Reputational Sources of the Accommodation Dilemma

Michael Goldfien[*] Ryan Powers[†] Tyler Pratt[‡] Jonathan Renshon[§]

December 5, 2024

Abstract

Recent years have been marked by a surge in non-cooperative behavior in world politics, but current theories of international cooperation do not adequately explain the incentives and motivations of states as they formulate responses to hostile behavior. We make two contributions. Conceptually, we argue that, while much work on international politics has demonstrated that beliefs about "type" can facilitate and sustain cooperation, there is a dark side to reputation. Conditional on bad behavior—such as abrogating a treaty—reputation shapes the "accommodation dilemma" faced by states in two ways, first through actual damage to the violator's reputation for cooperation and through observers' concerns about their own reputation for toughness should they be seen to "go easy" on offending states. Empirically, we present data from descriptive surveys of IR scholars showing suggestive evidence of both of our posited reputational mechanisms and helping to rule out other potential explanations. We also field a pre-registered conjoint experiment which demonstrates that (1) past behavior has an outsized influence on the probability and cost of future cooperation (2) this is true for both aggrieved parties and observers, suggesting the importance of reputation broadly and (3) these costs potentially flow through the two reputational mechanisms posited in our theory. By design, our conjoint also allows us to benchmark our theory against numerous other potential explanations for cooperation and accommodation.

[*] Assistant Professor, Department of National Security Affairs, US Naval War College. ✉: michael.goldfien@usnwc.edu 🏠: https://mgoldfien.com. Personal views, do not reflect those of the US Navy or Department of Defense.

[†] Assistant Professor, Department of International Affairs, School of Public and International Affairs, University of Georgia. ✉: ryan.powers@uga.edu 🏠: https://ryanpowers.net

[‡] Assistant Professor, Department of Political Science, University of North Carolina at Chapel Hill. ✉: tbpratt@unc.edu 🏠: https://tylerbpratt.com

[§] Board of Visitors Professor of Political Science, Department of Political Science, University of Wisconsin-Madison. ✉: renshon@wisc.edu. 🏠: http://jonathanrenshon.com. Authors listed in alphabetical order. All contributed equally.

In the last decade, the international system has witnessed a surge in high-profile and public non-cooperative behavior.[1] In his first term, Donald Trump withdrew the United States from the Paris Climate Agreement, the Trans-Pacific Partnership, Intermediate-Range Nuclear Forces Treaty, and the Iran nuclear deal, all while questioning the utility of international governance more broadly. Rodrigo Duterte aggressively attacked international human rights institutions, going so far as to threaten the arrest of UN investigators and withdraw the Philippines from the International Criminal Court. And, in a slow but doggedly committed fashion, a series of short-lived British governments made the United Kingdom the first state to exit the European Union. These high-profile examples dovetail with an emerging empirical literature suggesting that violations, abrogations and exits occur more often than conventional wisdom would suggest: 34% of alliances end via "opportunistic abrogation" (Leeds and Savun, 2007), states have left IOs 200 times since 1945 (Von Borzyskowski and Vabulas, 2019) and a full 20% of states abandon their coalition partners during wartime (Weisiger, 2016, 756).

Prevailing models of international cooperation argue that such non-cooperative behavior should generate meaningful costs for the offending state (Keohane, 1984; Axelrod, 1984; Martin, 1992; Simmons, 2000). Yet, if states are violating agreements with more frequency than our models suggest, it is also true that they seem to face a more varied set of responses than our theories suggest as well. When the Trump administration declared its abandonment of the North American Free Trade Agreement, Mexico and Canada responded not with punishment but by accommodating US demands in negotiations over the successor U.S.-Mexico-Canada Trade Agreement. However, other US non-cooperative behavior on trade – such as the imposition of tariffs on steel and aluminum – elicited retaliatory tariffs by both US allies and adversaries.

Other anecdotal evidence suggests similarly puzzling variation in responses to non-cooperation. Russia's 2014 invasion and annexation of Crimea, for example, provoked retaliatory sanctions and widespread exclusion from cooperative arrangements. In comparison, Russia's military aggression and seizure of Georgian territory in 2008 generated a much weaker international response. More systematically, von Borzyskowski and Vabulas (2025) show that there were 387 withdrawals from IOs between 1945-201, and that roughly half of those states *later rejoined the same IO*, suggesting

[1] We follow Dellmuth and Walter (2024) in defining non-cooperative behavior as a case in which "one state or a group of states unilaterally change[s] the costs and benefits associated with the status quo relations with another country, group of countries, or [international organizations]."

that the accommodation of violators happens with some frequency, theories of costs be damned.

This set of stylized facts motivates the "accommodation dilemma": *how do foreign audiences decide whether to punish or accommodate non-cooperation*? We argue that two distinct but concurrent reputational dynamics are key to understanding the politics of accommodation. First, non-cooperative behavior damages the reputation of the violating state. As the conventional wisdom suggests, states that reject international obligations earn a reputation for being untrustworthy and unlikely to comply with their commitments. We expect this reputational mechanism to materialize even among audiences that are not directly victimized by the non-cooperative behavior. The intensity and longevity of this reputation cost, however, may vary. We contend that a key moderating factor is whether audiences attribute a state's antagonistic behavior to a rogue leader or to an enduring shift in state-level preferences (Renshon, Dafoe and Huth, 2018).

Second, state responses to non-cooperation are shaped by their own strategic reputational concerns: the fear that accommodating violators will signal weakness and invite additional defections (Dellmuth and Walter, 2024). When this concern is highly salient — for instance, when responses to noncompliance are likely to be publicized or closely scrutinized by other states — we expect states to impose costly penalties for non-cooperative behavior. The hard line that some European states took during post-Brexit negotiations with the United Kingdom, for example, was motivated in part to deter other EU member states from considering leaving the institution.

To investigate the accommodation dilemma empirically, we draw on several sources of data across three different research designs. Each method addresses the costs and incentives of accommodating offending states from a different perspective. First, we field an original survey on scholars of International Relations that provides suggestive evidence of our two reputational mechanisms, as well as helping to rule out plausible alternatives such as a concern that cooperation wouldn't produce benefits for the states involved.

Second, we field a novel conjoint experiment that explores variation in citizens' preferred response to non-cooperative behavior. We present respondents with a hypothetical international agreement with a foreign state, varying the partner's past record of compliance as well as a large set of contextual features (e.g., the country's regime type, the ease of detecting violations, and whether the previous leader has been replaced).

We find that violators suffer a large penalty—$3 - 6x$ the size of next largest factor in our analysis—in both public willingness to forge new agreements and demands over the terms of those agreements. We also find that the consequences of violating agreements obtain even for those who only observe the violation, suggesting a key role for reputation. The conjoint results also provide indirect support for our specific reputational mechanisms: replacing a leader attenuates the effect of violation, consistent with a story in which the bad reputation of the state is partially reset through leader turnover and a story in which other states worry less about the reputational consequences of accommodating a violator since they can plausibly claim that the state has been reformed. Finally, we propose a parallel encouragement survey experiment to assess the two reputational mechanisms described above. We have not yet fielded this experiment and welcome feedback on its structure.

1 The Consequences of Non-Cooperation in World Politics

A long tradition in international relations research examines the motivations of states as they consider cooperative relationships with others. Early realist scholarship pointed to enduring characteristics of human nature (Morgenthau, 1948) or the international system (Waltz, 1979) that make cooperation difficult to reach and sustain in general. Others identified variables at the system, state, or issue area level that shape the relative attractiveness of cooperative arrangements. Jervis (1976, 1978), for example, argues that environments with greater transparency improve states' ability to detect cheating, increasing their willingness to cooperate with partners that may fail to comply. Fearon (1998) similarly identifies variation in concerns about cheating, arguing that the stakes of defection differ across policy domains. The presence of international regimes or institutions is believed to facilitate cooperation by clarifying joint expectations, lowering transaction costs, and overcoming information asymmetries (Keohane, 1984; Axelrod and Keohane, 1985; Oye, 1986; Martin and Simmons, 1998). Others point to domestic political institutions, often arguing that democratic states are particularly attractive partners for cooperation (Mansfield, Milner and Rosendorff, 2002; Milner and Kubota, 2005; Lipson, 2013).

For many IR scholars, the dominant factor driving the choice to cooperate is the past behavior of a state's potential cooperative partner (Keohane, 1984; Axelrod, 1984; Simmons, 2000; Sartori, 2002; Tomz, 2008; Crescenzi et al., 2012; Weisiger and Yarhi-Milo, 2015). Violations of international commitments provide a signal about the likelihood that a potential partner will stick to its word.

This reputational effect may be stronger when commitments are embedded in international law, which further clarifies expectations and generates a deeper sense of obligation (Abbott et al., 2000; Guzman, 2008). As a result, breaking or abandoning international legal agreements damages the reputation of the offending state and often triggers costly backlash, exclusion, retaliation, or material penalties (Martin, 1992; Chaudoin, 2014; Schmidt, 2023).

Despite the general consensus that non-cooperative actions should make a state a less attractive partner in future cooperative endeavors, the empirical record is mixed (Downs and Jones, 2002). Some hostile behaviors (e.g., North Korea's abandonment of the Nuclear Non-Proliferation Treaty (NPT) in 2003) trigger costly sanctions and expulsion from cooperative arrangements. Others (e.g., North Korean NPT violations in the early 1990s) generate an accommodative, rather than a punitive, response from other states. While recent scholarship has begun to grapple with variation in third-party responses (Renshon, Dafoe and Huth, 2018; Morse and Pratt, 2022; Donahue and Crescenzi, 2023), we have few general theories about the circumstances under which non-cooperation will engender punishment or accommodation. Dellmuth and Walter (2024) offer one such framework, conceptualizing responses to non-cooperation as an "accommodation dilemma" in which states must balance the loss of cooperative gains against the risk that acquiescence invites further hostile behavior.

What role does reputation play in shaping how states manage the accommodation dilemma? Conventional accounts emphasize the reputation of the offending state, explaining how other actors rationally update beliefs and adjust behavior following non-cooperation (e.g., Guzman, 2008). However, as Dellmuth and Walter (2024) note, victims and observers of non-cooperation are also likely to be worried about protecting their own reputations. Our paper focuses on the interaction of these distinct reputational mechanisms.

1.1 How Dual Reputational Dynamics Shape Responses to Non-Cooperation

We seek to explain how reputational concerns drive foreign audience responses to non-cooperative foreign policy behavior. Given our interest in reputational mechanisms, we focus on non-cooperative behavior that includes two key elements: it must *(1) represent a shock to expectations (i.e., a departure from the state's prior behavior)*, and *(2) reflect a clear hostility toward international commitments.* This behavior does not have to constitute a change in formal commitments (though it can); any action that is sufficiently conspicuous and hostile to cooperative commitments can qualify. An example is a U.S. leader publicly questioning whether the United States would come to defense of NATO countries.

Specifically, we examine how non-cooperative behavior shapes the perpetrator state's ability to secure cooperative agreements in the future. In general, high-profile defections from international commitments should make it more difficult to find willing cooperative partners. States that do enter into cooperative arrangements with the perpetrator are likely to require more demanding conditions to guard against the possibility of withdrawal or noncompliance. In other words, non-cooperative behavior should increase both the availability and the price of future cooperation for antagonistic states.

Formally, consider a scenario in which countries i and j interact in view of an audience, k, over some initial period of time t_1. If i engages in non-cooperative behavior at some point in t_1—for example, by brazenly violating a bilateral agreement with j—we expect country i to find fewer cooperative partners and confront more demanding terms in the future, both from j and from k.

In this environment, the reactions of foreign audiences (j and k) determine the intensity of the costs that the offending state i will face. As these actors formulate a response to non-cooperative behavior, we argue their motivations are influenced by two distinct reputational dynamics. These dynamics reflect the actual damage to the i's reputation, as well as concerns that j and k might have that cooperating with i will damage their own reputation for toughness (in being seen to accommodate a state that had engaged in bad behavior).

Ex-post damage to the offender's reputation The first reputational mechanism we examine occurs when non-cooperative behavior damages the offender's reputation for honoring international commitments. In the wake of hostile behavior, states update their expectations about the offender's likelihood of future compliance and adjust their behavior accordingly. This often means excluding the offender from cooperative endeavors where monitoring is difficult or non-compliance is particularly costly. It may also take the form of imposing more rigorous conditions on future agreements with the country.

