant estimates about the probability that allowing arms sales would lead to terrorist attacks in the U.S. and the killings of thousands of Eritrean civilians. Participants also assessed the probability of public disapproval, harming the U.S. reputation for resolve, and harming the U.S. reputation for moral authority.[16]

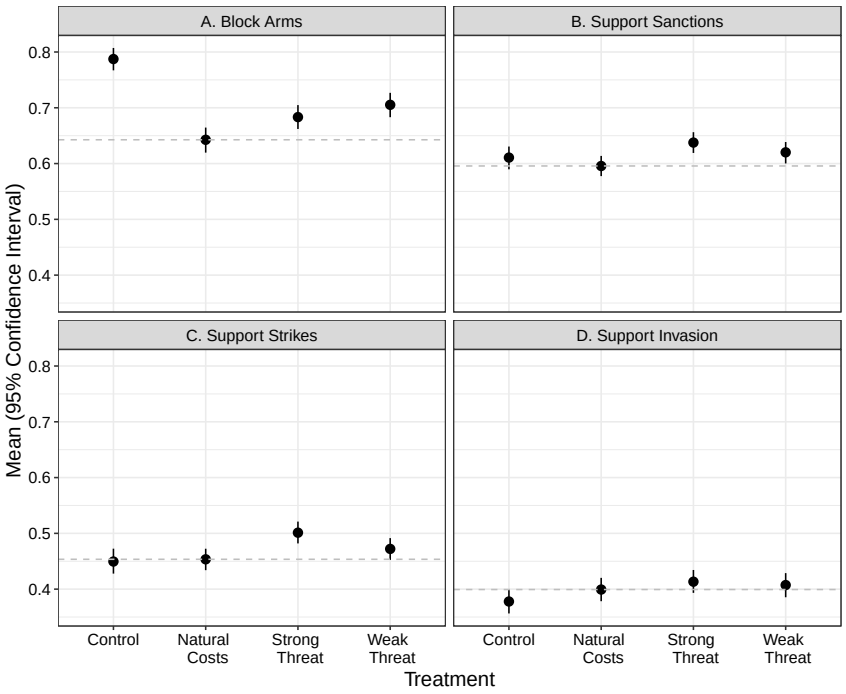
Assessing strong and weak threats

The treatment effects reinforce study 1's findings. Both threats increased the proportion of participants who elected to continue blocking arms sales, relative to natural costs: In the natural costs group, 32.6% of participants decided to end the embargo. The strong and weak threat increased that proportion by 4.9 ($p < 0.05$) and 7.6 percentage points ($p < 0.01$), respectively.[17] Consistent with psychological reactance, more participants preferred to maintain the embargo when Eritrea demanded that they end it. Panel **A** in Figure 6 depicts

[16]See the Appendix §B.2 for question wording.
[17]See Appendix §B.5

these treatment effects using the continuous outcome. The baseline control group lacked any incentives to end the arms embargo; participants accordingly preferred to continue it.

Figure 6: Coercion, Resistance, and Retaliation



Panels **B** and **C** in Figure 6 show that threats significantly increased support for retaliation. Compared to the natural costs condition, the strong threat increased support for imposing sanctions by 4.2 percentage points ($p < 0.01$) and support for striking Eritrean bases by 4.8 percentage points ($p < 0.01$). Nevertheless, panel **D** shows that participants across groups exhibited restraint on invading Eritrea to take control of the country. t.

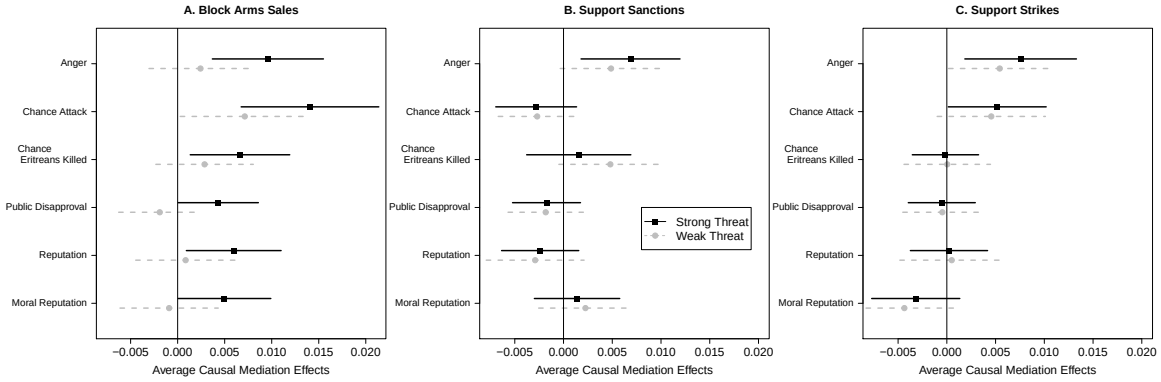
Reactance research argues that weaker, suggestive threats may provoke less backlash than those using dogmatic language (Gadarian, 2014). Indeed, a manipulation check confirmed that the strong threat induced greater reactance than the weak threat ($b = 0.06$, $p < 0.01$). Although both groups exhibited a similar tendency to resist Eritrea's demands (panel **A**), the weak threat had smaller and less consistent effects on retaliation compared to its dogmatic counterpart. The weak threat slightly increased support for sanctions ($b = 0.024$, $p = 0.08$) compared to natural costs, but not military strikes ($b = 0.018$, $p = 0.2$).

Collectively, these results suggest that any threat-to-choice language causes defiance, but support for aggression may depend on whether the coercer uses dogmatic language. We nevertheless urge caution in interpreting differences between the weak and strong threat treatments — although the 0.03-unit difference in support for strikes between the strong and weak threat groups corresponds to a low p-value ($p = 0.04$), the 0.02-unit difference in support for sanctions exceeds standard thresholds for statistical significance ($p = 0.19$).

The pathways from coercion to resistance

The direct effects on each mechanism provide suggestive support for our argument. Both threats again increased anger and heightened beliefs that arms sales would encourage a terrorist attack, though only the weak threat increased concerns that allowing arms sales would harm Eritrean civilians. Yet as we discuss in the Appendix §B.6, the threats also raised both moral and resolve-based reputation considerations.

Figure 7: Mechanisms of Coercion, Resistance, and Retaliation



Note: Lines depict 95% bootstrapped confidence intervals. Models include pre-treatment controls.

Figure 7 again presents estimates from causal mediation models that accommodate dependence between mechanisms (Imai and Yamamoto, 2013). The results suggest that the strong threat encouraged participants to reject Eritrea's demands and provoked retaliation

by prompting psychological reactance. Anger mediates the strong threat's effect on each outcome. Anger accounts for 20.64% of the strong threat's effect on support for blocking arms, 14.02% of the effect on sanctions support and 14.79% of the effect on support for strikes. The weak threat appears to run through a similar emotional pathway, though the estimates imply more statistical uncertainty: The weak threat's indirect pathway through anger explains 21.77% ($p < 0.1$) and 28.96% ($p < 0.05$) of its effects on sanctions and strikes, respectively, but we find a statistically insignificant ACME on resistance.

We find complementary evidence for counterarguing. Results suggest that threats caused resistance partly by encouraging participants to discount the benefits from conceding to the dictator. Both threats increased evaluations of the chance that allowing arms sales would lead to a terrorist attack (*chance attack*) ($b = 0.04$ and $b = 0.04$, both $p < 0.01$), and this factor accounts for 30.26% of and 10.59% of the strong and weak threat's effects on support for blocking arms (both $p < 0.05$). Panels **B** and **C** suggest that the strong threat increased support for retaliation via strikes when participants dismissed assurances about destroying the terrorists' capacity to attack the United States ($ACME = 0.005$, $p < 0.05$), but we did not find the same for sanctions ($ACME = -0.003$ $[-0.007, 0.001]$). Both threats also heightened perceptions that capitulating would increase the probability of civilian deaths in Eritrea (*chance Eritreans killed*; $b_{strong} = 0.02$, $p = 0.07$; $b_{weak} = 0.04$, $p < 0.01$), which partially mediated the strong threat's effect on resisting the dictator's demands ($ACME = 0.007, p < 0.05$). Given that this question highlights civilian lives, it is perhaps unsurprising that this mechanism does not contribute to support for military strikes.

The results reveal some support for alternative mediators as contributors to the strong threat's effect on defiance. Concerns that backing down would damage the U.S. reputation for resolve or moral authority accounted for 12.83 and 10.71% of the choice to block arms sales (both $p < 0.05$), but as the confidence bands depict in Figure 7, neither reputation significantly mediated the effect of the threats on retaliation. These mediation analyses supplement the main effects to provide additional suggestive evidence that the strong threat

— and, to a lesser extent, the weak threat — cause coercion failure via reactance.

Generalizability

Establishing the tendency for individuals to resist threats takes a necessary and important step toward applying reactance to international politics. Moreover, we establish some external validity by replicating our findings in two distinct IR scenarios.

Nonetheless, applying experimental results from the general population to foreign policy elites raises questions about "elite exceptionalism" — that elites "employ fundamentally different cognitive architectures" (Kertzer, Renshon and Yarhi-Milo, 2019, 19) — perhaps making them less susceptible to autonomy threats. If education trains elites to privilege rational calculations, for example, we would expect political knowledge, university education, or high income to weaken the effect of threats on defiance and aggression. Supplementary analyses reveal no consistent evidence that proxies for eliteness moderate the treatment effects. Of 54 interaction coefficients, only five meet standard statistical significance thresholds, and the direction varies: university education weakens the strong threat's effect on defiance in study 2 ($b = -0.08$, $p < 0.05$) and sanctions in study 1 ($b = -0.06$, $p < 0.05$), but knowledge increases both threats' effects on war support in study 2 (both $p < 0.05$). The overwhelming absence of differences conforms with mounting evidence that elites respond to experimental treatments much like ordinary citizens (Yarhi-Milo, 2018; Kertzer, 2020).

Another possibility is that higher stakes in real world decisions alter the dynamics of threat responses. However, we conjecture that reactance — and particularly its emotional component — grows stronger when actors feel personally coerced rather than vicariously coerced in hypothetical scenarios. Finally, we might expect cross-national or cultural variation in reactance that our U.S. sample cannot capture, a point we consider below.

Conclusion

Our results suggest that coercion operates at an inherent disadvantage in international politics. The act of coercing creates a psychological motivation to resist. We propose that reactance helps to explain why coercion fails in many cases and why even powerful actors find it difficult to employ successfully in others.

Our experiments exploited distinctive, unusual circumstances where plausible "natural costs" mimic coercive threats, allowing us to isolate psychological reactance from other contributing factors in IR settings. Despite large differences between the vignettes, both experiments yielded similar evidence that reactance causes coercion failure and that anger and counterarguing mediate the effects of threats. Interestingly, both studies suggest that anger played an especially important role in support for retaliation, whereas counterarguing more powerfully mediated coercion's effects on resistance.

For both scholars and practitioners, reactance is a novel explanation for why coercion so often fails. For instance, it is a plausible reason why punitive bombing so rarely succeeds (Pape, 1996), why states so rarely acquire even small pieces of territory via threats (Altman, 2017), and why even threats from nuclear powers often fail (Sechser and Fuhrmann, 2017). It can help explain why policymakers do — or why they should — attempt alternatives to coercion, whether positive inducements (Nincic, 2011) or forceful imposition. It offers insight into the obstacles policymakers must overcome – including anger and counterarguing – when they attempt coercion.