This first mechanism is consistent with the conventional wisdom of reputation costs following violations of international commitments (Keohane, 1984; Goldsmith, 2005; Tomz, 2008). It is triggered when a foreign audience learns about non-cooperative behavior by an offending state. As a result, we expect reputational effects to emerge among *both* direct victims of non-cooperation as well as third-party observers. This expectation is supported by recent experimental findings (e.g., Chen, Pevehouse and Powers, 2023; Morse and Pratt, 2024) but at odds with with some observational analyses that find costly responses are limited to victimized parties (Schmidt, 2023).

Ex-ante concern for third party reputations In addition to the realized damage to the offender's reputation, non-cooperation triggers a potential effect on the reputation of responding states. Foreign audiences have their own strategic reputational concerns about appearing weak in the face of antagonistic behavior. Accommodating non-cooperation signals weakness and could encourage further undesirable behavior by the offender or other states (Dellmuth and Walter, 2024). More generally, failing to respond to provocations by offending states may weaken a country's perceived resolve (Bloch and McManus, 2024).

This second mechanism provides an additional reason for actors to respond harshly to non-cooperation. When concerns about their reputation for toughness are salient, states are likely to take particularly visible punitive actions toward the offender. This may occur when responses to noncompliance are publicized or likely to be closely scrutinized by other states

Reputational concerns of responding states may be most intense among direct victims of non-cooperation. However, even observer states often want to ensure that non-cooperation does not set a damaging precedent that could undermine beneficial patterns of cooperation or make themselves look weak. Powerful states that are looked upon as enforcers of global cooperation may be especially sensitive to these concerns, though, alternatively, some states may be so powerful that they have *fewer* concerns about damaging their reputation for toughness.

We summarize the main empirical implications of these dynamics in Table 1. The first two rows reflect the general expectation that non-cooperation by country i will make it more difficult to forge cooperative arrangements with both the victim (j) and observers (k) of non-cooperation. Rows 3 and 4 reflect the two reputational mechanisms we argue drive these broader outcomes.

Moderators Focusing on the perpetrator state's reputation further suggests two moderating variables that should condition the effect of non-cooperative behavior on future cooperation costs.

Empirical implication #	*Prediction about:*	
1	outcome	i's non-cooperative behavior (aimed at j) at t_1 increases the price of cooperation j will demand from i at t_2
2	outcome	i's non-cooperative behavior at t_1 increases the price of cooperation that observers, k, will demand from i at t_2
3	mechanism	i's non-cooperative behavior raises its cost for future cooperation through the mechanism of damaging its reputation for fulfilling commitments
4	mechanism	i's non-cooperative behavior raises its cost for future cooperation by triggering the reputational concerns of future partners, who worry that accommodation of i will damage their reputation for toughness.
5	moderator	effect of i's non-cooperative behavior (on cooperation cost) will depend upon the visibility of i's behavior
6	moderator	effect of i's non-cooperative behavior (on cooperation cost) will depend upon observer judgments about whether i's behavior reflect underlying preferences of domestic populace

Table 1: Empirical Implications of the theory

First, since reputations require publicity (Dafoe, Renshon and Huth, 2014), how far the reputational damage extends—and who demands higher prices for future cooperation—depends on the visibility of the non-cooperative behavior. In a completely private interaction between i and j, only i's reputation *in the eyes of j* can be affected. In a completely public interaction, i's behavior toward j is observed by the audience k, and their broader reputation (in the eyes of the international community) is affected. We therefore expect that the effect of i's bridge-burning behavior (on cooperation cost) will depend upon the *visibility* of i's behavior.

Second, the effect of non-cooperation will depend on the reputational inferences drawn by both victims and observers. Both parties—j and the larger audience k—must assess the state's underlying cooperative type. This requires drawing accurate inferences about whether the state's non-cooperative posture will endure. In practice, this is a complex and difficult inferential task. Consider two stylized cases of non-cooperation. In the first case, non-cooperative behavior, especially from states that have generally upheld their cooperative commitments in the past signals a sharp and long-lasting change in the underlying policy preferences of the state. In the second case, non-cooperation reflects the preferences of a rogue leader whose policies are likely to be reversed in a future period. International audiences thus face a signal extraction problem. They observe a single leader's policy choice, but their optimal response depends upon the long-run policy preferences that emerge from the state's political system. To resolve this uncertainty, we argue that observers look for other signals of domestic political support for the leader's actions. In other words, the effect of i's non-cooperative behavior will depend upon observer judgments about whether i's behavior reflects underlying preferences of its domestic populace.

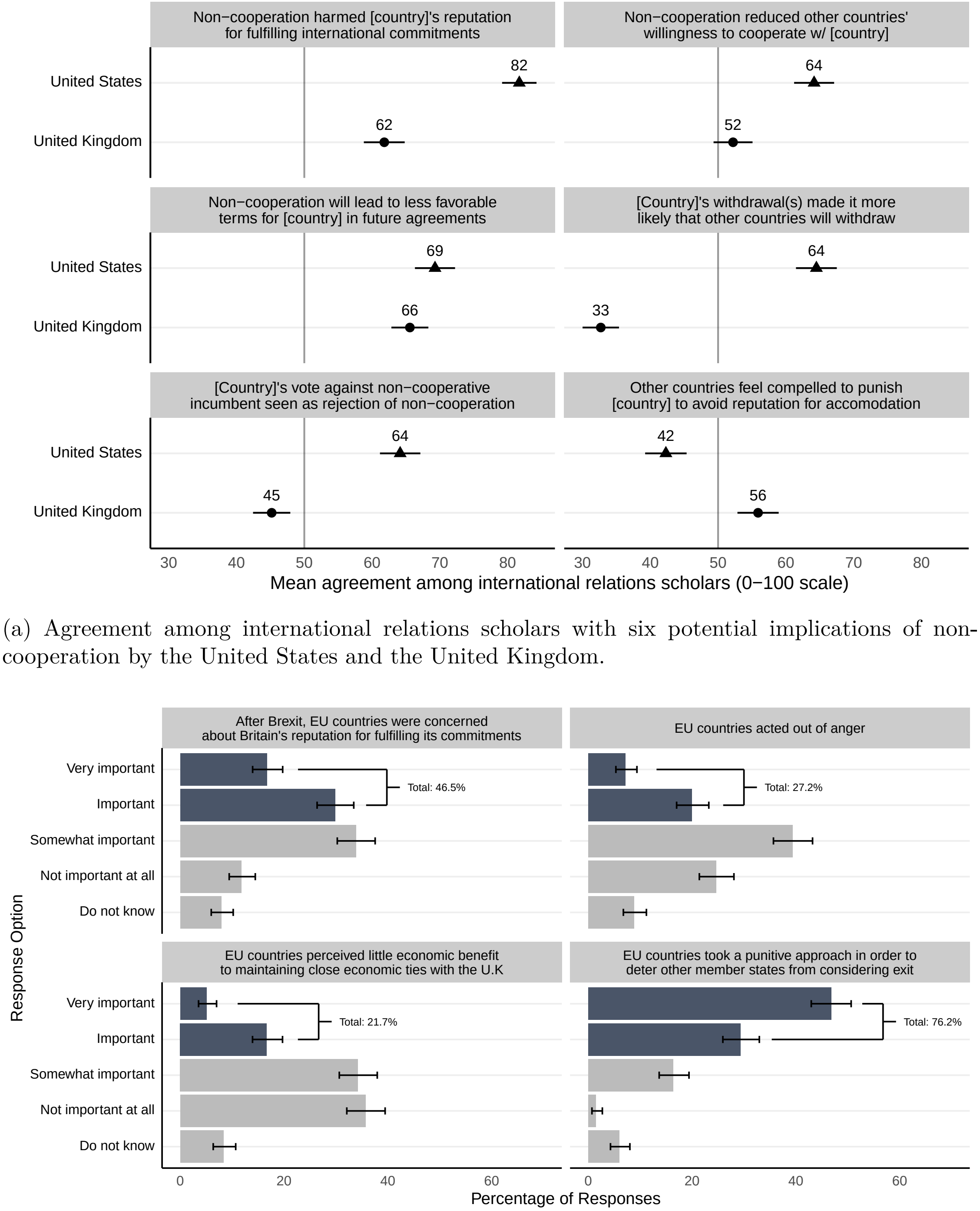
2 Descriptive Survey Data

As an initial, descriptive examination of our theoretical expectations, we embedded three questions in a Teaching, Research, and International Policy (TRIP) survey in September, 2024.[2] The TRIP survey solicits the opinion of international relations experts employed in political science departments or public policy schools in the United States. While not public or elected officials, these scholars do have requisite domain-specific knowledge to be considered elites in one important sense of the term (Kertzer and Renshon, 2022) and to allow us a first cut at evaluating our theory of the reputational dynamics involved in accommodation. The three questions focused on recent non-cooperative behavior on the international stage, specifically (1) U.S. withdrawals from international agreements in recent years, (2) the United Kingdom's withdrawal from the EU, and (3) the EU's response to Brexit. Figure1 (a) and (b) displays our key results.

We take three lessons from our descriptive survey questions. First, our respondents believed that leaving agreements stifles future cooperation, both by reducing the willingness of others to cooperate and by downgrading the terms under which agreements are offered. Roughly two-thirds of our sample (64%; see Figure 1 (a)) believed that recent withdrawals by the United States had reduced the willingness of other countries to enter into agreements with them while 69% stated that these same withdrawals have made it more difficult for the U.S. to negotiate favorable terms. Similar dynamics were at work in the UK case, where roughly the same amount (66%) believed that Brexit had led the EU to take a hard line against Britain in subsequent negotiations.

There is also support for both mechanisms at work in our theory. With respect to our first posited mechanism—that non-cooperative behavior harms the reputation of the violating state—there is overwhelming (82%) support among our respondents for the belief that recent withdrawals

[2] See survey text in Appendix A.1.



(a) Agreement among international relations scholars with six potential implications of non-cooperation by the United States and the United Kingdom.

(b) Perceived importance of factors motivating EU response to Brexit among international relations scholars

Figure 1: **Descriptive Survey Results from IR Scholars ($N \approx 670$).** Data from TRIP Snap Poll XXII fielded in late October 2024.

have harmed the U.S.' reputation for fulfilling its international commitments. There is also strong support (62%) for the same dynamics at work with Britain's reputation following Brexit. Finally, there is support for the second mechanism from our theory, namely that other states might be concerned about the potential harm to their reputation for toughness if they were to accommodate the U.S./UK following their withdrawals. While only 41% of respondents agreed with this logic in the U.S. case (perhaps as a result of either self-serving biases or outsized U.S. power), the results were far stronger for the UK (see Figure 1 (b)). 56% of respondents agreed that EU countries *have* felt the need to punish and condemn Brexit specifically in order to avoid developing a reputation for accommodating non-cooperation and a full 75% believed (i.e., chose this motive as "important" or "very important") that the EU was taking a "punitive approach in order to deter other members from exit."

This descriptive data helps sets the stage for the causal research designs we describe below. Broadly, our survey data from subject matter experts provide suggestive evidence of both the main relationship between withdrawals and future cooperation as well as the two reputational mechanisms that might underlie it. Simultaneously, they help to rule out other potential alternatives. For example, our respondents overwhelmingly (76%) agreed that deterring future states from exiting the EU was important or very important in driving responded to Brexit. Similarly almost half (47%) of respondents rated concerns about the UK's reputation for fulfilling commitments as important or very important. There was less support, however, for the notion that EU states took a hard-line approach either out of anger (26%) or reduced benefits to close economic ties with Britain (22%).

3 Causal Research Designs

To explore the causal effects of non-cooperative behavior, we turn to survey experimentation. Our goal is to test the empirical implications listed in Table 1. In particular, we will focus on the argument that a state's non-cooperative behavior increases the price of cooperation that observers will demand in a future period and that it does so through the mechanisms of (1) actual reputational damage to the violator state and (2) concern over reputation to the observer state.

3.1 Estimands

Table 2 captures our estimands, using a simple version of the framework suggested by Lundberg, Johnson and Stewart (2021). Theoretical estimands (τ) are the "questions outside of the data" and are a combination of a (1) unit specific quantity (a realized or potential outcome) and (2) a target population. The unit-specific quantity clarifies whether the object is to make descriptive or causal inferences, and the target population addresses the question: over whom or what do we aggregate that unit-specific quantity (Lundberg, Johnson and Stewart, 2021, 534)? The theoretical estimand τ is in some cases—for example in row 2—a difference in potential outcomes, and thus inherently unobservable.

The empirical estimand (θ) takes into account real world constraints and focuses only on observed quantities. For example, in row 1, our empirical estimand (θ) is informative of our theoretical estimand under the identification assumptions of the particular method. Here, that means randomization of Country i's history of keeping or abrogating agreements. We list empirical estimands for two different experimental designs: a conjoint study and a parallel encouragement factorial design. The value in explicitly stating these quantities is greater clarification about what research design is optimal, what sources of data ought to be used, and most importantly, what assumptions we must make in order to connect our theoretical to our empirical estimands.

4 Conjoint Experiment

Our goal is to estimate the how an actor's non-cooperative behavior affects their attractiveness as a cooperative partner in the future, with a focus on the role played by the two reputational mechanisms: the violator's own reputation for cooperation and the cooperative partner's concern for their reputation for toughness. We consider this study to be an initial, exploratory analysis of the role that reputation may play in linking past non-cooperative behavior to the prospects for future cooperation, as well as other factors that may inhibit cooperation or accommodation.

Two additional features of the experiment are of note. First, despite a host of work on the effects of different types of violation and responses to it, none that we are aware of have experimentally manipulated whether violations occurred alongside a significant number of other features in a highly powered design. Instead, some vary framing of, or responses to, violations (e.g., Chilton,

Question	**Theoretical Estimand (τ)**		**Empirical Estimands (θ)**	
	unit-specific quantity	*target pop.*	*Conjoint exp*	*Factorial exp*
What is the effect of “burning bridges” on costs of future cooperation?	causal: effect of a country violating an agreement on their prospects for future cooperation	all respondents (no restrictions)	average marginal component effect (AMCE) of “past behavior” attribute (levels: brazenly violated/rigorously complied) on support for cooperation with hypothetical Country **A** in Prolific sample.	average treatment effect (ATE) of Country **A**’s treaty abrogation (compared to no information and keeping commitments) on respondents’ “willingness to accept” price for cooperation in U.S. online convenience sample
Does burning bridges decrease support for cooperation through *actual damage to the violator’s reputation for fulfilling commitments*?	causal mechanism: Portion of total effect of violation on cooperation that goes through damage to the violator’s reputation for keeping commitments	all respondents (no restrictions)	conditional AMCE: interaction between *past behavior* attribute and *leadership turnover* attribute in online conjoint study using Prolific sample.	(“eliminated effect”) Difference between: (1) ATE: burning bridges (versus keeping commitments) under no reputation information (natural mediator arm) and (2) ACDE: same effect when the reputation is set to “poor reputation” in U.S. online convenience sample
Does burning bridges decrease support for cooperation through potential partners’ *concern for their reputation for toughness?*	causal mechanism: Portion of total effect of violation on cooperation that goes through (prospective) “concern for reputation for toughness”	all respondents (no restrictions)	conditional AMCE: interaction between *past behavior* attribute and *secret agreement* attribute in online conjoint study using Prolific sample.	(“eliminated effect”) Difference between: (1) ATE: burning bridges (versus keeping commitments) under no info about publicity condition (natural mediator arm) and (2) ACDE: same effect when the publicity is set to “secret agreement ” in U.S. online convenience sample

Table 2: **Theoretical and Empirical Estimands**

2014; Chaudoin, 2014; Chilton, 2015; Strezhnev, Simmons and Kim, 2019; Morse and Pratt, 2022; Tingley and Tomz, 2022), identity of the victims (Cohen and Powers, 2024) or related features such as past reputation (Donahue and Crescenzi, 2023; Powers, 2024). Few experimentally manipulate violations themselves (though, for examples, see Lupu and Wallace, 2019; Kim, 2019) and those that do rarely if ever consider more than small handful of features.[3] The advantage of our conjoint design is that we are able to calibrate the importance of violations of commitments in relation to a host of other theoretically important attributes as well as estimate how different features attenuate or amplify the effects of commitment violations (for which, one needs to experimentally vary the violation itself and power the design for interactions). Second, the results from this conjoint experiment inform our second experiment where we focus more directly on causal mechanisms.

In our conjoint experiment, we measure the preferences of American respondents over a cooperative agreement between the United States and a hypothetical Country A (Brutger et al., 2022, 2023). While there are trade-offs in some cases with single-profile designs compared to paired (see Hainmueller, Hangartner and Yamamoto, 2015) , they are still widely used (e.g., Huff and Kertzer, 2018; Jost and Kertzer, 2023; Goldfien, Joseph and McManus, 2023) and in this case is the design that accords with how respondents would encounter the information in the real world (we are rarely presented with two entirely different international agreements and asked what we would prefer). Figure 2 visualizes the survey flow, Table 4 lists our pre-registered outcomes and Appendix B contains details of information presented to respondents and randomization scheme.[4]

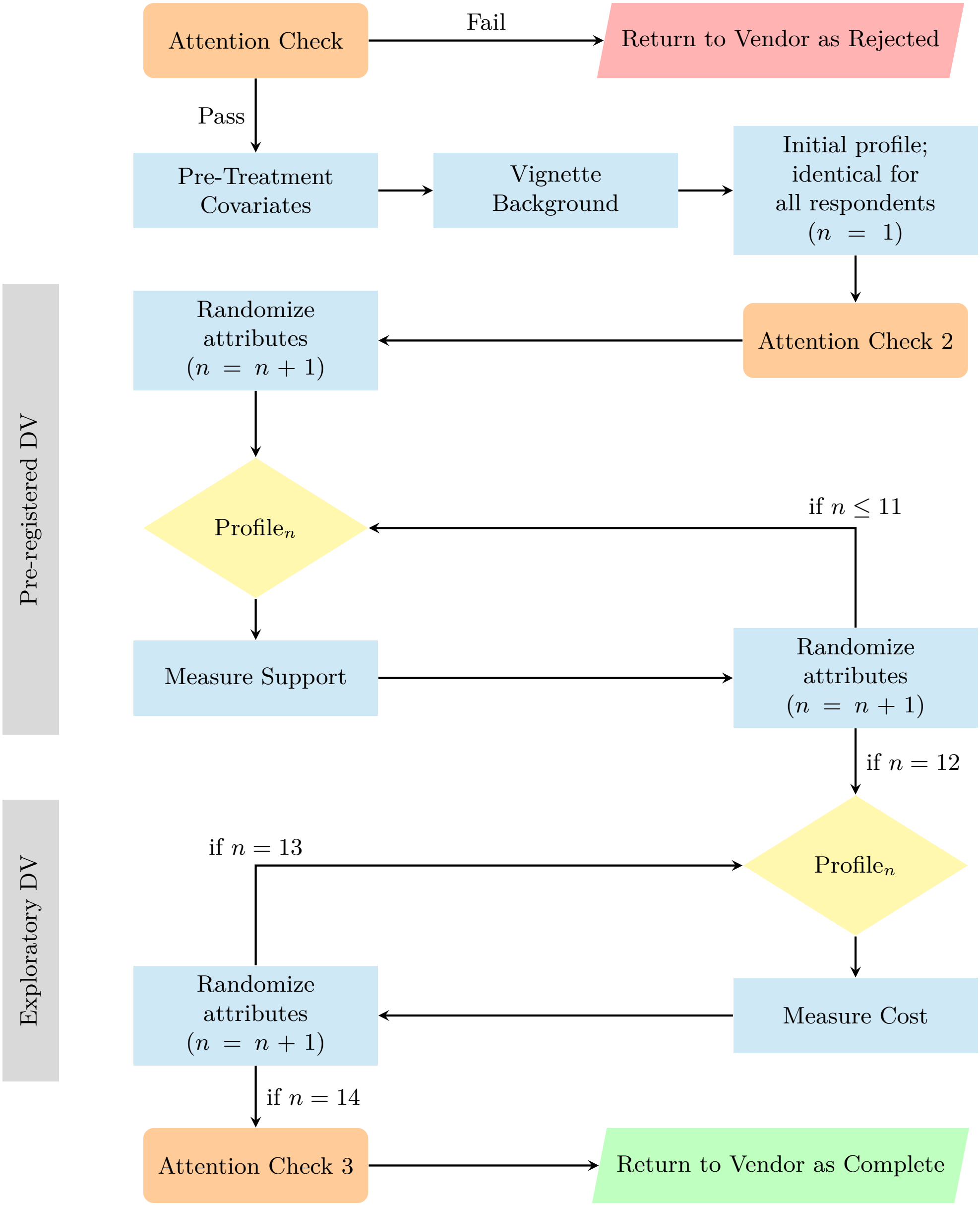
4.1 Logistics and Design

Attention We measured attentiveness both before, during and after the conjoint experiment. Prior to the experiment, we employ a traditional attention check question; those that fail this pre-treatment attention check will be removed from the survey. To measure attention *during* the conjoint (pre-treatment), we employ Kane and Costa's (2024) method of asking three factual questions about the vignette after a first task that does *not* vary across respondents. The correct answers are thus the same across all respondents and can be combined into a continuous (additive)

[3] Though conjoints have been used to address, for instance, support for different features of the trans Atlantic partnership (Hahm et al., 2019) or the factors that shape accommodation preferences towards the UK following Brexit (Jurado, León and Walter, 2022).

[4] A link to our pre-analysis plan can be found here: `https://osf.io/5urvq`. The PAP was amended *prior* to fielding to fix an issue with the wording of our main outcome question.

Figure 2: Consort Diagram for Pre-registered Conjoint Experiment



measure of attention that can be used to test the robustness of our findings without inducing post-treatment bias. To explore whether attention deteriorates substantially over the course of the study (something relevant for single-profile designs; Hainmueller, Hangartner and Yamamoto 2015, 2399), we use the method outlined in Powers (2024) and included a post-treatment attention check based on the *last* conjoint task that respondents complete. Respondents were not removed based on this, and because it's post-treatment, it was not included in main models of AMCEs.

Design The survey focuses on a hypothetical foreign policy scenario (see Appendix B.4) involving a potential cooperative agreement between the U.S. and a hypothetical Country A. We manipulate 15 factors/attributes related to Country A, its previous behavior, the potential cooperative agreement and the international system. Recent work suggests that results of conjoints are relatively stable and not highly contingent on the number of attributes or profiles shown at the same time (Jenke et al., 2021) and that "with respect to the number of attributes, the 'breaking points' of conjoint survey experiments appear to be outside the range of current practice" (Bansak et al., 2021). Each respondent judges 14 profiles (Bansak et al., 2018): the 1^{st} fixes attributes for an attention check, the next 11 feature randomized attributes and are used to test our pre-registered hypotheses, and the last 2 feature a different outcome for exploratory purposes.[5] Table 3 presents the attributes and levels of the conjoint.

Outcomes Our main outcome variable is support for the cooperative agreement with Country A (scaled $0 - 100$). For exploratory purposes, we are also interested in how reputational effects shape the cost of cooperation. It seems plausible that, for many states, the main effect of a bad reputation is not the availability of cooperative partners, but instead in having to agree to more onerous terms in order to secure a cooperative agreement in the first place. These more onerous terms might be denominated in side payments, policy costs, or the distribution of the cooperative surplus. In the final two profiles (#13 & 14), they are asked to design their preferred agreement by using a slider (0-100) to allocate the benefits of the treaty between the United States and Country A. Following previous work (Kertzer, Renshon and Yarhi-Milo, 2021), the order of attributes is randomized across respondents, but held constant for each respondent across all profiles they see

[5] Bansak et al. (2018, 113) find that "we see no significant decline in the core attributes' effects as the number of tasks increases."

in order to facilitate legibility and comprehension.[6]

Table 3: Conjoint Attributes and Levels

	BACKGROUND FEATURES	
	Randomized attribute	**Levels**
Attributes of Country A	(B.1) Size	(1) small (2) large
	(B.2) Economic development	(1) advanced (2) developing
	(B.3) Geographic Region	(1) Latin America (2) Europe (3) Africa (4) Middle East (5) Asia
Attributes of new agreement	(B.4) Agreement is about…	(1) economics (reducing tariffs) (2) environmental (reducing carbon emissions) (3) security (defense spending)
	THEORY-RELEVANT FEATURES	
System-level attributes	(T.1) International system characterized by…	(1) Outsized U.S. power (2) U.S.-China competition (3) IO-led order
Past Behavior of Country A	(T.2) Previous treaty was with…	(1) United States (2) Country B
	(T.3) Did A uphold previous treaty?	(1) rigorous compliance (2) brazen violations
A's Domestic Politics	(T.4) Regime type	(1) democracy (2) autocracy
	(T.5) Leader's fate	(1) same leader (2) new leader with different views
	(T.6) Support for int'l regime?	(1) no add'l info (2) challenge
Attributes of new agreement	(T.7) Publicity of agreement	(1) public & observable (2) confidential & not observable
	(T.8) Agreement would produce…benefits	(1) moderate (2) very significant
	(T.9) Detecting cheating is...	(1) easy (2) hard
	(T.10) Failing to detect cheating will be...	(1) quite costly (2) minimally costly
	(T.11) The treaty...	(1) is open-ended (2) will expire in five years

Note: All dimensions randomized independently across profiles.