In turn, future research might examine what factors lessen reactance to inform policymakers seeking to calibrate threats and scholars examining variation in coercion outcomes. First, do threats using less controlling language or autonomy-affirming clauses succeed more often (Quick and Considine, 2008)? Study 2 did not detect a significant difference between strong and weak threats on the choice to capitulate, though we found some evidence that cloaking a threat in diplomatic language reduces support for retaliation. Other message-specific manipulations may yield different results. Second, cultural and identity variables

merit examination: In collectivist cultures, for example, threats from outsiders produce more reactance than threats from insiders (e.g., Graupmann et al., 2012), whereas elsewhere even in-group members provoke resistance (Matland and Murray, 2013). Third, leader characteristics might moderate reactance and its effect on coercion failure. Psychologists find that "trait reactance" predisposes some people toward cross-situational resistance (e.g., Cherulnik and Citrin, 1974; Dillard and Shen, 2005). Although exploratory analyses reveal no systematic interactions between threat treatments and traits like gender, partisanship, or U.S. region, measuring personality traits could yield insights into which targets are likelier to capitulate. Fourth, does reactance depend on the salience of a freedom, akin to smokers reacting more strongly against health warnings (Hall et al., 2016, 737)? Or do leaders value their sovereign freedom to make foreign policy choices across issues? Although we find similar evidence for reactance in two distinct scenarios, future research might manipulate issue salience and threat language while examining important cultural and dispositional variables.

Finally, we hope that future research will investigate reactance as a cause of war. Reactance, after all, makes it harder for states to reach the coercive bargains that are so important for avoiding wars (Fearon, 1995). Consequently, one side will wrongly believe that pressure can cow the other into granting concessions, only to instead encounter resistance and sometimes retaliation. Alternatively, challengers may forgo attempts at coercive bargaining that could have averted war, instead pursuing their objectives on the battlefield because they anticipated reactance. Either way, it stands to reason that reactance would contribute to causing wars.

References

Achen, Christopher H and Duncan Snidal. 1989. "Rational deterrence theory and comparative case studies." *World Politics* pp. 143–169.

Altman, Dan. 2017. "By Fait Accompli, Not Coercion: How States Wrest Territory from Their Adversaries." *International Studies Quarterly* 61(4):881–891.

Altman, Dan. 2020. "The Evolution of Territorial Conquest after 1945 and the Limits of the Territorial Integrity Norm." *International Organization* 74(3):490–522.

Arreguin-Toft, Ivan. 2001. "How the Weak Win Wars: A Theory of Asymmetric Conflict." *International security* 26(1):93–128.

Art, Robert J and Kelly M Greenhill. 2018. "The Power and Limits of Compellence: A Research Note." *Political Science Quarterly* 133(1):77–98.

Behrouzian, Golnoosh, Erik C Nisbet, Aysenur Dal and Ali Çarkoğlu. 2016. "Resisting censorship: How citizens navigate closed media environments." *International Journal of Communication* 10:23.

Berejikian, Jeffrey D. 2002. "A cognitive theory of deterrence." *Journal of Peace Research* 39(2):165–183.

Betts, Richard K. 2010. *Nuclear Blackmail and Nuclear Balance*. Brookings Institution Press.

Borghard, Erica D and Shawn W Lonergan. 2017. "The logic of coercion in cyberspace." *Security Studies* 26(3):452–481.

Brehm, Jack W. 1966. *A theory of psychological reactance*. New York: Academic Press.

Brutger, Ryan, Joshua D. Kertzer, Jonathan Renshon, Dustin Tingley and Chagai Weiss. 2021. "Abstraction and Detail in Experimental Design.".
URL: *https://drive.google.com/file/d/1QVlgtci8vPhMV2uO36DAVyBjbmkG9xss/view*

Bullock, John G and Donald P Green. 2021. "The Failings of Conventional Mediation Analysis and a Design-Based Alternative." *Advances in Methods and Practices in Psychological Science* 4(4):25152459211047227.

Chamberlain, Dianne Pfundstein. 2016. *Cheap threats: why the United States struggles to coerce weak states*. Georgetown University Press.

Cherulnik, Paul D and Murray M Citrin. 1974. "Individual difference in psychological reactance: The interaction between locus of control and mode of elimination of freedom." *Journal of Personality and Social Psychology* 29(3):398.

Dafoe, Allan, Jonathan Renshon and Paul Huth. 2014. "Reputation and Status as Motives for War." *Annual Review of Political Science* 17:371–393.

Dafoe, Allan, Sophia Hatz and Baobao Zhang. 2021. "Coercion and Provocation." *Journal of Conflict Resolution* 65(2-3):372–402.

Dillard, James Price and Lijiang Shen. 2005. "On the nature of reactance and its role in persuasive health communication." *Communication Monographs* 72(2):144–168.

Downes, Alexander B. 2018. "Step Aside or Face the Consequences: Explaining the Success and Failure of Compellent Threats to Remove Foreign Leaders." *Coercion: The Power to Hurt in International Politics* pp. 93–114.

Engs, Ruth and David J Hanson. 1989. "Reactance theory: A test with collegiate drinking." *Psychological Reports* 64(3_suppl):1083–1086.

Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes." *American political science review* 88(3):577–592.

Fearon, James D. 1995. "Rationalist explanations for war." *International Organization* 49(3):379–414.

Fortna, Virginia Page. 2015. "Do Terrorists Win? Rebels' Use of Terrorism and Civil War Outcomes." *International Organization* 69(3):519–556.

Gadarian, Shana Kushner. 2014. Beyond the water's edge: Threat, partisanship, and media. In *The Political Psychology of Terrorism Fears*, ed. Samuel Justin Sinclair and Daniel Antonius. New York: Oxford University Press pp. 67–84.

Gadarian, Shana Kushner and Bethany Albertson. 2014. "Anxiety, immigration, and the search for information." *Political Psychology* 35(2):133–164.

George, Alexander and William Simons. 1994. *The limits of coercive diplomacy*. Westview Press.

Graupmann, Verena, Eva Jonas, Ester Meier, Stefan Hawelka and Markus Aichhorn. 2012. "Reactance, the self, and its group: When threats to freedom come from the ingroup versus the outgroup." *European journal of social psychology* 42(2):164–173.

Greenhill, Kelly M. 2010. *Weapons of Mass Migration: Forced Displacement, Coercion, and Foreign Policy*. Cornell University Press.

Hall, Marissa G, Paschal Sheeran, Seth M Noar, Kurt M Ribisl, Laura E Bach and Noel T Brewer. 2016. "Reactance to health warnings scale: development and validation." *Annals of Behavioral Medicine* 50(5):736–750.

Hall, Todd H. 2011. "We will not swallow this bitter fruit: Theorizing a diplomacy of anger." *Security Studies* 20(4):521–555.

Hall, Todd H and Andrew AG Ross. 2015. "Affective politics after 9/11." *International Organization* pp. 847–879.

Haun, Phil. 2015. *Coercion, Survival, and War: Why Weak States Resist the United States*. Stanford University Press.

Holmes, Marcus and Keren Yarhi-Milo. 2017. "The psychological logic of peace summits: How empathy shapes outcomes of diplomatic negotiations." *International Studies Quarterly* 61(1):107–122.

Horowitz, Michael and Dan Reiter. 2001. "When Does Aerial Bombing Work? Quantitative Empirical Tests, 1917-1999." *Journal of Conflict Resolution* 45(2):147–173.

Huddy, Leonie and Stanley Feldman. 2011. "Americans respond politically to 9/11: understanding the impact of the terrorist attacks and their aftermath." *American Psychologist* 66(6):455.

Hufbauer, Gary Clyde, Jeffrey J Schott and Kimberly Ann Elliott. 1990. *Economic Sanctions Reconsidered: History and Current Policy*. Vol. 1 Peterson Institute.

Imai, Kosuke and Teppei Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis* 21(2):141–171.

Jervis, Robert. 1976. *Perception and Misperception in International Politics*. Princeton University Press.

Jones, Seth G and Martin C Libicki. 2008. *How Terrorist Groups End: Lessons for Countering Al Qa'ida*. Vol. 741 Rand Corporation.

Kertzer, Joshua D. 2020. "Re-Assessing Elite-Public Gaps in Political Behavior." *American Journal of Political Science* .

Kertzer, Joshua D, Jonathan Renshon and Keren Yarhi-Milo. 2019. "How Do Observers Assess Resolve?" *British Journal of Political Science* pp. 1–23.

Khong, Yuen Foong. 1992. *Analogies at War: Korea, Munich, Dien Bien Phu, and the Vietnam Decisions of 1965*. Princeton University Press.

Krause, Peter. 2013. "The Political Effectiveness of Non-State Violence: A Two-Level Framework to Transform a Deceptive Debate." *Security Studies* 22(2):259–294.

LaFree, Gary, Laura Dugan and Raven Korte. 2009. "The impact of British counterterrorist strategies on political violence in Northern Ireland: Comparing deterrence and backlash models." *Criminology* 47(1):17–45.

Laurin, Kristin, Aaron C Kay, Devon Proudfoot and Gavan J Fitzsimons. 2013. "Response to restrictive policies: Reconciling system justification and psychological reactance." *Organizational Behavior and Human Decision Processes* 122(2):152–162.

Lerner, Jennifer S and Dacher Keltner. 2000. "Beyond valence: Toward a model of emotion-specific influences on judgement and choice." *Cognition & Emotion* 14(4):473–493.

Lindsay, Jon R. 2013. "Stuxnet and the Limits of Cyber Warfare." *Security Studies* 22(3):365–404.

Lupton, Danielle L. 2020. *Reputation for Resolve: How Leaders Signal Determination in International Politics*. Cornell University Press.

Mack, Andrew. 1975. "Why Big Nations Lose Small Wars: The Politics of Asymmetric Conflict." *World politics* 27(2):175–200.

Maoz, Ifat, Andrew Ward, Michael Katz and Lee Ross. 2002. "Reactive devaluation of an "Israeli" vs."Palestinian" peace proposal." *Journal of Conflict Resolution* 46(4):515–546.

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

Matland, Richard E and Gregg R Murray. 2013. "An experimental test for "backlash" against social pressure techniques used to mobilize voters." *American Politics Research* 41(3):359–386.

Mercer, Jonathan. 2010. *Reputation and International Politics*. Cornell University Press.

Miller, Nicholas L. 2014. "The secret success of nonproliferation sanctions." *International Organization* 68(4):913–944.

Miron, Anca M and Jack W Brehm. 2006. "Reactance theory-40 years later." *Zeitschrift für Sozialpsychologie* 37(1):9–18.

Morgan, T Clifton, Navin Bapat and Yoshiharu Kobayashi. 2014. "Threat and Imposition of Economic Sanctions 1945-2005: Updating the TIES Dataset." *Conflict Management and Peace Science* 31(5):541–558.

Nezlek, John and Jack W Brehm. 1975. "Hostility as a function of the opportunity to counteraggress 1." *Journal of personality* 43(3):421–433.

Nincic, Miroslav. 2011. *The logic of positive engagement*. Cornell University Press.

Nisbet, Erik C, Kathryn E Cooper and R Kelly Garrett. 2015. "The partisan brain: How dissonant science messages lead conservatives and liberals to (dis) trust science." *The ANNALS of the American Academy of Political and Social Science* 658(1):36–66.

Nooruddin, Irfan. 2002. "Modeling selection bias in studies of sanctions efficacy." *International Interactions* 28(1):59–75.