[6] See Appendix B.4 for details on how text is presented and randomized.

Recruitment and Statistical Power We recruited a general population sample of 1,800 adults based in the United States via Prolific[7] and fielded the experiment from November 22-26[8], 2024. We arrived at this number by starting with the fact that the most demanding of our pre-registered tests, in terms of power, are the 2 × 2 interaction terms (Hypotheses 2-4 in Table 4). For more details on power calculations, see pre-analysis plan.

4.2 The Direct and Indirect Consequences of Non-Cooperation

As a first cut at estimating the quality of our sample, we consider our attention checks (See Appendix B.2). 94% of respondents passed the initial, pre-treatment check required to stay in the study, validating the relatively high quality of Prolific samples on this dimension. We also find that attention was high for our pre-treatment conjoint-specific attention check ($\mu = 79\%$) that was embedded in the first profile, though attentiveness did decline moderately between the 1^{st} and 13^{th} profile (from 79% to 66%).[9]

The critical importance of past behavior Turning to our substantive results as summarized in Table 4, our pre-registered expectations are largely borne out. We begin with the main effect of past non-cooperative behavior on support for future cooperation. We predicted that non-cooperative behavior would make it harder and more costly for states to secure future cooperative deals. To generate the AMCE needed to test this prediction, we regress support (pre-registered) and reservation price (exploratory) on a complete battery of attribute level indicators using OLS. We present the resulting AMCEs in Figure 3. As per the suggestion in Liu and Shiraito (2023) and as pre-registered, we present AMCEs both with adjustment for multiple hypotheses (Benjamini and Hochberg, 1995) and without.

Averaging over all other manipulations, revealing that a state had "brazenly violated" a past treaty lowers support for a newly proposed treaty with that same state by 20.8 points on our

[7] In a recent comparison of samples, Douglas, Ewell and Brauer (2023) find that "compared to MTurk, Qualtrics, or an undergraduate student samples (i.e., SONA), participants on Prolific and CloudResearch were more likely to pass various attention checks, provide meaningful answers, follow instructions, remember previously presented information, have a unique IP address and geolocation, and work slowly enough to be able to read all the items." In another comparison, Eyal et al. (2021) find that—among Amazon Mechanical Turk, CloudResearch, Prolific, Qualtrics and Dynata— "only Prolific provided high data quality on all measures." See also similar results in Albert and Smilek (2023).

[8] After the first day of fielding, the pay rate was increased from $1.5 to $3 per respondent.

[9] This may overestimate inattention in the last profile since it's possible that the questions concerned factors—e.g., secrecy—that respondents judged to be less important.

$0 - 100$ scale and raised the reservation price by 6 points.[10] We thus find strong support for H_1: past violations make future cooperation harder and more costly to secure. Notably, the main effect of past behavior was the *largest single AMCE recovered* in both our support and our reservation price analyses, exceeding the next largest AMCE—regime type—by a factor of about $3X$ in the case of support and about $6X$ in the case of the reservation price. Though these effect size differences may partially reflect precise treatment wordings, the results nonetheless highlight the salience of past behavior for respondent support of cooperative partnerships.

The other AMCEs are consistent with benchmark theories of cooperation. Support is lower when agreements offer fewer benefits, when cheating is costly and hard to detect, when the partner country is an autocracy, and when the partner country's leader had recently expressed skepticism of international cooperation.

The far-reaching consequences of non-cooperation Next, we turn to more targeted predictions regarding the conditions under which states might be willing to accommodate counterparts with a history of non-cooperative behavior. The first of these concerns whether non-cooperative behavior matters even when it is aimed at a third party, answering the question of whether direct experience with non-cooperation is necessary to trigger consequences or if simply observing it is sufficient. To examine this, we use OLS to regress our outcomes on a complete battery of attribute treatment level indicators and the interaction of past partner ("another country" or "the United States") and past behavior ("brazen violator" or "faithful complier"). We present the resulting conditional AMCE estimates in Figure 4a(a).

When the past non-cooperative behavior is against another country, support for the agreement declines by 19.4 points.[11] While the magnitude of this effect is somewhat larger—22 points—when non-cooperation is directed at the respondent's country, the broader takeaway is clear: the effect of non-cooperative behavior spills over across dyads. We obtain similar results using our exploratory DV: past violations increase the reservation price by a similar magnitude when it is directed at the respondent's country (6 points) as when it is directed against a third country (7 points; see Appendix B). That past violations have roughly similar effects whether directed at the respondent's country or a third party suggests a key role for reputation (as opposed to other factors, such as

[10] Support: 95% CI:20.1, 21.6; $p < .001$; BH adj. $p < .001$.

[11] 95% CI:18.3, 20.6; $p < .001$; BH adj. $p < .001$

anger). That is, we conjecture that public and high-profile violations damage states' reputations for cooperation generally, and this seems to have dramatic effects on the willingness of other states to agree to new cooperative deals.

Two distinct reputational mechanisms Our final two hypotheses relate to two factors interactions that are meant to indirectly test the plausibility of our reputational mechanisms: the actual damage to Country A's reputation for fulfilling commitments and the concerns of respondents over their reputation for toughness. To investigate these, we focus on two attributes: the domestic politics of the partner state (via leader turnover) and the features of the new agreement (whether it is secret or public). We interpret the first interaction—past violation × leader turnover—as a bundle of both mechanisms, and the second as implicating only respondents' concerns over their reputation for toughness. We generate these ACIEs in two different ways. First, we regress our outcomes on a complete battery of attribute level treatment indicators while interacting leader turnover ("remained in power", "voted out", or "removed) and secrecy ("secret" and "public") with past behavior ("brazen violator" and "faithful compiler"). We refer to this as our "restricted model" because it implicitly assumes that there are no relevant interactions other than that between past behavior and the other attribute of interest (leader turnover or secrecy). We also present "unrestricted" estimates that relax the "no other relevant interactions" assumption.[12] We present both because they differ in magnitude but are both broadly consistent with how we articulated the relevant contrasts of interest in our pre-analysis plan. We present the results in Figures 4b (leader turnover) and 4c (secrecy).

With respect to leader turnover, our expectation was that the negative effect of past violation would be smaller when followed by leader replacement in the partner state (H_3). The replacement of such a leader, in our story, undermines the inference that non-cooperative preferences are an enduring feature of the partner state as well as provides plausible cover for respondents to cooperate with the state without damaging their own reputation for toughness.

The middle panel of Figures 4b depicts the effect of non-cooperative behavior conditional on leader turnover and the right panel displays the pre-registered interactive quantities of interest.

[12] We are motivated here by Leeper, Hobolt and Tilley (2020). Their discussion focused on sub-groups defined by respondent-level characteristics, our sub-groups are defined by exposure to either secrecy or leader turnover. Ultimately, it may be useful to use the approach outlined in Egami and Imai (2019) to identify the relevant attribute interactions.

Consistent with our expectations, we learn from these estimates that leader turnover can significantly attenuate the effect of past violations. When the cooperative partner is an autocracy, replacing the leader increases support by between $22 - 24$ points on our $0 - 100$ scale (depending on model choice), with similar effects for democracies (range of between $20 - 22$ points).[13] Leader replacement also reduced the average reservation price by 7 percentage points in the autocratic case and 5 percentage points in the democratic case (Appendix B). These results are broadly consistent with our argument that leader turnover can either help to repair a state's reputation or provide cover for potential new partners to accommodate the state without concern for their own reputations.

With regard to the secrecy mechanism in isolation (not bundled with other reputational concerns as in H_3), we expected that confidential or private agreements would lower the negative effect of past violations on respondent support (H_4), providing indirect support of our second posited mechanism: concern for one's reputation for toughness. A secret agreement means that accommodation is less visible, and thus respondent concern for their state's reputation for toughness may be less salient.

Our results provide some evidence consistent with this expectation. The left-most panel of Figure 4c depicts marginal means, showing the now familiar drop in support for a newly proposed deal when the cooperative partner is revealed to have violated in the past, and that, in general, public deals are preferred to private arrangements. The middle panel shows that this effect obtains whether or not the agreement is secret while the right-most column, the ACIE, tells us that the effect of past non-cooperation is mitigated to a small degree by making the agreement secret. Both the estimates from the restricted (1.7 points) and unrestricted (3.2) models are statistically significant using the conventional $p < .05$ cutoff, though our BH corrections causes (only) the estimate from the restricted model to fall below the significance threshold. The effect of secrecy on reservation price was, similarly, in the expected direction but substantively small and not statistically significant.

We take this evidence as consistent with our theory, though it's not direct and other interpretations are possible given the design. For example, while a secret agreement should reduce concerns for respondents that a wider audience would attribute weakness to them, it leaves open the possibility that the violating state would themselves revise their beliefs about respondents' reputation

[13] Autocracies: 95% CI:20.2, 24.1; $p < .001$; BH adj. $p < .001$. Democracies: 95% CI:17.8, 21.6; $p < .001$; BH adj. $p < .001$.

		Expectation	**Test of...**		**Findings**
H_1	main effect (AMCE)	Violating previous agreement reduces support for cooperation with Country A	Core theoretical prediction (IV → DV)	✓	Past violations reduce support for new agreement by 20.8 points on our 0–100 scale. (95% CI: 20.1, 21.6; $p < .001$; BH adj. $p < .001$). See Figure 3.
H_2	conditional AMCE	Violating previous agreement reduces support for cooperation with Country A even when defection is only observed (third party).	Reputational mechanism (broadly defined)	✓	When the past agreement is with “another country,” past violations reduce support for new agreement by 19.4 points on our 0–100 scale. (95% CI:18.3, 20.6; $p < .001$; BH adj. $p < .001$). See Figure 4a.
H_3	interaction effect (ACIE)	Effect of past violation is lower when leadership turnover occurs	both mechanisms in our theory (indirect)	✓	Replacing the leader reduces the magnitude of the past violation effect by 19.7 points when the leader is “voted out” (democracy) and by 22.2 points when the leader is “removed” (dictatorship). Democracies: (95% CI:17.8, 21.6; $p < .001$; BH adj. $p < .001$) Dictatorship: (95% CI:20.2, 24.1; $p < .001$; BH adj. $p < .001$). See Figure 4b.
H_4	interaction effect (ACIE)	Effect of past violation is lower when current agreement is secret	second mechanism in theory (concern over reputation for toughness)	✓/✗	Making the agreement secret reduces the magnitude of the past violation effect by between $1.7 - 3.2$ points, but only significant before BH adjustment. (95% CI:0.1, 3.2; $p = 0.042$; BH adj. $p = 0.056$) See Figure 4c.

Table 4: **Outcomes for pre-registered hypotheses in Conjoint Design**

for toughness. In addition, it may be the case that respondents did not understand how an secret agreement could even be workable, especially over the long run, and so viewed it as largely irrelevant to the broader question of whether they support the agreement or not.

Durable and conditional effects of past behavior on accommodation Two other aspects of the effects of past violations on future accommodation are notable. First, our results suggest that the effects of past non-cooperation are durable and long-lasting. Overall willingness to cooperate is lower for states who have exhibited non-cooperation in the past and replacing the leader, even in a democracy, does not return support for cooperation to the "faithful complier" baseline.