Nyhan, Brendan and Jason Reifler. 2010. "When corrections fail: The persistence of political misperceptions." *Political Behavior* 32(2):303–330.

Pape, Robert A. 1996. *Bombing to Win: Air Power and Coercion in War*. Cornell University Press.

Pape, Robert A. 1997. "Why Economic Sanctions Do Not Work." *International security* 22(2):90–136.

Pape, Robert A. 2003. "The Strategic Logic of Suicide Terrorism." *American political science review* 97(3):343–361.

Pauly, Reid B. C. 2019. Stop or I'll Shoot, Comply and I Won't: Coercive Assurance in International Politics PhD thesis Massachusetts Institute of Technology.

Peffley, Mark and Jon Hurwitz. 2007. "Persuasion and resistance: Race and the death penalty in America." *American Journal of Political Science* 51(4):996–1012.

Petty, Richard E. and Duane T. Wegener. 1999. The Elaboration Likelihood Model: Current Status and Controversies. In *Dual-Process Theories in Social Psychology*, ed. Shelly Chaiken and Yaacov Trope. The Guilford Press pp. 41–72.

Powell, Robert. 1990. *Nuclear Deterrence Theory: The Search for Credibility*. Cambridge University Press.

Press, Daryl Grayson. 2005. *Calculating Credibility: How Leaders Assess Military threats*. Cornell University Press.

Quick, Brian L and Jennifer R Considine. 2008. "Examining the use of forceful language when designing exercise persuasive messages for adults: A test of conceptualizing reactance arousal as a two-step process." *Health communication* 23(5):483–491.

Quick, Brian L and Michael T Stephenson. 2007. "Further evidence that psychological reactance can be modeled as a combination of anger and negative cognitions." *Communication Research* 34(3):255–276.

Rains, Stephen A. 2013. "The nature of psychological reactance revisited: A meta-analytic review." *Human Communication Research* 39(1):47–73.

Rosenberg, Benjamin D and Jason T Siegel. 2018. "A 50-year review of psychological reactance theory: Do not read this article." *Motivation Science* 4(4):281.

Salehyan, Idean. 2009. *Rebels without borders: transnational insurgencies in world politics*. Cornell University Press.

Schelling, Thomas C. 1966. *Arms and Influence*. Yale University Press.

Sechser, Todd S. 2010. "Goliath's Curse: Coercive Threats and Asymmetric Power." *International Organization* 64(4):627–660.

Sechser, Todd S. 2011. "Militarized Compellent Threats, 1918-2001." *Conflict Management and Peace Science* 28(4):377–401.

Sechser, Todd S and Matthew Fuhrmann. 2017. *Nuclear weapons and coercive diplomacy*. Cambridge University Press.

Silvia, Paul J. 2006. "Reactance and the dynamics of disagreement: Multiple paths from threatened freedom to resistance to persuasion." *European Journal of Social Psychology* 36(5):673–685.

Slantchev, Branislav L. 2011. *Military threats: the costs of coercion and the price of peace*. Cambridge University Press.

Slantchev, Branislav L. 2012. "Audience cost theory and its audiences." *Security Studies* 21(3):376–382.

Snyder, Jack and Erica D Borghard. 2011. "The Cost of Empty Threats: A Penny, Not a Pound." *American Political Science Review* 105(3):437–456.

Stein, Janice Gross. 1992. "Deterrence and compellence in the Gulf, 1990-91: A failed or impossible task?" *International Security* 17(2):147–179.

Sullivan, Patricia. 2012. *Who Wins?: Predicting Strategic Success and Failure in Armed Conflict*. Oxford University Press.

Tomz, Michael R and Jessica LP Weeks. 2020. "Human rights and public support for war." *The Journal of Politics* 82(1):182–194.

Yarhi-Milo, Keren. 2018. *Who Fights for Reputation: The Psychology of Leaders in International Conflict*. Princeton University Press.

Online Appendix
The Psychology of Coercion Failure: How Reactance Explains Resistance to Threats

Contents

A Study 1

A.1 Sample Characteristics

We recruited respondents via Prolific Academic. 1,662 American respondents, located in the U.S., completed informed consent.[1] We removed 1) 16 observations that failed the reCAPTCHA bot detection; and 2) 205 that failed 1-2 pre-treatment attention checks. Nine participants completed 30 and 88% of the survey. Although we retain those partial responses where possible, these adjustments produce an effective sample of 1,433. Table A1 presents the demographic characteristics of our sample.

Table A1: Study 1 Sample Characteristics

	Percent Sample
Male	51.85
18-24	28.51
25-34	32.98
35-44	18.87
45-54	10.90
55+	8.74
University graduate	52.16
White	64.83
Income: <$30,000	24.37
Income: $30-60,000	27.09
Income: $60-100,000	27.51
Income: Over $100,000	21.02

A.2 Manipulation Checks and Information Equivalence

We included two factual manipulation checks. One asked participants to identify whether or not someone from Navalia had threatened them, and another asked them to recall the number of soldiers who would die if the ship sank. Across groups, 94% correctly identified whether or not someone from Navalia threatened them, and 98% of respondents correctly recalled the number of soldiers who would die if the ship sank, confirming that participantsreceived and understood the treatments.

The threat treatment raised reactance

We included a subjective manipulation check designed to measure psychological reactance. We modified 5 standard items from psychological research to suit the vignette (Dillard and Shen, 2005; Hall et al., 2016), asking participants whether they agree with the following statements and creating an additive index from their responses ($\alpha = 0.87$): 1) Navalia tried to make a decision for me.; 2) Navalia tried to manipulate me.; 3) Navalia tried to pressure me.; 4) I felt like the Navalia was trying to take away my freedom to make a choice.; 5) Navalia is treating me like a fool.

The threat treatment increased scores on the reactance scale ($b = 0.14$, $p < 0.01$), but the reputation treatment did not ($b = 0.003$, $p = 0.79$). Model 1 of Table A2 presents the results for the four separate treatment groups and confirms that threats raise reactance.

The treatments do not systematically alter perceptions of power nor stakes

Three additional questions probed whether participants understood the material risks associated with their decision to continue to the island or turn back. These questions help to establish whether the threat or

[1] Participants received $1.27 for the 7-minute survey.

Table A2: Study 1 Reactance and Information Equivalence

	Reactance	Navalia Stronger	Turn back: Soldiers Live	Turn back: Keep Island
	(1)	(2)	(3)	(4)
Natural Costs/Reputation	0.03	0.01	0.01	−0.01
	(0.02)	(0.01)	(0.02)	(0.02)
Threat/No Reputation	0.17**	0.003	0.02	−0.04*
	(0.02)	(0.01)	(0.02)	(0.02)
Threat/Reputation	0.15**	0.01	0.01	−0.004
	(0.02)	(0.01)	(0.02)	(0.02)
Controls	✓	✓	✓	✓
Constant	0.41**	0.47**	0.79**	0.26**
	(0.04)	(0.02)	(0.04)	(0.04)
N	1,428	1,428	1,426	1,426
R^2	0.12	0.01	0.03	0.03

*p < .05; **p < .01

Note: Models display OLS coefficients. All dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is natural costs, no reputation. All models include controls for militarism, isolationism, gender, race, age, party identification, income, and education, suppressed for space.

reputation treatments affect participant beliefs about other background aspects of the vignette, which would limit our ability to infer information equivalence across groups (Dafoe, Zhang and Caughey, 2018).

The first question asked participants to evaluate Navalia's power, relative to their own country. Model 2 in Table A2 shows that the treatments did not alter participant beliefs about Navalia's power.

Participants also estimated the chances that 1) their country's soldiers will die and that 2) they will lose the island and its resources should they turn the ship around. Higher values indicate participant estimates that if they turn back, their country's soldiers "definitely will not" die (*soldiers live* in Model 3 of Table A2) and they "definitely will not" lose the island (*keep island* in Model 4).

Notably, all participants learn that the only risk to their soldiers' lives comes from continuing to the island, and that turning back means losing the territory to Navalia. Measuring participant perceptions about these elements has two advantages. First, we want to evaluate whether the threat treatment implicitly altered participant perceptions about the stakes (potential information leakage). Second, each item provided potential fodder for reactance-induced counter-arguing. As we wrote in the pre-analysis plan, people might downgrade the risks or overweight the benefits from the prohibited course of action (RQ1a and RQ1b).

Results in Model 3 shows that none of the treatments affect participants' estimates about whether turning back will allow their country's soldiers to live. Most participants correctly deduced that turning back would mean safety. Most participants also recognized that turning back means losing the island (overall mean= 0.17). But Model 4 hints at one potential manifestation of counterarguing, in that the threat/no reputation treatment increased participants' certainty that turning back would mean losing the island — perhaps as another cognitive justification for moving forward against Navalia's demands. Yet we only find evidence for this effect in the reputation absent group, suggesting caution in this interpretation. Instead, the results in Table A2 collectively suggest that the threat treatments increased reactance and provide additional confidence that the vignettes controlled for relative power and stakes, thereby isolating reactance from alternative explanations.

A.3 Main Treatment Effects

Table A3 presents OLS estimates for the effects of each treatment on our dependent variables — the continuous measure for confidence in the decision to continue to the island (Models 1-3), support for sanctions

(Models 4-6), and support for war (Models 7-9). Models 2, 5, and 8 show that the reputation treatment does not moderate the effect of the threat, and Models 3, 6, and 9 show that the results hold when we adjust for a panel of control variables.

Table A3: Study 1: Treatment Effects on Coercion Failure and Retaliation

	Continue to Island			Support Sanctions			Support War		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reputation	0.02	0.02	0.01	−0.005	−0.002	−0.01	0.03*	0.03	0.02
	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
Threat	0.09**	0.08**	0.09**	0.05**	0.06**	0.05**	0.06**	0.06**	0.06**
	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
Rep x Threat		0.01			−0.01			−0.01	
		(0.04)			(0.03)			(0.03)	
Controls			✓			✓			✓
Constant	0.39**	0.39**	0.22**	0.66**	0.66**	0.53**	0.29**	0.29**	0.10*
	(0.02)	(0.02)	(0.07)	(0.01)	(0.01)	(0.04)	(0.01)	(0.01)	(0.04)
N	1,440	1,440	1,429	1,440	1,440	1,429	1,440	1,440	1,429
R^2	0.01	0.01	0.14	0.01	0.01	0.09	0.02	0.02	0.22

*p < .05; **p < .01

Note: Models display OLS coefficients, all dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is natural costs, no reputation. Models 3, 6, and 9 include controls for for militarism, isolationism, party identification, gender, race, age, university, and income, omitted for space.

A.4 Mechanisms of Coercion Failure in Study 1

We test reactance as a cause of coercion failure first by designing experiments that compare coercive threats to natural cost counterparts. These vignettes control for other factors and isolate the causal effect of the threat itself on whether participants choose to make a policy concession or support retaliation against their coercer. Next, we included a battery of post-treatment questions to assess the mechanisms for this causal effect — how do threats shape the propensity to resist? We test multiple mechanisms because we aim to both evaluate whether threats increase psychological reactance and test alternative explanations.