In fact, because of how our treatments are designed, we are likely *over estimating* the extent to which leader turnover attenuates the effect of past violations. Generating an unbiased estimate of the extent to which leader turnover moderates the effect of bad past behavior requires fixing expectations of future compliance in each of the conditions where the cooperative partner was a "faithful complier" in the past. However, in our design, new leaders always had different preferences from the leaders they replaced. The upshot is that the conditional effect of treaty violation that we estimate in Figure 4b is the product both of *increased* expectations of compliance when the past leader was a violator and *decreased* expectations of compliance when the past leader was a complier.[14]

One way to approximate a comparison more directly related to our theory is by using the "faithful complier who remains in power" as the baseline for *all* the other conditions. Doing so yields conditional AMCEs closer to about 19 points in the case of dictatorships ($66 - 47 = 19$) or about 22 points for democracies ($66 - 44 = 22$). This implies that leader replacement eliminates about $\frac{1}{3}$ of the effect of past non-cooperative behavior. In sum, while we find support for our argument that leader turnover dampens the effect of past "bad" behavior, there is reason to believe that the effect of non-cooperation is even more durable than the estimates based on our pre-registered contrasts suggest.

Second, exploratory analyses reveal that reputation and other factors that condition the accom-

[14] Replacing an old leader who was a "brazen violator" with a new leader with different preferences should *increase* expectations of future compliance. When an old leader is a "faithful complier," respondents are likely to view a new leader with "different views on most issues" skeptically, lowering their expectations of future compliance. Our ultimate goal is to estimate how increased expectations of future compliance shape support for the agreement *relative to a baseline in which expectations of future compliance remained high*, but our estimate depicted in Figure 4b does not give us that exact comparison.

modation dilemma do not operate in isolation. Figure 5 plots the "effect of past non-cooperative behavior" conditional on the other conjoint attribute levels. These factors seem to interact in intuitive but—from the perspective of the cooperation literature—under-appreciated ways.

Specifically, Figure 5 shows that the negative effect of past treaty violations on support for a new agreement is smaller when the new agreement offers very significant benefits, when the cost of cheating is low, and when the agreement is time limited. The negative consequences of past violations are also lower for security/trade (compared to environmental) agreements, for dictatorships (compared to democracies) and when the international system is described as being characterized by intense US-China competition. These results are preliminary and should be taken with a grain of salt, but nonetheless suggest a fruitful avenue of future inquiry into the role of reputation in determining the prospects for international cooperation.

5 Parallel Encouragement Design

Our conjoint experiment provided strong evidence on the importance of past behavior in determining respondents' support for future cooperation. Because of our design, the conjoint results are, by definition, robust to variation across many other important dimensions relating to the agreement, the states involved and the international system. Our conjoint design also provided initial, indirect support in favor of both reputational mechanisms broadly and the two specific reputational mechanisms from our theory. Finally, the conjoint study also provides us rich data on which to build a factorial experimental design, highlighting features that we can include to give our vignette verisimilitude and realism without influencing the effects of our treatments.

However, mechanisms are infamously difficult to study (Bullock, Green and Ha, 2010; Green, Ha and Bullock, 2010). Our conjoint study took an indirect approach that was in the spirit of "implicit mediation" (Bullock and Green, 2021): varying treatments in ways that implicate or attenuate particular causal pathways. To the extent that "a wide array of different X-induced changes in M coincide with X-induced changes in Y (Bullock and Green, 2021, 5), the proposed mechanism is more plausible. Below we describe our plans for more direct tests of our theoretical mechanisms using a parallel encouragement design.

5.1 Analytical Framework

Our theoretical mechanisms focus on reputation: our first mechanism is the actual damage to reputations for cooperation sustained by states that violate agreements and our second mechanism is the ex-ante concern for their reputation for toughness that potential accommodators must consider. In our follow-up experiment, we trace the impact of these mechanisms—taking inspiration from Acharya, Blackwell and Sen (2018)—through the use of a parallel encouragement design. In a parallel encouragement design, one fields two studies: in both studies, the treatment and control operate as normal, but in the "fixed/encouraged" mediator study, we provide additional information about the mediator (while in the "natural mediator" study, we do nothing additional). Our experiment is a parallel *encouragement* design because we cannot perfectly manipulate the mediator.[15]

In experiments, one usually focuses on an ATE: for example, the difference, for subject i, in "support for cooperation" with a hypothetical country that had abrogated a treaty versus a country that had kept its international commitments. This quantity represents the total treatment effect. In contrast, the *average controlled direct effect* (ACDE) is the difference in "support for cooperation" between these two countries—one who had abrogated a treaty and the other who had not—when respondents are provided with additional information that fixes the mediator.[16] One might do this, for example, by adding information stipulating that the violator's reputation for fulfilling commitments is very bad. Because that information about the mediator is constant across treatment/control, the treatment effect in this study is termed the ACDE: the part of the total effect that is *not* due to either mediation or interaction with the mediator.

The *eliminated effect*—the difference between the ATE and the ACDE—is the difference between these two quantities: (1) effect of a country that had burned bridges (versus kept commitments) under no reputation information (natural mediator arm) and the (2) the same effect when the reputation is set to "damaged reputation." Eliminated effects are a combination of the causal process (indirect effect) and (2) the causal interaction (sometimes called the "reference interaction").[17] Any

[15] In order to estimate the quantities we care about, we need only add the assumption of monotonicity, which states that there are no "perverse" effects of providing information about the mediator. This would be violated if, for instance, providing information that a country's reputation was damaged by abrogating agreements *boosted* estimates of that country's reputation for cooperation. Either way, imperfect manipulation of the mediator leads to (conservative) underestimates of the true mediating effect.

[16] For a more in-depth treatment, see Glynn (2021, 264-65).

[17] A positive *indirect effect* would suggest that abrogating agreements impacts judgments about reputation and that this changed costs for cooperation. A positive *reference interaction* would imply inferring reputational damage

findings of an eliminated effect in our case would imply that there are either indirect effects of abrogating agreements on the cost of cooperation *through inferences about reputation* or that there are positive interactions between abrogating agreements and reputation.

Figure 6 illustrates what such a design might look like for our second mechanism. In the "encouraged mediator" arm, we attempt to manipulate the mediator via additional text that describes the potential agreement as secret, noting that because it is secret, other countries would not be able to change their beliefs about the United States or its reputation. The total treatment effect is obtained via a difference in means between Arms 3 and 4, while the ACDE would be the difference in means in the "encouraged mediator arms": 1 and 2. The eliminated effect would be equal to: (Arm 3 − 4) − (Arm 2 − Arm 1). Assumptions required to identify the eliminated effect are noted in Appendix C.1.

6 Discussion

This project makes a number of contributions to research on compliance and international reputation. Conceptually, we advance the literature by highlighting two reputational mechanisms simultaneously implicated by non-cooperative acts, distinguishing two pathways through which past violations can undermine future cooperation. We also theorize that the salience of these mechanisms is crucially moderated by domestic political conflict and leader turnover, bringing insights from the international security literature (Renshon, Dafoe and Huth, 2018; Goldfien, Joseph and McManus, 2023; Myrick, 2024) into the realm of international cooperation.

Empirically, the paper situates reputational concerns among many factors thought to influence the attractiveness of cooperation on the international stage. Alongside a partner's past compliance record, our conjoint design manipulated a host of other factors that past work suggests are key factors in shaping demand for cooperation including including the magnitude of the cooperative surplus, ease of monitoring compliance, the cost of partner defection, and partner regime type, agreement duration, and the structure of the international system. The results represent a step forward for the literature in assessing the relative importance of these factors. In addition, we explore whether and how many of these same factors interact to make cooperation more or less

caused a larger direct effect of abrogating agreements.

likely.

Finally, this project sheds light on the implications of populist backlash to the rules-based international order. Dishearteningly, our preliminary results suggest that non-cooperative policies of the sort that have been commonplace under populist leaders can greatly undermine observers' support for future cooperation. However, our results also suggest a more optimistic take on the current moment: leader turnover can substantially offset the reputational harm of a state's past violations. Therefore, if and when internationalist leaders regain power, a return to cooperation may well be on the table after all.

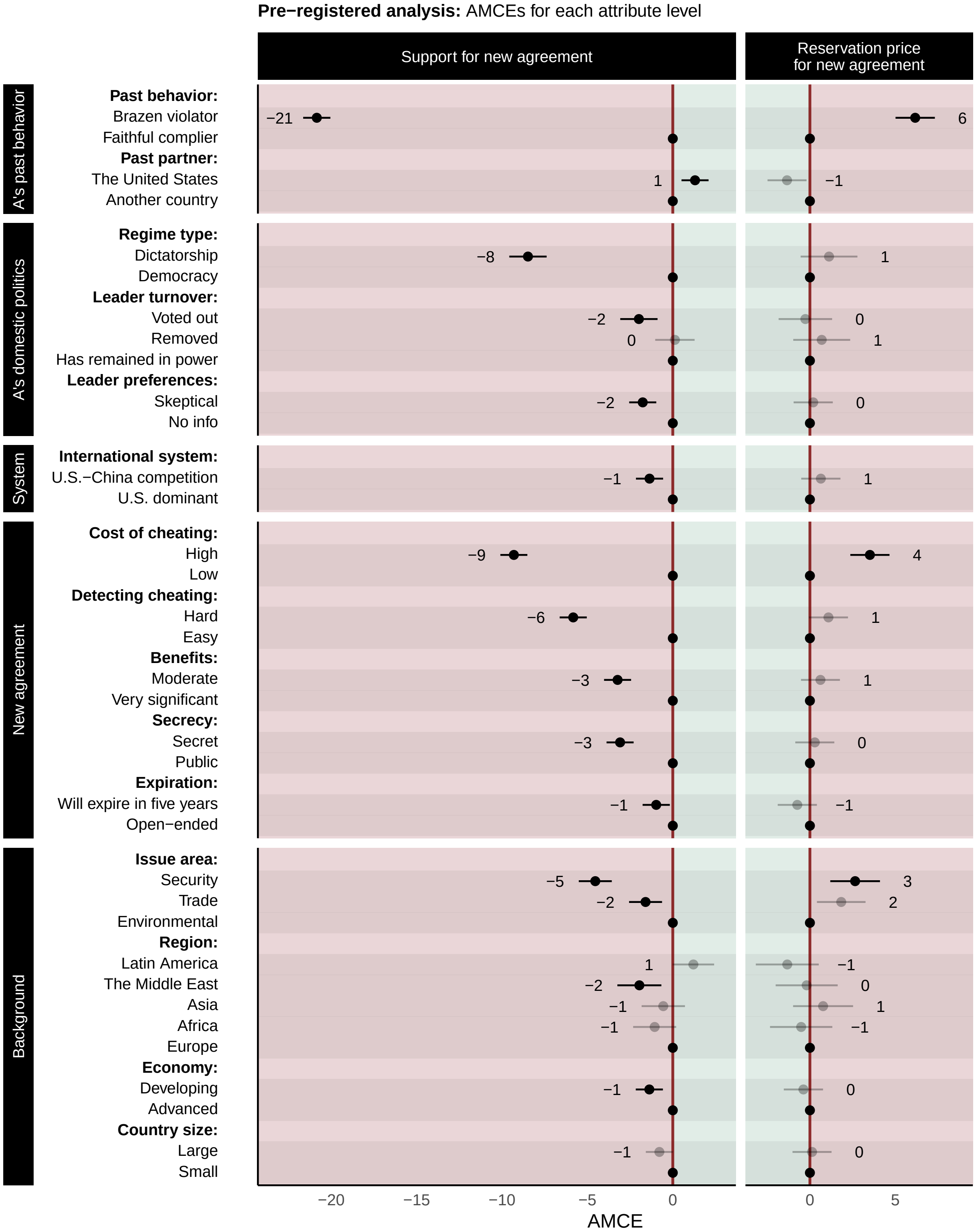
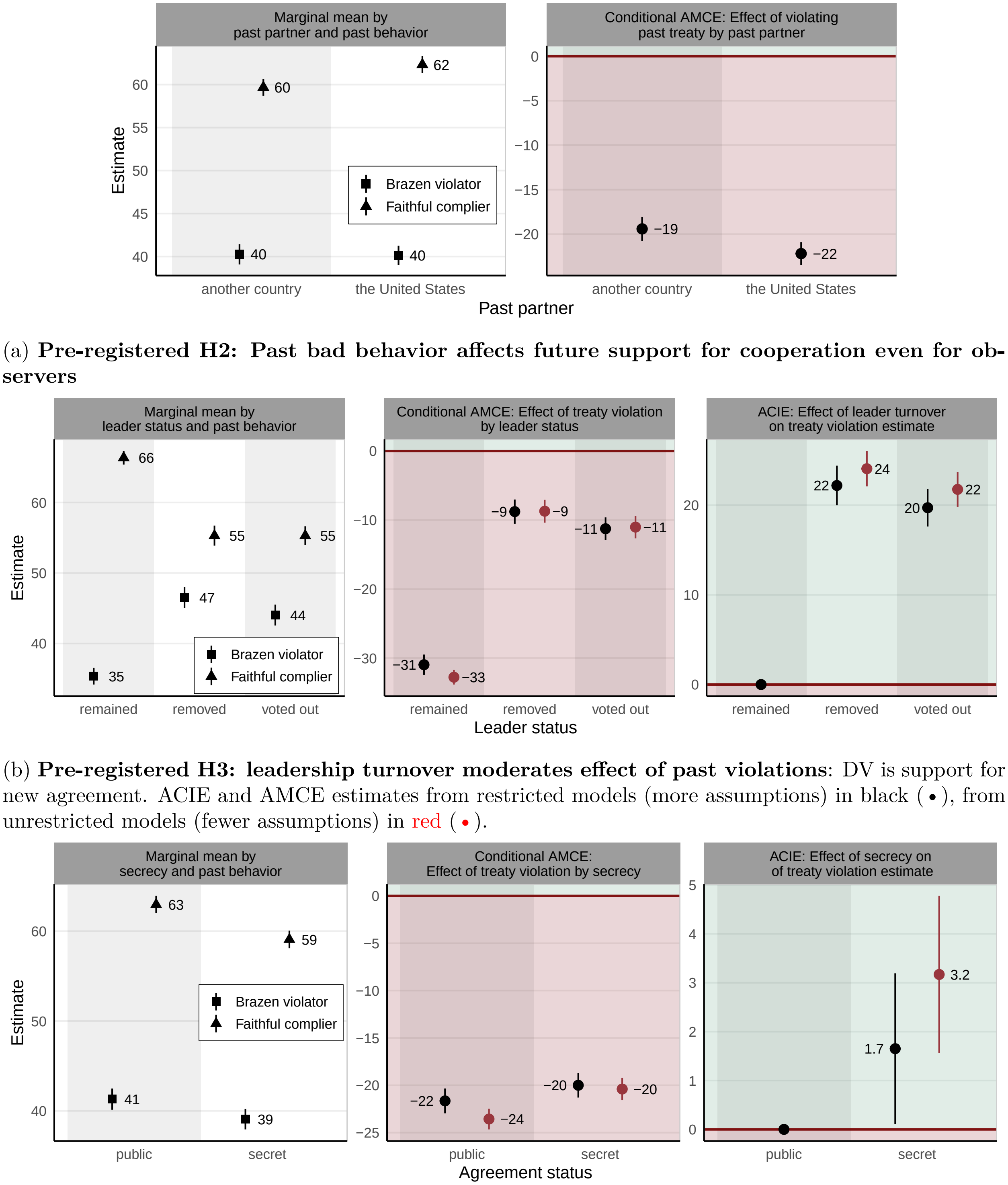