First, evidence about whether the treatment directly affects a potential mediator provides important information about the likely and unlikely causal pathways: Evidence that the threat treatment affects a given mediator is consistent with what we would expect if that mechanism contributes to the treatment effect. Evidence that the threat treatment *does not* affect a mediator, conversely, matches what we would expect if the mechanism does not mediate the relationship between the threat treatment and the outcome variables (Bullock and Green, 2021, 12-13). Table A4 displays estimates from OLS models that regress each of the 6 proposed mechanisms separately on three treatment dummies (natural costs, no reputation as the omitted category) and the panel of control variables.[2]

As we describe in the paper, the results show that the threat treatments had positive and significant effects on two key indicators for reactance: anger (Model 1) and perceptions that the threat has been exaggerated (Model 2). The reputation treatments increased concerns about whether turning back would cause their country's reputation to suffer (Model 5) or lead the public to disapprove of their leadership (Model 6). The results from Models 5 and 6 support hypotheses 5a-5d from our pre-analysis plan.

We included questions to gauge participants' perceptions about the island's value (Model 3) and priority associated with protecting their soldiers' lives (Model 4) as potential manifestations of counterarguing. However, we do not find evidence that the threat treatment uniquely affects either variable. Regarding perceptions of the island's value, the vignette's explicit statement about the island's value might contribute

[2]A parallel analysis and exploratory factor analysis on the three counterarguing measures suggested that they did not load together. Per our pre-analysis plan, we therefore analyzed each separately.

to the null effect. More broadly, these results suggest the need for additional work to determine the best ways to measure reactance-induced counterarguing in an interstate coercion scenario, given the many potential manifestations of this cognitive mechanism (Hall et al., 2016).

Table A4: Treatment Effects on Mechanisms

	Anger	Threat Exaggerated	Island Valuable	Protect Soldiers	Reputation Suffer	Public Disapproval
	(1)	(2)	(3)	(4)	(5)	(6)
Natural Costs/Reputation	0.02	0.001	−0.01	0.04*	0.10**	0.10**
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
Threat/No Reputation	0.06**	0.12**	0.01	0.03*	0.01	0.01
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
Threat/Reputation	0.04*	0.11**	−0.001	0.03*	0.13**	0.13**
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
Controls	✓	✓	✓	✓	✓	✓
Constant	0.36**	0.18**	0.30**	0.16**	0.46**	0.40**
	(0.05)	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)
N	1,423	1,428	1,428	1,428	1,426	1,426
R^2	0.05	0.12	0.05	0.09	0.07	0.07

*p < .05; **p < .01

Note: Models display OLS coefficients. All dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is natural costs, no reputation. Models control for militarism, isolationism, party identification, gender, race, age, university, and income, omitted for space.

Causal mediation and sensitivity analyses for study 1

Second, we conducted a series of nonparametric mediation analyses (Imai and Yamamoto, 2013; Tingley et al., 2014) to estimate the average causal mediation effect for each mediator. We separately evaluated the effect of the threat and reputation treatments via each mechanism, while controlling for the other treatment.

Although our pre-analysis plan specified that we would evaluate mediation via the product-of-coefficients approach (Baron and Kenny, 1986), diagnostic tests suggested evidence for causal dependence between the mechanisms: As we might expect from manipulating reputation and audience costs together or when different aspects of countarguing might work together, for example, we observed some evidence for significant relationships between alternative mechanisms after regressing each mediator on the treatments, pre-treatment covariates, and other mechanisms (Imai and Yamamoto, 2013, 167).[3] We therefore turned to a causal mediation model that specifically accounts for potential dependence between mediators. We present estimates using the pre-registered product-of-coefficients approach below, noting that we draw similar conclusions from both methods but adopt the more rigorous alternative for our main inferences.

Notably, although these causal mediation models allow dependece between mechanisms, they require an additional assumption: that the mediator's effect does not depend on treatment assignment — the absence of an interaction between the treatment and mediator. This section presents results from sensitivity analyses that assess the robustness of these estimates to potential violations of that assumption.

The sensitivity analyses calculate two quantities of interest. The σ value indicates "the standard deviation (SD) of the individual-level coefficient for the treatment–mediator interaction" (Imai and Yamamoto, 2013, 158). The sensitivity analyses vary this unobserved quantity, calculating upper and lower bounds on the ACME estimate as σ increases to evaluate how much heterogeneity in the interaction between the treatment (threat or reputation) and mediator would render the ACME estimate non-significant. The $\tilde{R}^2$ value shows the proportion of total variance in the dependent variable that the treatment-mediator interaction would explain if modeled (Imai and Yamamoto, 2013, 159); it indicates the importance of the potential interaction,

[3]Analyses available in the replication materials.

insofar as it provides a concrete threshold for maintaining the original conclusion of a significant ACME. Its upper bound represents a situation where the unobserved interaction explains all residual variance in the model (ibid., 161). Higher values on either quantity indicate that the ACME estimate is robust to larger violations of the assumption.

We conduct sensitivity analysis for statistically significant ($p < 0.05$) ACME estimates in our main analyses. To our knowledge the field has not reached a consensus regarding which σ and $\tilde{R}^2$ values constitute reasonable robustness. We present the sensitivity results while placing them in context of the total variance explained in the main treatment effect models. Results from the sensitivity tests suggest that the importance of threat exaggeration is relatively robust and that the ACME estimates on anger are more sensitive to the homogeneous interaction assumption.

The bottom three rows in Table A5 display the sensitivity analyses for the threat's effects via counterarguing. The results suggest that our conclusions about counterarguing are especially robust with respect to the effect on coercion failure via continuing to the island. The ACME lower bound estimate crosses zero when σ equals 0.192, 26.6% of its largest possible value; the confidence interval includes zero at $\tilde{R}^2$ values greater than 0.035. Our conclusions retain support unless the interaction explains more than 3.5% of the total variance in participants' choice about whether to continue to the island — a reasonably high bar relative to R^2 from the model regressing the outcome the treatments (0.01; see Table A3).

The top two rows in Table A5 display the same quantities with respect to anger as the mediator connecting the threat treatment to support for war and sanctions. The results suggest that relaxing the homogeneous interaction assumption could produce a null ACME estimate when the σ value reaches 0.05 for the sanctions outcome and 0.049 for war support — 10.1% of its largest possible value in each case. The ACME estimate for anger crosses zero when the interaction explains 0.8 and 0.7% of the variance in support for sanctions or war, respectively. Although these findings suggest that the ACME estimates for anger may be more fragile than their counterarguing counterparts, they imply that the treatment-mediator interaction must explain nearly as much variance in the dependent variables as the experimental treatments to update our conclusions. Moreover, the robust direct treatment effects on anger remain an important source of support for the conclusion that anger partly mediates the threat's effects.

Table A5: Sensitivity Analyses for Threat via Reactance Mechanisms

Mechanism	DV	ACME (Avg) (1)	CI Low (2)	CI High (3)	σ (4)	Max σ (5)	$\tilde{R}^2$ (6)	Max $\tilde{R}^2$ (7)
Anger	Sanctions	0.012	0.004	0.02	0.05	0.494	0.008	0.831
Anger	War	0.008	0.002	0.014	0.049	0.484	0.007	0.701
Threat Exaggerated	Continue	0.046	0.033	0.059	0.192	0.722	0.035	0.49
Threat Exaggerated	Sanctions	-0.006	-0.012	0	0.062	0.62	0.008	0.831
Threat Exaggerated	War	0.012	0.006	0.018	0.061	0.608	0.007	0.701

Note: Columns 1-3 display the estimated ACMEs and 95% bootstrapped confidence intervals for factors that significantly mediated the strong threat's effect in study 2, estimated assuming the homogeneous interaction assumption using the mediation package in R (Tingley et al., 2014). Column 4 displays the value of the sensitivity parameter *sigma* at which the confidence interval for the ACME would include 0 and represents the sensitivity with respect to interaction heterogeneity (Imai and Yamamoto, 2013, 162). Column 5 displays the maximum possible value for *sigma*. Column 6 displays the value for $\tilde{R}^2$ at which the ACME estimate bounds would include 0, and column 7 displays the largest possible value for $\tilde{R}^2$. $\tilde{R}^2$ indicates how much variance the treatment-mediator interaction must explain to render the ACME non-significant.

Finally, Table A6 examines whether the effect of reputation through reputation is reasonably robust to violations of the homogeneous interaction assumption. The ACME estimate on continuing to the island includes 0 when σ crosses 0.099 — when the interaction explains 1.5% of the variance in participants' choice. For the war support outcome, the confidence interval includes 0 for σ values greater than 0.048, when the interaction explains more than 0.7% of the variance.

Table A6: Sensitivity Analyses for Reputation Treatment via Reputation and Audience Cost Mechanisms

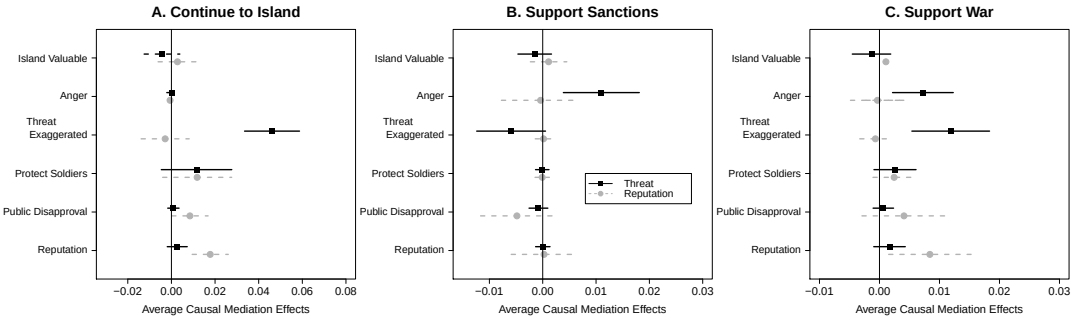
Mechanism	DV	ACME (Avg) (1)	CI Low (2)	CI High (3)	σ (4)	Max σ (5)	$\tilde{R}^2$ (6)	Max $\tilde{R}^2$ (7)
Reputation	Continue	0.022	0.013	0.031	0.099	0.571	0.015	0.487
Reputation	Support War	0.01	0.003	0.017	0.048	0.482	0.007	0.702
Public Disapproval	Continue	0.009	0	0.018	0.06	0.596	0.005	0.487

Note: Columns 1-3 display the estimated ACMEs and 95% bootstrapped confidence intervals for factors that significantly mediated the strong threat's effect in study 2, estimated assuming the homogeneous interaction assumption using the mediation package in R (Tingley et al., 2014). Column 4 displays the value of the sensitivity parameter *sigma* at which the confidence interval for the ACME would include 0 and represents the sensitivity with respect to interaction heterogeneity (Imai and Yamamoto, 2013, 162). Column 5 displays the maximum possible value for *sigma*. Column 6 displays the value for $\tilde{R}^2$ at which the ACME estimate bounds would include 0, and column 7 displays the largest possible value for $\tilde{R}^2$. $\tilde{R}^2$ indicates how much variance the treatment-mediator interaction must explain to render the ACME non-significant.

Perceptions of island's value do not mediate treatment effects

Our pre-analysis plan did not specify conditions under which we would exclude any of the proposed mechanisms from the mediation analyses, a minor oversight. Insofar as the absence of direct effects provides strong suggestive evidence against perceptions of the island's variable as a mechanism for any of the treatment effects, it makes little sense to include it and we proceeded without the item in our main analyses (Zhao, Lynch Jr and Chen, 2010). Still, we estimated the causal mediation models while accounting for perceptions about the island's value as an alternative mediator to probe the robustness of our results. Figure A1 presents the results, which are consistent with our primary estimates.

Figure A1: Including Perceptions of Island's Value as Mediator does not Affect ACME Estimates



Note: Lines depict 95% confidence intervals. ACME estimates generated using the medation pacakage in R (Tingley et al., 2014; Imai and Yamamoto, 2013). Models include the panel of control variables.

Product of coefficients mediation analysis

This section presents results from our pre-registered product-of-coefficients mediation analysis. Table A7 displays results from three OLS models that regress each dependent variable on the treatments, the 5 mechanisms affected by the treatments, and the control variables. Figure A2 displays the indirect effect estimates calculated by multiplying estimates from models in Table A7 by the direct effects in Table A4. The lines depict 95% bootstrapped confidence intervals for the indirect effect estimates (Preacher and Hayes, 2008). The results from this analysis support the conclusions from the causal mediation analysis in the manuscript, and we include them here for completeness and transparency.

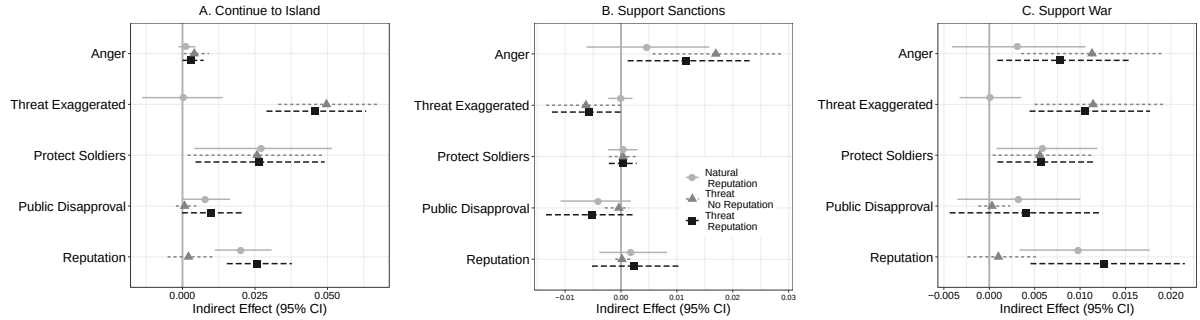
Table A7: Effect of Mediators on Dependent Variables

	Continue to Island	Support War	Support Sanctions
	(1)	(2)	(3)
Natural Costs/Reputation	-0.05^{*}	-0.001	-0.01
	(0.02)	(0.02)	(0.02)
Threat/No Reputation	0.01	0.04^{*}	0.04^{*}
	(0.02)	(0.02)	(0.02)
Threat/Reputation	-0.01	0.04^{*}	0.04^{*}
	(0.02)	(0.02)	(0.02)
Anger	0.06^{*}	0.18^{**}	0.27^{**}
	(0.03)	(0.02)	(0.02)
Threat Exaggerated	0.40^{**}	0.09^{**}	-0.05^{*}
	(0.03)	(0.02)	(0.03)
Value Soldiers	0.76^{**}	0.17^{**}	0.01
	(0.04)	(0.03)	(0.03)
Public Disapproval	0.08^{*}	0.03	-0.04
	(0.04)	(0.03)	(0.03)
Reputation	0.19^{**}	0.09^{**}	0.02
	(0.03)	(0.03)	(0.03)
Controls	✓	✓	✓
Constant	-0.12^{*}	-0.07	0.45^{**}
	(0.05)	(0.04)	(0.04)
N	1,421	1,421	1,421
R^2	0.51	0.30	0.17

$^{*}p < .05$; $^{**}p < .01$

Note: Models display OLS coefficients. All dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is natural costs, no reputation.Models control for militarism, isolationism, party identification, gender, race, age, university, and income, omitted for space.

Figure A2: Indirect Effects



Note: Indirect effects relative to natural costs/no reputation (bootstrapped 95% confidence intervals; Preacher and Hayes, 2008). Models include pre-treatment controls.

B Study 2

B.1 Sample Characteristics

We recruited an effective sample of 3,526 participants over a two-week period in August 2020 via Lucid Theorem. Lucid Theorem is an online survey platform that works with a network of companies to select a diverse set of participants that resemble the U.S. population on key demographic characteristics, including age, gender, race, and region of residence. We screened out participants with foreign IPs or VPNs prior to the informed consent process (Burleigh, Kennedy and Clifford, 2018). We completed data collection in two batches, which retained the same instrument but evaluated participant attentiveness pre-treatment in different ways.

The first batch did not include a pre-treatment attention check question, and we relied only on time spent reading the preliminary scenario description (prior to treatment) to evaluate engagement. We removed participants who spent less than 2.1 seconds reading the vignette ($< 4\%$). While this batch was in the field, survey researchers began raising concerns about data quality and inattention on Lucid Theorem (Aronow et al., 2020). For the second batch, we therefore followed guidance to include a pre-treatment attention check and screen out inattentive respondents before they advanced in the survey. Of those who consented to participate, 51.7% failed the attention check and terminated from the survey. We have no reason to expect that attentiveness moderates our treatment effects and rather that it decreases the precision of our estimates. We pool data from the two batches for analysis, but also present the results for a) the sample that includes those who spent < 2.1 seconds reading the pre-treatment vignette and b) the separate second batch to show that our findings remain robust. Table B8 presents demographic data for the effective sample compared to the U.S. adult population.

Table B8: Sample Characteristics compared to U.S. Adult Population

	Sample	Population
Female	0.533	0.508
Male	0.467	0.492
18-24	0.115	0.119
25-44	0.365	0.342
45-64	0.345	0.326
65+	0.175	0.212
College/university or Higher	0.446	0.306
White	0.736	0.763

B.2 Coercion Vignette

Following the pre-treatment questionnaire, all participants received a prompt instructing them that "On the next page, you are going to read about a hypothetical foreign policy issue that is similar to situations the U.S. has faced in the past and will probably face again. For scientific validity, the situation that you will read does not refer to any specific event in history or in the news today. For the next questions, we'd like you to imagine that you are the U.S. Ambassador to Eritrea. You are responsible for making policy decisions toward Eritrea." A single question then asked participants to affirm that they "agree to read the details very carefully, and then give your most thoughtful answers," before presenting the following scenario on the next page:

Here is the situation:

- A terrorist group affiliated with ISIS has sworn to attack the United States.
- This terrorist group is setting up bases in the country of Eritrea.
- Because U.S. intelligence agencies have been unable discover the exact locations of these bases, you need the cooperation of the Eritrean government to remove them.
- Eritrea is ruled by an oppressive dictator whose secret police have killed hundreds of innocent civilians in towns seen as less loyal to him.
- Consequently, you (the United States) have been blocking all sales of weapons to Eritrea.
- If the dictator had these weapons, he would use them to massacre thousands of civilians.

Next, we randomly assigned participants to one of four groups. Because the page advanced, a short summary preceded all four treatments:

Just to review:

- A terrorist group targeting the United States is setting up bases in Eritrea, and you need the cooperation of the Eritrean dictator to remove them.
- The United States is blocking the sale of weapons to Eritrea. With those weapons, the dictator would massacre thousands of civilians in areas seen as less loyal to him.

In the *baseline control* group, participants read the following:

- Allowing weapons sales to Eritrea would happen secretly. It would not receive attention from the media or other countries.
- The CIA assesses that future conflicts with Eritrea are unlikely.

In the *natural costs* group, participants a prompt directing them that although the dictator did not issue a threat, allowing arms sales could alleviate the terrorist threat:

- Although the dictator hasn't said anything to you, you know that **his government needs the weapons to be able to destroy the terrorist group's bases.**
- If you change U.S. policy to allow Eritrea to buy weapons, the CIA is confident that the Eritrean government would then remove the bases.
- Allowing weapons sales to Eritrea would happen secretly. It would not receive attention from the media or other countries.
- The CIA assesses that future conflicts with Eritrea are unlikely.
- If the bases are not destroyed, the CIA assesses that the terrorist group will attack the United States in the next year, **likely killing dozens of Americans.**

The *strong threat* treatment included a forceful, dogmatic demand similar to threat treatment from study 1:

- In a private meeting earlier today, the dictator threatened you, "**You must immediately stop meddling with Eritrea! If you do not allow us to buy the weapons, I will let the terrorist group operate freely.**"
- If you change U.S. policy to allow Eritrea to buy weapons, the CIA is confident that the Eritrean government would then remove the bases.
- Allowing weapons sales to Eritrea would happen secretly. It would not receive attention from the media or other countries.
- The CIA assesses that future conflicts with Eritrea are unlikely.
- If the bases are not destroyed, the CIA assesses that the terrorist group will attack the United States in the next year, **likely killing dozens of Americans.**

The *weak threat* treatment used less dogmatic language but conveyed the same threat:

- In a private meeting earlier today, the dictator appealed to you: "**I hope that you consider allowing us to buy the weapons. It's your choice. But if you do not, I will have to let this terrorist group operate freely.**"
- If you change U.S. policy to allow Eritrea to buy weapons, the CIA is confident that the Eritrean government would then remove the bases.
- Allowing weapons sales to Eritrea would happen secretly. It would not receive attention from the media or other countries.
- The CIA assesses that future conflicts with Eritrea are unlikely.
- If the bases are not destroyed, the CIA assesses that the terrorist group will attack the United States in the next year, **likely killing dozens of Americans.**

Dependent Variables

Following the treatment text, a single question asked participants "What do you think the U.S. should do in this situation?" and allowed them to choose either to "Continue to block weapons sales to Eritrea" or "Allow weapons sales to Eritrea." Immediately after selecting their preferred policy, participants reported "On a scale from 0-10, how strongly do you feel about this policy choice?" We combined responses from these branched questions to create a continuous dependent variable that ranges from 0 (strongly support allowing weapons sales) to 1 (strongly support blocking weapons sales).

To measure support for retaliation, a prompt asked participants "In response to this situation, to what extent would you support or oppose each of the following:" "The U.S. imposing sanctions against Eritrea?", "U.S. military strikes against Eritrean military bases?", and "The U.S. invading Eritrea to take control of the country?" We presented these three items in random order and measured each on a 7-point scale from "strongly oppose" to "strongly support."

B.3 Manipulation Checks

Threats increased reactance

To evaluate whether our coercion treatments induced reactance, we asked participants whether they agree or disagree that the dictator tried to make a decision for them, tried to manipulate them, tried to pressure them, used forceful language, or was trying to take away their freedom to choose. Responses ranged from "not applicable" to "strongly agree" on a 5-item scale, and we combined these items into an additive scale ($\alpha = 0.92$; rescaled to range from 0 to 1). Both the strong ($b = 0.21$, $p < 0.01$) and weak ($b = 0.16$, $p < 0.01$) threats increased reactance relative to natural costs. Moreover, participants in the strong threat treatment expressed more reactance than their counterparts who received the weak threat ($b = 0.06$, $p < 0.01$), in

line with evidence that reactance increases with more dogmatic language or vivid threats (Gadarian, 2014; Hall et al., 2016). This subjective manipulation check confirms that our treatments manipulate the targeted construct, and that strong threats increased reactance more than weak threats.

Factual recall

A factual manipulation check, included after the dependent variables, asked participants: "In the scenario you read, did the Eritrean dictator directly demand something from the United States?" Options included "Yes", "No", "The dictator asked for something, but did not demand it", and "I'm not sure." In the strong threat group, 69% answered the question correctly, compared to 59% in the natural costs group, 62% in the control condition, and 41% in the weak threat group. 37% of participants exposed to the weak threat interpreted the dictator's request as a demand — illustrating the challenge involved in communicating coercive threats. We include all respondents in our analyses to avoid concerns about post-treatment bias associated with removing respondents who fail comprehension checks (Montgomery, Nyhan and Torres, 2018). The substantive results of our main treatment effects do not change if we exclude those who failed the post-treatment comprehension check, with one important exception: The strong threat increases support for sanctions and strikes more than the weak threat, a finding consistent with reactance.