Figure 3: **Average Marginal Causal Effect (AMCE):** a summary measure of the overall effect of an attribute (relative to reference category), averaging over the effect of all other attributes. Black circles (•) indicate that the AMCE remains statistically significant after BH correction for multiple comparisons.



(a) **Pre-registered H2: Past bad behavior affects future support for cooperation even for observers**

(b) **Pre-registered H3: leadership turnover moderates effect of past violations**: DV is support for new agreement. ACIE and AMCE estimates from restricted models (more assumptions) in black (●), from unrestricted models (fewer assumptions) in red (●).

(c) **Pre-registered H4: Secrecy moderates the effect of past treaty violation**: DV is support for new agreement. ACIE and AMCE estimates from restricted models (more assumptions) in black (●), from unrestricted models (fewer assumptions) in red (●).

Figure 4: H2-H4

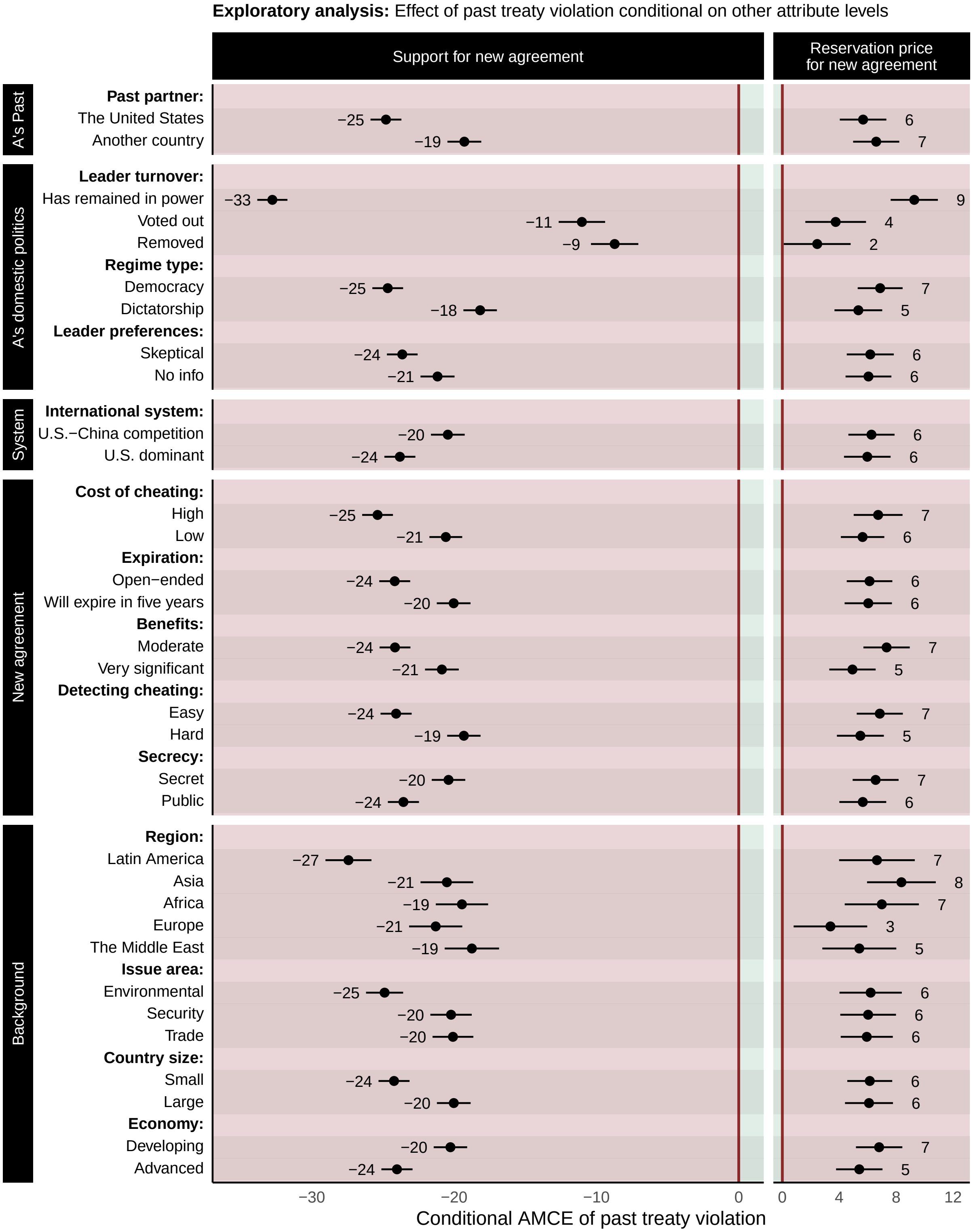


Figure 5: **Conditional Average Marginal Component Effect (cAMCE):** the effect of past treaty violation conditional on other attribute levels. Black circles (•) indicate that the AMCE remains statistically significant after BH correction for multiple comparisons (all estimates do remain significant) .

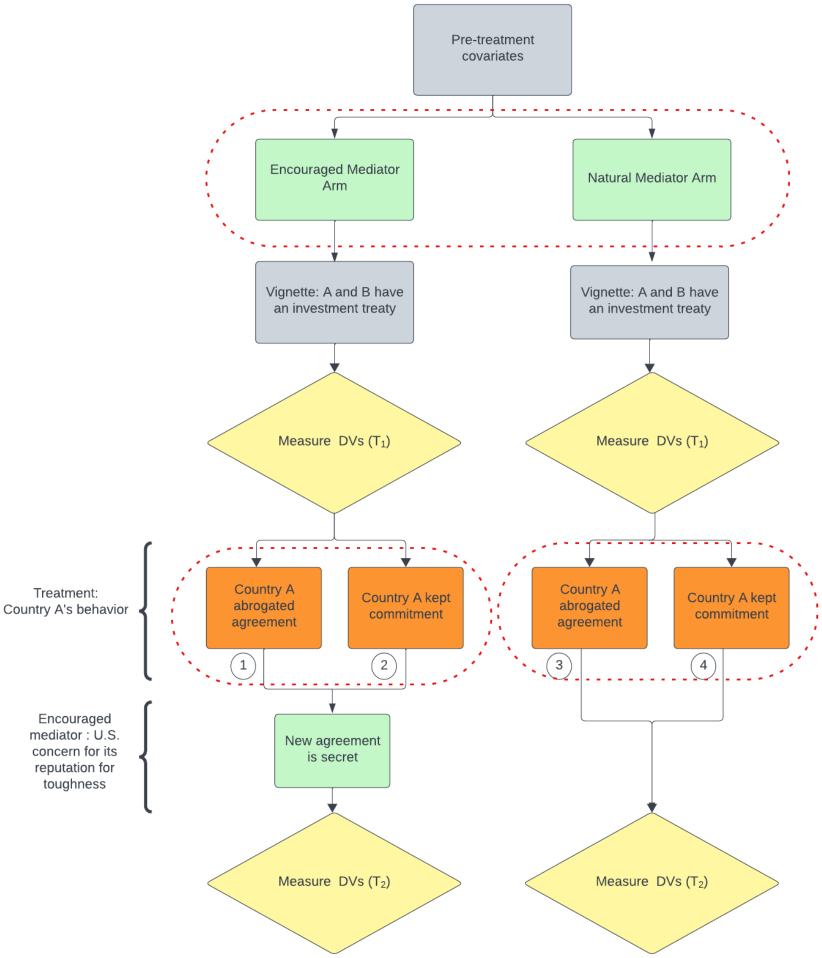

Pre-treatment covariates
Encouraged Mediator Arm
Natural Mediator Arm
Vignette: A and B have an investment treaty
Vignette: A and B have an investment treaty
Measure DVs (T_1)
Measure DVs (T_1)
Treatment: Country A's behavior
Country A abrogated agreement
Country A kept commitment
Country A abrogated agreement
Country A kept commitment
1
2
3
4
Encouraged mediator : U.S. concern for its reputation for toughness
New agreement is secret
Measure DVs (T_2)
Measure DVs (T_2)

Figure 6: Parallel Encouragement Consort Diagram

References

Abbott, Kenneth W, Robert O Keohane, Andrew Moravcsik, Anne-Marie Slaughter and Duncan Snidal. 2000. "The concept of legalization." *International organization* 54(3):401–419.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2018. "Analyzing causal mechanisms in survey experiments." *Political Analysis* 26(4):357–378.

Albert, Derek A and Daniel Smilek. 2023. "Comparing attentional disengagement between Prolific and MTurk samples." *Scientific Reports* 13(1):20574.

Axelrod, Robert. 1984. *The Evolution of Cooperation.* New York: Basic Books.

Axelrod, Robert and Robert O Keohane. 1985. "Achieving cooperation under anarchy: Strategies and institutions." *World politics* 38(1):226–254.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins and Teppei Yamamoto. 2018. "The number of choice tasks and survey satisficing in conjoint experiments." *Political Analysis* 26(1):112–119.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins and Teppei Yamamoto. 2021. "Beyond the breaking point? Survey satisficing in conjoint experiments." *Political Science Research and Methods* 9(1):53–71.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57(1):289–300.

Bloch, Chase and Roseanne W McManus. 2024. "Denying the Obvious: Why Do Nominally Covert Actions Avoid Escalation?" *International Organization* pp. 1–25.

Brutger, Ryan, Joshua D Kertzer, Jonathan Renshon and Chagai M Weiss. 2022. *Abstraction in experimental design: Testing the tradeoffs.* Cambridge University Press.

Brutger, Ryan, Joshua D Kertzer, Jonathan Renshon, Dustin Tingley and Chagai M Weiss. 2023. "Abstraction and detail in experimental design." *American Journal of Political Science* 67(4):979–995.

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.

Bullock, John G, Donald P Green and Shang E Ha. 2010. "Yes, but what's the mechanism?(don't expect an easy answer)." *Journal of personality and social psychology* 98(4):550.

Chaudoin, Stephen. 2014. "Promises or policies? An experimental analysis of international agreements and audience reactions." *International Organization* 68(1):235–256.

Chen, Frederick R, Jon CW Pevehouse and Ryan M Powers. 2023. "Great expectations: the Democratic advantage in trade attitudes." *World Politics* 75(2):316–352.

Chilton, Adam S. 2014. "The Influence of International Human Rights Agreements on Public Opinion: An Experimental Study, 15 Chi." *J. Int'l L* 110.

Chilton, Adam S. 2015. "The laws of war and public opinion: An experimental study." *Journal of Institutional and Theoretical Economics: JITE* pp. 181–201.

Cohen, Harlan and Ryan Powers. 2024. "Judicialization and public support for compliance with international commitments." *International Studies Quarterly* 68(3):sqae078.

Crescenzi, Mark JC, Jacob D Kathman, Katja B Kleinberg and Reed M Wood. 2012. "Reliability, reputation, and alliance formation." *International Studies Quarterly* 56(2):259–274.

Dafoe, Allan, Jonathan Renshon and Paul Huth. 2014. "Reputation and status as motives for war." *Annual Review of Political Science* 17(1):371–393.

Dellmuth, Lisa and Stefanie Walter. 2024. "How Domestic and International Actors Respond to Non-cooperation." *Unpublisehd manuscript* .

Donahue, Bailee and Mark JC Crescenzi. 2023. "Weathering the Storm: Discordant Learning about Reputations for Reliability." *Foreign Policy Analysis* 19(2):orac037.

Douglas, Benjamin D, Patrick J Ewell and Markus Brauer. 2023. "Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA." *Plos one* 18(3):e0279720.

Downs, George W and Michael A Jones. 2002. "Reputation, compliance, and international law." *The Journal of Legal Studies* 31(S1):S95–S114.

Egami, Naoki and Kosuke Imai. 2019. "Causal interaction in factorial experiments: Application to conjoint analysis." *Journal of the American Statistical Association* .

Eyal, Peer, Rothschild David, Gordon Andrew, Evernden Zak and Damer Ekaterina. 2021. "Data quality of platforms and panels for online behavioral research." *Behavior research methods* pp. 1–20.

Fearon, James D. 1998. "Bargaining, enforcement, and international cooperation." *International organization* 52(2):269–305.

Glynn, Adam N. 2021. "Advances in experimental mediation analysis." *Advances in experimental political science* pp. 257–270.

Goldfien, Michael A, Michael F Joseph and Roseanne W McManus. 2023. "The Domestic Sources of International Reputation." *American Political Science Review* 117(2):609–628.

Goldsmith, Jack L. 2005. *The Limits of International Law.* Oxford University Press.

Green, Donald P, Shang E Ha and John G Bullock. 2010. "Enough already about "black box" experiments: Studying mediation is more difficult than most scholars suppose." *The Annals of the American Academy of Political and Social Science* 628(1):200–208.

Guzman, Andrew T. 2008. *How International Law Works: A Rational Choice Theory.* New York: Oxford University Press.

Hahm, Hyeonho, Thomas König, Moritz Osnabruegge and Elena Frech. 2019. "Who settles disputes? Treaty design and trade attitudes toward the Transatlantic Trade and Investment Partnership (TTIP)." *International Organization* 73(4):881–900.

Hainmueller, Jens, Dominik Hangartner and Teppei Yamamoto. 2015. "Validating vignette and conjoint survey experiments against real-world behavior." *Proceedings of the National Academy of Sciences* 112(8):2395–2400.

Huff, Connor and Joshua D Kertzer. 2018. "How the public defines terrorism." *American Journal of Political Science* 62(1):55–71.

Jenke, Libby, Kirk Bansak, Jens Hainmueller and Dominik Hangartner. 2021. "Using eye-tracking to understand decision-making in conjoint experiments." *Political Analysis* 29(1):75–101.

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

Jervis, Robert. 1978. "Cooperation under the security dilemma." *World politics* 30(2):167–214.

Jost, Tyler and Joshua D Kertzer. 2023. "Armies and influence: Elite experience and public opinion on foreign policy." *Journal of Conflict Resolution* p. 00220027231203565.

Jurado, Ignacio, Sandra León and Stefanie Walter. 2022. "Brexit dilemmas: Shaping postwithdrawal relations with a leaving state." *International Organization* 76(2):273–304.

Kane, John V and Mia Costa. 2024. "Being Careful with Conjoints: Accounting for Inattentiveness in Conjoint Experiments." working paper.

Keohane, Robert O. 1984. *After Hegemony.* Princeton University Press.

Kertzer, Joshua D and Jonathan Renshon. 2022. "Experiments and surveys on political elites." *Annual Review of Political Science* 25:529–550.

Kertzer, Joshua D, Jonathan Renshon and Keren Yarhi-Milo. 2021. "How do observers assess resolve?" *British Journal of Political Science* 51(1):308–330.

Kim, Matthew Dale. 2019. "Reputation and compliance with international human rights law: Experimental evidence from the US and South Korea." *Journal of East Asian Studies* 19(2):215–238.

Leeds, Brett Ashley and Burcu Savun. 2007. "Terminating alliances: Why do states abrogate agreements?" *The Journal of Politics* 69(4):1118–1132.

Leeper, Thomas J, Sara B Hobolt and James Tilley. 2020. "Measuring subgroup preferences in conjoint experiments." *Political Analysis* 28(2):207–221.

Lipson, Charles. 2013. *Reliable partners: How democracies have made a separate peace.* Princeton University Press.

Liu, Guoer and Yuki Shiraito. 2023. "Multiple hypothesis testing in conjoint analysis." *Political Analysis* 31(3):380–395.

Lundberg, Ian, Rebecca Johnson and Brandon M Stewart. 2021. "What is your estimand? Defining the target quantity connects statistical evidence to theory." *American Sociological Review* 86(3):532–565.

Lupu, Yonatan and Geoffrey PR Wallace. 2019. "Violence, nonviolence, and the effects of international human rights law." *American Journal of Political Science* 63(2):411–426.

Mansfield, Edward D, Helen V Milner and B Peter Rosendorff. 2002. "Why Democracies Cooperate More: Electoral Control and International Trade Agreements." *International Organization* 56(3):477–513.

Martin, Lisa L. 1992. *Coercive Cooperation: Explaining Multilateral Economic Sanctions.* Princeton, NJ: Princeton University Press.

Martin, Lisa L and Beth A Simmons. 1998. "Theories and empirical studies of international institutions." *International organization* 52(4):729–757.

Milner, Helen V and Keiko Kubota. 2005. "Why the move to free trade? Democracy and trade policy in the developing countries." *International organization* 59(1):107–143.

Morgenthau, Hans. 1948. *Politics Among Nations: The Struggle for Power and Peace.* Alfred A Knopf.

Morse, Julia C and Tyler Pratt. 2022. "Strategies of contestation: International law, domestic audiences, and image management." *The Journal of Politics* 84(4):2080–2093.

Morse, Julia C and Tyler Pratt. 2024. "Smoke and Mirrors: Denials, Norm Challenges, and Contested Noncompliance." *Unpublisehd manuscript* .

Myrick, Rachel. 2024. "Public reactions to secret negotiations in international politics." *Journal of Conflict Resolution* 68(4):703–729.

Oye, Kenneth A. 1986. *Cooperation under anarchy.* Princeton University Press.

Powers, Ryan. 2024. “Is context pretext? Institutionalized commitments and the situational politics of foreign economic policy.” *The Review of International Organizations* pp. 1–29.

Renshon, Jonathan, Allan Dafoe and Paul Huth. 2018. “Leader influence and reputation formation in world politics.” *American Journal of Political Science* 62(2):325–339.

Sartori, Anne E. 2002. “The might of the pen: A reputational theory of communication in international disputes.” *International Organization* 56(1):121–149.

Schmidt, Averell. 2023. “Damaged relations: How treaty withdrawal impacts international cooperation.” *American Journal of Political Science* .

Simmons, Beth A. 2000. “International Law and State Behavior: Commitment and Compliance in International Monetary Affairs.” *American Political Science Review* 94(4):819–835.

Strezhnev, Anton, Beth A Simmons and Matthew D Kim. 2019. “Rulers or rules? international law, elite cues and public opinion.” *European Journal of International Law* 30(4):1281–1302.

Tingley, Dustin and Michael Tomz. 2022. “The effects of naming and shaming on public support for compliance with international agreements: an experimental analysis of the Paris Agreement.” *International Organization* 76(2):445–468.

Tomz, Michael. 2008. “Reputation and the effect of international law on preferences and beliefs.” *Unpublished manuscript* .

Von Borzyskowski, Inken and Felicity Vabulas. 2019. “Hello, goodbye: When do states withdraw from international organizations?” *The Review of International Organizations* 14:335–366.

von Borzyskowski, Inken and Felicity Vabulas. 2025. *Exit from International Organizations.* New York: Cambridge University Press.

Waltz, Kenneth. 1979. *Theory of International Politics.* Addison-Wesley.

Weisiger, Alex. 2016. “Exiting the Coalition: When Do States Abandon Coalition Partners during War?” *International Studies Quarterly* 60(4):753–765.

Weisiger, Alex and Keren Yarhi-Milo. 2015. "Revisiting reputation: How past actions matter in international politics." *International Organization* 69(2):473–495.

Appendix

Table of Contents

A Descriptive Survey of TRIP IR Scholars

A.1 Survey Questions

All respondents completed question block **A**; respondents are randomized into question blocks **B** or **C** with probability .5.

A: After deciding to leave the EU, the United Kingdom sought to renegotiate its economic ties to the EU on more favorable terms. Many EU countries were unwilling to accommodate these demands and took a hard line in post-Brexit negotiations. In your view, how important were the following factors to the countries that adopted this initial non-accommodation posture?

(Not important at all, Somewhat Important, Important, Very important, Don't know)

1. After Brexit, EU countries were concerned about Britain's reputation for fulfilling its commitments
2. EU countries took a punitive approach in order to deter other member states from considering exit
3. EU countries acted out of anger
4. EU countries perceived little economic benefit to maintaining close economic ties with the UK.

B: As you know, the UK public voted to withdraw from the European Union in 2016 and UK government completed that withdrawal in 2020. Using the slider below, please indicate your level of agreement with the with the following statements.

A response of 0 indicates no agreement at all and a response of 100 indicates total agreement.

1. Brexit has harmed the UK's reputation for fulfilling its international commitments.
2. Brexit has reduced other countries' willingness to enter into agreements with the UK
3. Brexit has made it more difficult for the UK to negotiate favorable terms in future international agreements.
4. Brexit has made it more likely that other countries would withdraw from similar agreements.

5. The election of the Labor party in 2024 was a rejection of Brexit.
6. Other countries have felt the need to punish or condemn Brexit in order to avoid developing a reputation for accommodating non-cooperation.

- **C:** The US has withdrawn from several international initiatives in recent years, including the Trans-Pacific Partnership, the Joint Comprehensive Plan of Action (Iran Nuclear Deal), the Open Skies Treaty, and the Paris Climate Agreement.

  Using the slider below, please indicate your level of agreement with the with the following statements. A response of 0 indicates no agreement at all and a response of 100 indicates total agreement.