B.4 Measuring Mechanisms of Coercion Failure

To measure anger, a prompt instructed participants to "Please indicate your feelings toward the Eritrean dictator when thinking about this situation. To what extent do you feel...": Angry, Anxious, Furious, Sad, Excited, Irritated, Annoyed (presented in random order). Participants reported their emotions on a scale from 0 ("not at all) to 10 ("very much"), and we created an additive scale that combines scores for angry, furious, irritated, and annoyed (we included the others as distractors, and do not include them in analyses).

Finally, we measured cognitive reactance alongside alternative explanations by asking participants to estimate "If the U.S. decides to allow Eritrea to buy weapons, what are the chances that each of the following things will happen": (in random order) "The American public will disapprove of the president."; "The US reputation for resolve will suffer in the eyes of other countries."; "There will be a terrorist attack in the U.S."; "Thousands of Eritrean civilians will be killed."; or "The U.S. reputation for moral authority will suffer in the eyes of other countries." Participants recorded their estimates on a 5-point scale from "definitely will not" to "definitely will."

B.5 Treatment effects

First, Table B9 presents results from two OLS models that regress the binary DV on the experimental treatments. The results show that both the strong and weak threats increase the proportion of people who chose to continue blocking arms — potentially endangering the U.S. homeland. Our results are consistent across specifications of the dependent variable, providing additional confidence in our conclusions.

Table B10 presents results from OLS models that regress each dependent variable on our treatments both without (Models 1, 3, 5, and 7) and with (Models 2, 4, 6, and 8) a panel of pre-treatment controls. The results show that both threat treatments had positive, significant effects on support for blocking arms sales, imposing sanctions, and launching strikes against Eritrea. By contrast, dispositional militarism plays the biggest role in explaining support for invading Eritrea to take over the country ($b = 0.348$, $p < 0.01$) — revealing limits to reactance-induced support for retaliation.

B.6 Mechanisms of Coercion Failure in Study 2

This section includes a) models estimating the direct effects of the threat treatments on each respective mechanism; b) sensitivity analyses for the causal mediation models presented in the manuscript; and c) estimates from a product-of-coefficients analysis that assumes causal independence between mediators.

Table B9: Study 2 Treatment Effects

	Block Arms (binary)	
	(1)	(2)
Control	0.188**	0.197**
	(0.021)	(0.022)
Strong Threat	0.049*	0.060**
	(0.020)	(0.021)
Weak Threat	0.076**	0.086**
	(0.020)	(0.021)
Controls		✓
Constant	0.674**	0.758**
	(0.014)	(0.052)
N	3,487	3,224
R^2	0.024	0.039
Constant	0.669**	0.743**

*p < .05; **p < .01
Note: Models display OLS coefficients. All dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is the natural costs condition. Controls for militarism, isolationism, gender, race, age, party identification, census region, income, education, and political knowledge suppressed for space.

Table B10: Study 2: Treatment Effects

	Block Arms Sales		Support Sanctions		Support Strikes		Support Invasion	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Control	0.145**	0.151**	0.015	0.015	−0.004	−0.004	−0.021	−0.014
	(0.016)	(0.016)	(0.014)	(0.014)	(0.015)	(0.014)	(0.015)	(0.015)
Strong Threat	0.041**	0.048**	0.042**	0.048**	0.048**	0.051**	0.014	0.019
	(0.015)	(0.015)	(0.014)	(0.013)	(0.014)	(0.014)	(0.015)	(0.014)
Weak Threat	0.063**	0.068**	0.024	0.023	0.018	0.020	0.008	0.015
	(0.015)	(0.016)	(0.014)	(0.013)	(0.014)	(0.014)	(0.015)	(0.014)
Controls		✓		✓		✓		✓
Constant	0.643**	0.688**	0.596**	0.386**	0.454**	0.143**	0.399**	0.088*
	(0.010)	(0.039)	(0.010)	(0.033)	(0.010)	(0.035)	(0.010)	(0.036)
N	3,487	3,224	3,467	3,224	3,467	3,224	3,467	3,224
R^2	0.025	0.039	0.003	0.130	0.005	0.134	0.002	0.139

*p < .05; **p < .01
Note: Models display OLS coefficients. All dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is the natural costs condition. Controls for militarism, isolationism, gender, race, age, party identification, census region, income, education, and political knowledge, suppressed for space.

Direct effects on mechanisms

We first regress each mechanism on our treatment variables, and present the results in Tables B11 and B12 below. The results show that our treatments had significant effects on multiple potential mechanisms — these include large, positive effects on anger and participant evaluations of whether selling arms to the

Table B11: Study 2 Treatment Effects on Reactance Mechanisms

	Reactance Scale		Anger		Chances U.S. Attack		Chances Eritrean Deaths	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Control	-0.024^{*}	-0.023^{*}	0.031^{*}	0.034^{**}	0.036^{**}	0.034^{**}	0.037^{**}	0.037^{**}
	(0.011)	(0.011)	(0.013)	(0.013)	(0.013)	(0.013)	(0.012)	(0.012)
Strong Threat	0.214^{**}	0.215^{**}	0.088^{**}	0.089^{**}	0.045^{**}	0.044^{**}	0.021	0.020
	(0.010)	(0.010)	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)
Weak Threat	0.157^{**}	0.163^{**}	0.070^{**}	0.069^{**}	0.039^{**}	0.040^{**}	0.036^{**}	0.040^{**}
	(0.011)	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)
Controls		✓		✓		✓		✓
Constant	0.529^{**}	0.377^{**}	0.568^{**}	0.460^{**}	0.541^{**}	0.399^{**}	0.707^{**}	0.519^{**}
	(0.007)	(0.026)	(0.009)	(0.031)	(0.008)	(0.031)	(0.008)	(0.028)
N	3,441	3,224	3,398	3,203	3,417	3,223	3,415	3,221
R^2	0.170	0.239	0.018	0.063	0.005	0.067	0.004	0.050

$^{*}p < .05$; $^{**}p < .01$

Note: Models display OLS coefficients. All dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is the natural costs condition. Additional controls include militarism, isolationism, gender, race, age, party identification, census region, income, education, and political knowledge, suppressed for space.

Eritreans would affect the chance of a terrorist attack alongside smaller but significant effects on concerns about whether selling arms to the dictator would cause the U.S. reputation for resolve or moral authority to suffer.

The results reveal two important findings. First, both strong and weak threats prompted anger and assertions that allowing arms sales would increase the chance of a terrorist attack in the United States. Participants who received a threat reacted with annoyance, irritation, and frustration with anger scores 9 and 7 percentage points higher than their counterparts in the natural costs group, respectively (both $p < 0.01$).

Consistent with counterarguing, both threats also caused participants to discount the potential benefits from conceding to the dictator. The strong and weak threats increased estimates that allowing arms sales would lead to a terrorist attack in the U.S. ($b = 0.04$ and $b = 0.04$, both $p < 0.01$). The weak threat heightened perceptions that allowing arms sales would harm Eritrean civilians ($b = 0.04$, $p < 0.01$), though the strong threat did not exert a statistically significant effect compared to the natural costs group ($p = 0.07$). Reactance's cognitive route runs through counterarguing, a pathway that includes discounting and dismissing risks while exaggerating costs associated with following commands.

Although we find no evidence that coercive threats increased expectations of public disapproval for allowing arms sales, threats did increase concerns that doing so would harm U.S. reputations for resolve and moral authority. Importantly, as we discuss in the manuscript, the indirect effects of reputational concerns account for only a small portion of the observed treatment effect on the block arms dependent variable, but do not significantly mediate the threats' effects on support for sanctions nor strikes. These results could indicate that, despite a design that minimizes reputation in principle, some respondents nevertheless remained convinced that reputation was not at stake. However, it is also possible that, when asked to provide a rationalization for decisions motivated by psychological reactance, respondents turned to reputational justifications (among others) rather than recognize or admit to non-rational psychological motives. Indeed, expressed concerns about reputation could be a form of counterarguing — though evaluating that intriguing possibility would require additional research.

Table B12: Study 2 Treatment Effects on Alternative Mechanisms

	Reputation Resolve		Moral Reputation		Public Disapproval	
	(1)	(2)	(3)	(4)	(5)	(6)
Control	0.026*	0.028*	0.030*	0.032*	0.029*	0.030*
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Strong Threat	0.031**	0.034**	0.038**	0.037**	0.005	0.005
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Weak Threat	0.047**	0.055**	0.043**	0.050**	0.024*	0.027*
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Controls		✓		✓		✓
Constant	0.645**	0.508**	0.662**	0.561**	0.649**	0.507**
	(0.008)	(0.030)	(0.008)	(0.030)	(0.008)	(0.030)
N	3,417	3,223	3,417	3,223	3,415	3,221
R^2	0.005	0.046	0.005	0.048	0.002	0.038

*p < .05; **p < .01
Note: Models display OLS coefficients. All dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is the natural costs condition. Additional controls include militarism, isolationism, gender, race, age, party identification, census region, income, education, and political knowledge, suppressed for space.

Sensitivity analyses for causal mediation analyses

Tables B13 and B14 display results from sensitivity analyses evaluating the robustness of the mediation results to violations of the treatment-mediator homogeneity assumptions. We include sensitivity results for ACME estimates that met a 0.05 statistical significance threshold.

The top three rows in Table B13 shows that relaxing the homogeneous interaction assumption could produce a null ACME estimate for anger when the σ value reaches 0.086 for blocking arms sales, 0.075 for the sanctions outcome and 0.078 for military strikes. The ACME estimate for anger crosses zero when the interaction explains 0.9%, 0.8% and 0.9% of the variance in support for resistance, sanctions, and war, respectively. The confidence interval on the ACME estimate for the weak threat's effect on support for strikes via anger — displayed in the top row of Table B14 includes 0 at σ values greater than 0.08, and $\tilde{R}^2$ values greater than 0.009. The counterarguing mechanisms (*chance attack* and *chance Eritreans killed*) and significant alternative mediators (*resolve reputation* and *moral reputation*) exhibit similar degrees of sensitivity, though the strong threat's effect via perceptions of the chance of a terrorist attack remains significant for larger values on the sensitivity parameters than the weak threat.

In general, the sensitivity analyses underscore that identifying ACME estimates in the presence of multiple mediators poses methodological challenges and that small treatment-mediator interactions may render results non-significant. Still, we caution that the small $\tilde{R}^2$ values required to uncover non-significant ACME estimates, relative to the largest possible value on this parameter, are partly a product of the substantial residual variance in models of experimental data. Moreover, evidence from both studies 1 and 2 reveal significant direct effects on our dependent variables and key mechanisms, alongside limited direct effects on mediators associated with alternative explanations. These findings support our theory while the sensitivity analyses suggest that future research should probe reactance mechanisms with alternative experimental designs that directly account for unit heterogeneity.

Mediation Analyses: Support for Invasion

Figure B3 displays the indirect effect estimates for the strong and weak threats on support for invading Eritrea. Although neither treatment had a direct effect on this form of retaliation, we find some evidence for small but significant indirect effects through reactance mechanisms.[4] To the extent that either threat

[4]See Zhao, Lynch Jr and Chen (2010) for a discussion about indirect-only mediation.

Table B13: Sensitivity Analyses for Effect of Strong Threat via Each Mechanism

Mechanism	DV	ACME (Avg) (1)	CI Low (2)	CI High (3)	σ (4)	Max σ (5)	$\tilde{R}^2$ (6)	Max $\tilde{R}^2$ (7)
Anger	Block Arms	0.01	0.004	0.015	0.086	0.852	0.009	0.889
Anger	Sanctions	0.007	0.002	0.012	0.075	0.742	0.008	0.833
Anger	Strikes	0.008	0.002	0.013	0.078	0.778	0.009	0.845
Chance Attack	Block Arms	0.014	0.006	0.022	0.094	0.935	0.009	0.889
Chance Attack	Strikes	0.005	0	0.01	0.086	0.854	0.009	0.845
Chance Eritreans Killed	Block Arms	0.007	0.001	0.012	0.079	0.784	0.009	0.889
Resolve Reputation	Block Arms	0.006	0.001	0.011	0.083	0.829	0.009	0.889
Moral Reputation	Block Arms	0.005	0	0.01	0.081	0.81	0.009	0.889

Note: Columns 1-3 display the estimated ACMEs and 95% bootstrapped confidence intervals for factors that significantly mediated the strong threat's effect in study 2, estimated assuming the homogeneous interaction assumption using the mediation package in R (Tingley et al., 2014). Column 4 displays the value of the sensitivity parameter *sigma* at which the confidence interval for the ACME would include 0 and represents the sensitivity with respect to interaction heterogeneity (Imai and Yamamoto, 2013, 162). Column 5 displays the maximum possible value for *sigma*. Column 6 displays the value for $\tilde{R}^2$ at which the ACME estimate bounds would include 0, and column 7 displays the largest possible value for $\tilde{R}^2$. $\tilde{R}^2$ indicates how much variance the treatment-mediator interaction must explain to render the ACME non-significant.

Table B14: Sensitivity Analyses for Effect of Weak Threat via Each Mechanism

Mechanism	DV	ACME (Avg) (1)	CI Low (2)	CI High (3)	σ (4)	Max σ (5)	$\tilde{R}^2$ (6)	Max $\tilde{R}^2$ (7)
Anger	Strikes	0.005	0	0.011	0.081	0.807	0.009	0.843
Chance Attack	Block Arms	0.007	0	0.014	0.097	0.966	0.009	0.892

Note: Columns 1-3 display the estimated ACMEs and 95% bootstrapped confidence intervals for factors that significantly mediated the strong threat's effect in study 2, estimated assuming the homogeneous interaction assumption using the mediation package in R (Tingley et al., 2014). Column 4 displays the value of the sensitivity parameter *sigma* at which the confidence interval for the ACME would include 0 and represents the sensitivity with respect to interaction heterogeneity (Imai and Yamamoto, 2013, 162). Column 5 displays the maximum possible value for *sigma*. Column 6 displays the value for $\tilde{R}^2$ at which the ACME estimate bounds would include 0, and column 7 displays the largest possible value for $\tilde{R}^2$. $\tilde{R}^2$ indicates how much variance the treatment-mediator interaction must explain to render the ACME non-significant.

increased anger or perceptions about the chance that allowing arms sales wold increase the chance of a terrorist attack, participants expressed more support for invading Eritrea to take over the country. Similar to its indirect effect on support for military strikes, we again find that concerns about the U.S. moral reputation suppressed the effect of threats on support for retaliation.

Product-of-coefficients Estimates

Finally, we present indirect effects estimates from a product-of-coefficients analysis. The results bear substantial similarity to the ACME estimates derived from our main analyses. We present these results to illustrate that our conclusions are robust to an alternative estimation strategy; indeed, our main analyses present the most conservative estimates.

We calculated the indirect effects by multiplying the direct effect of each treatment on the mechanism by the mechanism's effect on the outcome, and bootstrapping 95% confidence intervals (Preacher and Hayes, 2008). Table B15 presents the effect of each mediator on the three dependent variables affected by the treatments while controlling for the treatments, other mechanisms, and a panel of pre-treatment controls. Figure B4 presents the indirect effect estimates

Figure B3: Mechanisms of Threats' Effects on Support for Invasion, Compared to Natural Costs

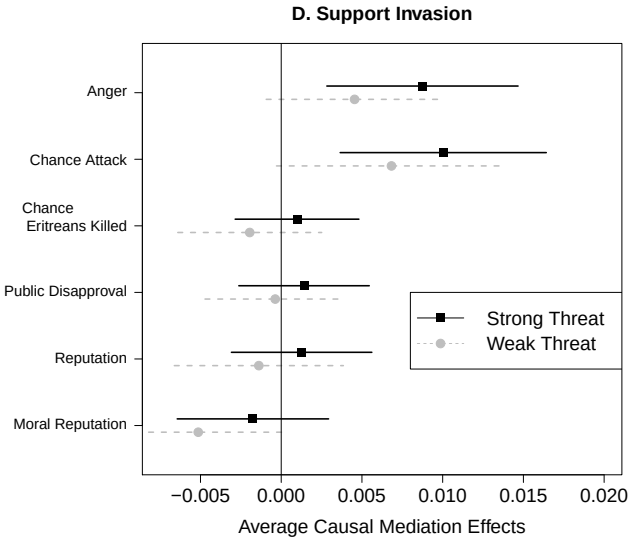
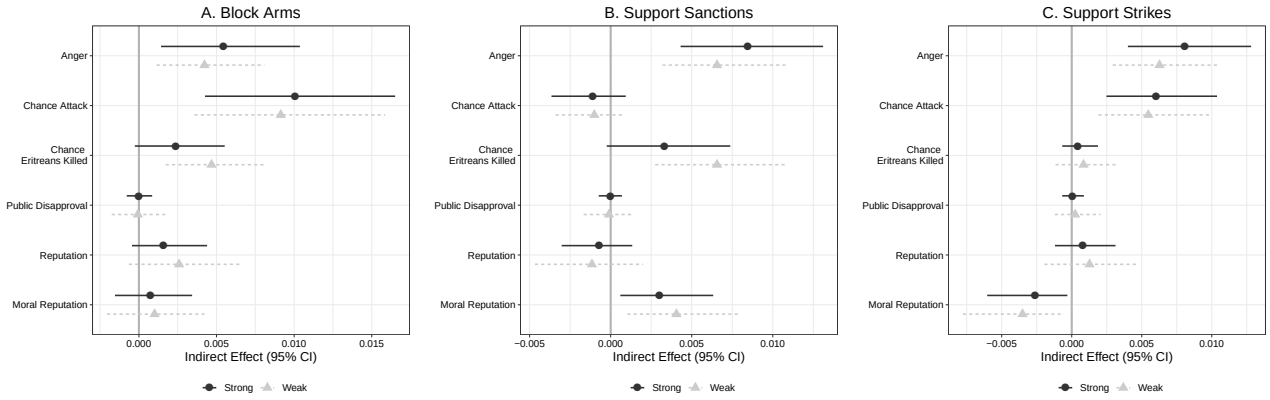


Figure B4: Mechanisms of Coercion, Resistance, and Retaliation



Note: Indirect effects relative to natural costs (bootstrapped 95% confidence intervals). Models include pre-treatment controls.

B.7 Study 2 Results for Full Sample

The results in the manuscript exclude people who failed a pre-treatment instructional manipulation check question or attention screen. Failing the instructional manipulation check ended the survey, but we include those from batch 1 who failed the attention screen (spending less than 2.1 seconds reading the opening vignette pre-treatment) to show that our results are robust to including these inattentive respondents. Weak threats have slightly weaker effects in this sample — which aligns with our expectation that reactance-induced behavior increases with stronger, more dogmatic language.

Table B15: Effect of Mediators on Dependent Variables

	Block Arms	Sanctions	Strikes	Invade
	(1)	(2)	(3)	(4)
Control	0.137**	0.005	−0.011	−0.023
	(0.016)	(0.014)	(0.014)	(0.015)
Strong Threat	0.027	0.036**	0.039**	0.005
	(0.015)	(0.013)	(0.014)	(0.014)
Weak Threat	0.046**	0.007	0.008	0.004
	(0.015)	(0.013)	(0.014)	(0.014)
Anger	0.061**	0.095**	0.091**	0.086**
	(0.023)	(0.020)	(0.021)	(0.021)
Chance U.S. Attack	0.229**	−0.025	0.137**	0.209**
	(0.025)	(0.022)	(0.023)	(0.023)
Reputation for Resolve	0.047	−0.021	0.023	0.001
	(0.030)	(0.026)	(0.027)	(0.028)
Moral Reputation	0.020	0.082**	−0.071*	−0.077**
	(0.030)	(0.026)	(0.027)	(0.028)
Public Disapproval	−0.002	−0.004	0.009	0.035
	(0.027)	(0.024)	(0.025)	(0.025)
Chance Eritreans Killed	0.118**	0.165**	0.021	−0.015
	(0.027)	(0.024)	(0.025)	(0.026)
Controls	✓	✓	✓	✓
Constant	0.467**	0.235**	0.067	0.001
	(0.041)	(0.036)	(0.038)	(0.039)
N	3,196	3,196	3,196	3,196
R^2	0.106	0.165	0.154	0.172

*p < .05; **p < .01

Note: Models display OLS coefficients. All dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is the natural costs condition. Additional controls include militarism, isolationism, gender, race, age, party identification, census region, income, education, and political knowledge, suppressed for space.

B.8 Study 2 Results for Attentive Subsamples

We assess the treatment effects in two attentive subsamples from the data to evaluate the extent to which our results are robust to varying the inclusion criteria. First, we can assess the treatment effects among only the most attentive respondents by restricting our sample to the second batch of Lucid respondents, which included a pre-treatment attention check to screen out the least attentive participants. The results provide strong and consistent support for our theory. Both the strong and weak threats increased support for blocking arms sales and issuing sanctions, but only the strong threat increased support for military strikes against Eritrea.

Finally, the survey included one post-treatment factual manipulation check to gauge comprehension. We found that 58% of our respondents correctly answered this question. Although we find strong evidence that threats increased scores on our reactance scale — an important subjective manipulation check that corresponds to standard metrics used in reactance research in psychology and communication studies — we can also probe whether our results hold among participants who paid the closest attention to detail while reading the vignette. Figure B5 shows that indeed, threats induce resistance and retaliation in this subset of attentive respondents. Indeed, the two primary differences between this subset and the full sample suggest stronger support for our hypotheses — we observe larger treatment effects in general, and larger differences between the strong threat and weak threat in support for retaliation. Still, conditioning on posttreatment variables could produce biased estimates (Montgomery, Nyhan and Torres, 2018), and we report the intent-

Table B16: Treatment Effects (Full Sample)

	Block Arms		Support Sanctions		Support Strikes		Support Invasion	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Control	0.142**	0.147**	0.022	0.019	−0.004	−0.004	−0.024	−0.016
	(0.015)	(0.016)	(0.014)	(0.013)	(0.014)	(0.014)	(0.015)	(0.014)
Strong Threat	0.047**	0.053**	0.046**	0.050**	0.041**	0.045**	0.012	0.018
	(0.014)	(0.014)	(0.013)	(0.012)	(0.013)	(0.013)	(0.014)	(0.013)
Weak Threat	0.061**	0.069**	0.029*	0.029*	0.018	0.020	0.009	0.017
	(0.014)	(0.015)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.013)
Controls		✓		✓		✓		
Constant	0.646**	0.684**	0.585**	0.401**	0.456**	0.158**	0.402**	0.103**
	(0.010)	(0.037)	(0.009)	(0.032)	(0.009)	(0.033)	(0.010)	(0.034)
N	3,911	3,602	3,888	3,602	3,888	3,602	3,888	3,602
R^2	0.023	0.037	0.004	0.137	0.004	0.126	0.002	0.137

*p < .05; **p < .01
Note: Models display OLS coefficients. All dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is the natural costs condition. Additional controls include militarism, isolationism, gender, race, age, party identification, census region, income, education, and political knowledge, suppressed for space.

Table B17: Treatment Effects, Second Batch Only

	Block Arms		Sanctions		Strikes		Invasion	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Control	0.187**	0.189**	0.032	0.031	−0.029	−0.029	−0.016	−0.015
	(0.025)	(0.025)	(0.023)	(0.022)	(0.023)	(0.023)	(0.025)	(0.024)
Strong Threat	0.065**	0.063**	0.066**	0.053**	0.055**	0.054**	0.036	0.038
	(0.022)	(0.023)	(0.021)	(0.020)	(0.021)	(0.021)	(0.022)	(0.021)
Weak Threat	0.075**	0.079**	0.042*	0.043*	0.009	0.024	0.016	0.034
	(0.023)	(0.023)	(0.021)	(0.020)	(0.022)	(0.021)	(0.023)	(0.022)
Controls		✓		✓		✓		✓
Constant	0.627**	0.701**	0.590**	0.439**	0.441**	0.156**	0.355**	−0.044
	(0.016)	(0.061)	(0.015)	(0.053)	(0.015)	(0.055)	(0.016)	(0.057)
N	1,470	1,384	1,465	1,384	1,465	1,384	1,465	1,384
R^2	0.037	0.075	0.007	0.139	0.010	0.143	0.004	0.156

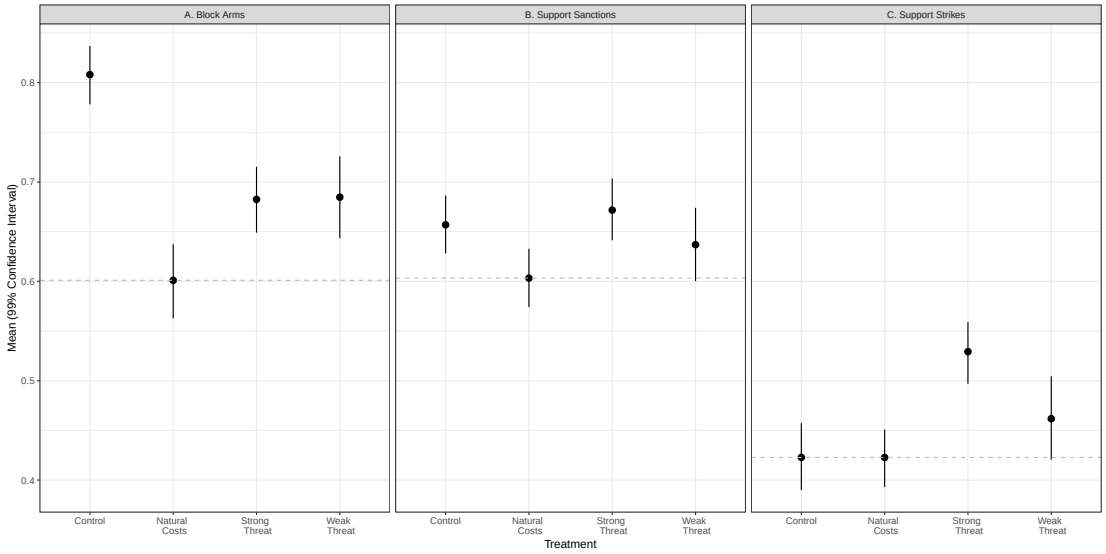
*p < .05; **p < .01
Note: Models display OLS coefficients. All dependent variables and continuous independent variables rescaled to range from 0 to 1. Reference category for treatment dummies is the natural costs condition. Additional controls include militarism, isolationism, gender, race, age, party identification, census region, income, education, and political knowledge, suppressed for space.

to-treat results in the manuscript.

C Anxiety does not mediate the effect of threats on coercion failure

Psychological reactance proposes that threats cause defiance via intertwined emotional and cognitive pathways, marked by anger and counterarguing. We found substantial support for these propositions across our

Figure B5: Outcome Variables by Treatment — Passed Comprehension Check



Note: Figure displays the mean with 99% bootstrapped confidence intervals in each of the four treatment groups for our three main dependent variables.

two experiments. We also controlled for and measured several alternative mechanisms associated with the IR literature on coercion failure, largely ruling them out as causes for coercion failure in the two experiments.

Yet political psychology opens the door to another possible path through which emotions might affect how people respond to threats — the degree to which threats cause anxiety. The appraisal tendency framework treats anxiety as a separate discrete emotion from anger, for example (Lerner and Keltner, 2000). Whereas anger is an approach emotion that encourages people to take risks and confront threats (Skitka et al., 2006), anxiety predicts risk aversion and can prompt support for *inaction*. In that respect, any indirect effects through anxiety on our dependent variables should point in the opposite direction from anger. On the other hand, anxiety also corresponds to openness. Anxious people surveil their environment for information, hoping to reduce their uncertainty about the anxiety-inducing situation. If threats increase anxiety more than natural costs, people might be more willing to change their minds and justify non-compliance and aggression (Marcus, Neuman and MacKuen, 2000).

We conduct an additional mediation analysis to examine whether anxiety contributes to coercion failure. Evidence that anxiety either has no indirect effect on our dependent variables — or that its effect points in the opposite direction from anger — would lend additional confidence in our argument that reactance mechanisms drive coercion failure in our experiments.

We turned to a single item included in our emotions battery as a distractor to test this possibility: the extent to which participants report feeling "anxious." Notably, we did not design our study to test a theory about anxiety and therefore lack a multi-item scale. As with our other analyses, we rescaled this variable to range from 0 to 1 and estimated the average causal mediation effect assuming causal dependence between anxiety and the set of alternative mediators using the mediation package in R (Tingley et al., 2014; Imai and Yamamoto, 2013). Moreover, these analyses are exploratory, and not pre-registered for study 1.

The results provide no evidence that anxiety explains resistance to threats nor support for the various forms of retaliation in our experiments. We find only negative signs on the statistically significant indirect effects for anxiety: In study 1, 95% confidence intervals for the threat's effect mediated via anxiety all contain 0. In study 2, anxiety has a significant *negative* mediating effect on support for sanctions comparing the strong threat to natural costs ($ACME = -0.005$ $[-0.009, -0.001]$). These findings, available in the replication materials, suggest at most that anxiety-inducing threats might lead people to favor capitulation rather than resistance. This finding comports with a body of research applying the appraisal tendency framework to foreign policy public opinion and coercion.

D Summary of Expectations and Findings

Table D18: Summary of Main Effect Expectations and Findings, Relative to Natural Costs

	Resistance	Supported?	Sanctions	Supported?	War/Strikes	Supported?	Invasion	Supported?
Study 1:								
Threat/reputation	+	✓	+	✓	+	✓	NA	NA
Threat/no reputation	+	✓	+	✓	+	✓	NA	NA
Natural Costs/reputation	+	no (null)	+	no (null)	+	✓	NA	NA
Study 2:								
Strong threat	+	✓	+	✓	+	✓	+	no (null)
Weak threat	+	✓	+	✓	+	✓	+	no (null)
Strong minus weak	+	no (null)	+	no (null)	+	✓	+	no (null)

Note: Plus signs indicate expected positive treatment effect, relative to the natural costs comparison group. Null indicates that we anticipated no effect. In study 1, comparing the threat/no reputation treatment to the natural costs/no reputation group provides the most important test of our theory and shows that the threat on its own increased resistance via anger and counterarguing without raising reputation or audience concerns.

Table D19: Summary of Mechanism Expectations and Findings, Relative to Natural Costs

	Anger	Supported?	Counterarguing	Supported?	Resolve Reputation	Supported?	Public Disapproval	Supported?
Study 1:								
Threat	+	✓(indirect on retaliation)	+	✓(via threat exaggerated)	+	✓	+	mixed
Reputation	null	✓	null	✓	+	✓	+	mixed
Study 2:								
Strong threat	+	✓	+	✓	null	mixed	null	✓
Weak threat	+	✓(indirect on retaliation)	+	✓(indirect on resistance)	null	mixed	null	mixed

Note: Plus signs indicate expected positive direct effect on mechanism *and* expected indirect effect via the mechanism on each outcome variable, relative to the natural costs comparison group. Null indicates that we anticipated no effect. Across both studies, we find widespread support for reactance mechanisms mediating the effect of threats. Although both threats had small direct effects on reputation for resolve in study 2, most ACME estimates were non-significant.

References

Aronow, Peter M., Josh Kalla, Lilla Orr and John Ternovski. 2020. "Evidence of Rising Rates of Inattentiveness on Lucid in 2020.".
URL: *https://osf.io/preprints/socarxiv/8sbe4/*

Baron, Reuben M and David A Kenny. 1986. "The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." *Journal of Personality and Social Psychology* 51(6):1173–1182.

Bullock, John G and Donald P Green. 2021. "The Failings of Conventional Mediation Analysis and a Design-Based Alternative." *Advances in Methods and Practices in Psychological Science* 4(4):25152459211047227.

Burleigh, Tyler, Ryan Kennedy and Scott Clifford. 2018. "How to screen out VPS and international respondents using Qualtrics: A protocol." *Available at SSRN* .

Dafoe, Allan, Baobao Zhang and Devin Caughey. 2018. "Information equivalence in survey experiments." *Political Analysis* 26(4):399–416.

Dillard, James Price and Lijiang Shen. 2005. "On the nature of reactance and its role in persuasive health communication." *Communication Monographs* 72(2):144–168.

Gadarian, Shana Kushner. 2014. Beyond the water's edge: Threat, partisanship, and media. In *The Political Psychology of Terrorism Fears*, ed. Samuel Justin Sinclair and Daniel Antonius. New York: Oxford University Press pp. 67–84.

Hall, Marissa G, Paschal Sheeran, Seth M Noar, Kurt M Ribisl, Laura E Bach and Noel T Brewer. 2016. "Reactance to health warnings scale: development and validation." *Annals of Behavioral Medicine* 50(5):736–750.

Imai, Kosuke and Teppei Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis* 21(2):141–171.

Lerner, Jennifer S and Dacher Keltner. 2000. "Beyond valence: Toward a model of emotion-specific influences on judgement and choice." *Cognition & Emotion* 14(4):473–493.

Marcus, George E, W Russell Neuman and Michael MacKuen. 2000. *Affective Intelligence and Political Judgment*. University of Chicago Press.

Montgomery, Jacob M, Brendan Nyhan and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62(3):760–775.

Preacher, Kristopher J and Andrew F Hayes. 2008. "Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models." *Behavior Research Methods* 40(3):879–891.

Skitka, Linda J, Christopher W Bauman, Nicholas P Aramovich and G Scott Morgan. 2006. "Confrontational and preventative policy responses to terrorism: Anger wants a fight and fear wants" them" to go away." *Basic and Applied Social Psychology* 28(4):375–384.

Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele and Kosuke Imai. 2014. "Mediation: R package for causal mediation analysis." *Journal of Statistical Software* 59:1–38.

Zhao, Xinshu, John G Lynch Jr and Qimei Chen. 2010. "Reconsidering Baron and Kenny: Myths and truths about mediation analysis." *Journal of Consumer Research* 37(2):197–206.