  1. These withdrawals have harmed America's reputation for fulfilling its international commitments
  2. These withdrawals have reduced other countries' willingness to enter into agreements with the US
  3. These withdrawals have made it more difficult for the United States to negotiate favorable terms in future international agreements
  4. These withdrawals have made it more likely that other countries would withdraw from similar agreements
  5. Electing Joseph Biden in 2020 was viewed by the international community as a rejection of the these withdrawals
  6. Other countries may feel the need to punish or condemn U.S. behavior in order to avoid developing a reputation for accommodating non-cooperation

A.2 TRIP Demographics

Table 5: TRIP Demographic Distribution

	TRIP Snap Poll 21	U.S. IR scholar population
Level	**Percentage (n=703)**	**Percentage (n)**
Gender		
Female	25.9% (182)	31.0% (1556)
Male	71.3% (501)	67.3% (3385)
Prefer not to answer	2.8% (20)	1.7% (85)
Rank		
Adjunct	2.1% (15)	4.4% (222)
Assistant Professor	8.1% (57)	11.6% (583)
Associate Professor	30.2% (212)	25.9% (1301)
Emeritus	6.4% (45)	7.6% (379)
Full Professor	45.0% (316)	40.4% (2029)
Lecturer or Senior Lecturer	3.8% (27)	3.1% (158)
Other	3.6% (25)	5.3% (265)
Visiting Instructor/Visiting Assistant Professor	0.9% (6)	1.6% (80)
University type		
National Liberal Arts College	10.6% (71)	11.1% (525)
National Research University	70.0% (467)	66.0% (3128)
Regional Liberal Arts College	3.1% (21)	3.1% (145)
Regional Research University	16.2% (108)	19.8% (940)
Political party		
Democrat	63.4% (446)	-
Independent	23.5% (165)	-
Republican	3.4% (24)	-
Other	4.8% (34)	-
Prefer not to answer	4.8% (34)	-

B Conjoint Experiment

B.1 Additional conjoint results

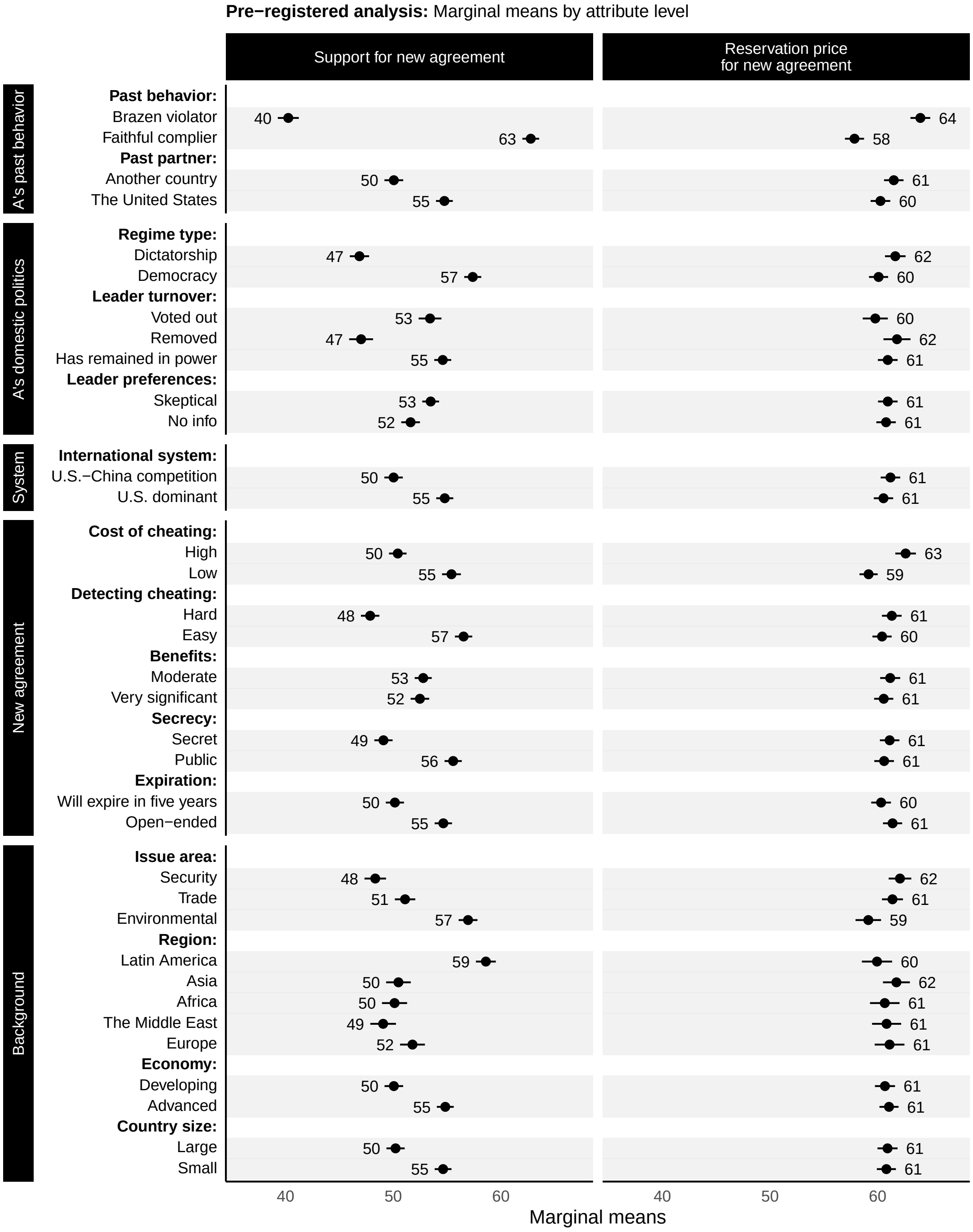
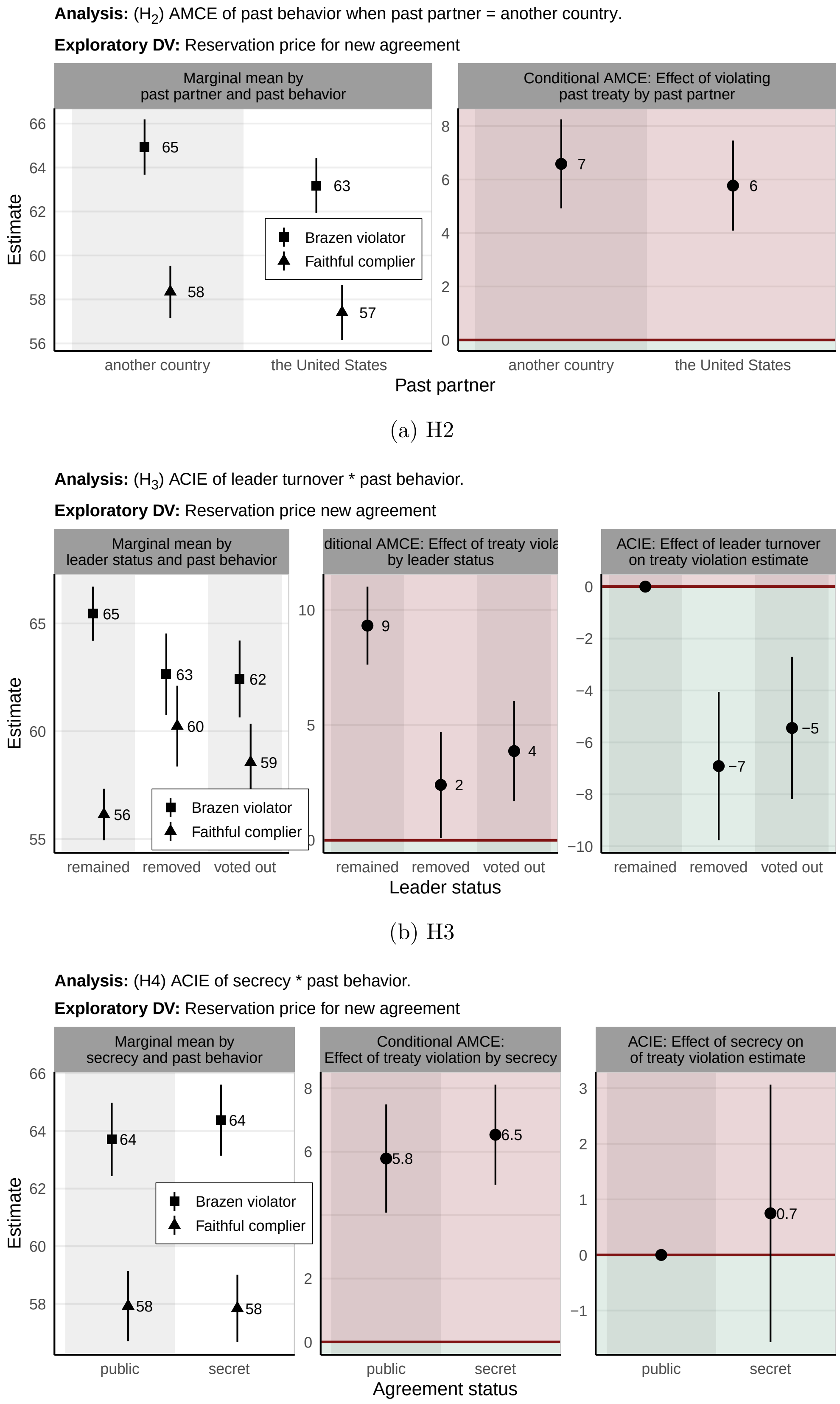


Figure 7: **Marginal Means:** describes the level of favorability toward profiles that have a particular feature level, ignoring all other features. Left column (pre-registered DV) uses data from the first 11 profiles, while the right column (exploratory DV) uses data from last 2 profiles.



(a) H2

(b) H3

(c) H4

Figure 8: H2-H4

B.2 Attention checks

Refer to these with **??** and **??**

Table 6: Pre-treatment attention check.

Outcome	**% (n)**
Pre-treatment	
Failed	5.6% (101)
Passed	94.4% (1,713)

Table 7: Attention check results for after first scenario and after last scenario.

	Scenario 1	Scenario 13
	%(n)	**% (n)**
Attribute: Leader		
Passed	79.6% (1,364)	67.2% (1,151)
Failed	20.3% (347)	32.7% (560)
Attribute: Past behavior		
Passed	83.8% (1,435)	77.6% (1,329)
Failed	16.2% (278)	22.4% (384)
Attribute: Secrecy		
Passed	72.5% (1,242)	54.3% (931)
Failed	27.5% (471)	45.7% (782)

B.3 Prolific Demographics

Table 8: Demographic Breakdown

Level	Percentage (n=1713)
Education	
Some high school or less	0.8% (14)
High school graduate	13.1% (224)
Some college	23.0% (394)
2 year degree (e.g., Associates degree)	13.1% (224)
4 year degree (e.g., BA, BS)	32.4% (555)
Post-grad (e.g., JD, MD, PhD, MA, etc)	17.6% (302)
Gender	
Male	48.3% (828)
Female	50.6% (866)
Other	1.1% (19)
Race	
White	68.7% (1176)
Black or African American	14.5% (249)
Indigenous	2.3% (39)
Asian	7.9% (135)
Some other race	5.7% (97)
Prefer not to answer	1.0% (17)
Age Range	
18-20	0.0% (0)
20-29	18.9% (324)
30-39	20.2% (346)
40-49	16.2% (278)
50-59	19.1% (327)
60-69	16.8% (287)
70-79	4.7% (81)
80+	0.4% (6)
NA	3.7% (64)
Party ID	
Democrat	33.6% (575)
Republican	28.4% (486)
Independent	37.0% (634)
Other	1.1% (18)
Ideology	
Very liberal	12.8% (219)
Liberal	21.2% (363)
Slightly liberal	11.0% (189)
Moderate, middle of the road	22.5% (385)
Slightly conservative	10.8% (185)
Conservative	14.7% (252)
Very conservative	7.0% (120)
Income	
Less than $30,000	16.9% (289)
Between $30,000 and $59,999	28.2% (483)
Between $60,000 and $149,999	40.9% (701)
$150,000 or more	11.6% (199)
Prefer not to say	2.4% (41)
Region	
Midwest	19.4% (332)
Northeast	17.3% (296)
South and Central	42.3% (725)
West	21.0% (360)

B.4 Vignette Text

.

In this portion of the survey, we will ask you to consider a series of scenarios that the United States could face in the future.

In the scenarios, the **United States must decide whether to cooperate with another country** on a particular set of policy issues and on what terms.

We will ask you to consider the details of the situation and whether or not the United States should cooperate with the country in question.

Some of the details the scenarios may be important to you, while others may be less so. We will ask you to evaluate thirteen scenarios.

Each scenario is independent from all of the others. We would like you to consider each one as an entirely new scenario under a different president and in a different context.

After reading this page, respondents proceed to complete the conjoint tasks.

Scenario Introduction

The United States is considering negotiating a new [security/environmental/economic] agreement with another country that we will call "Country A."

(1) About Country A and the International System: [The United States and China are the two most powerful countries and compete to influence the behavior of other countries around the world./The United States is the most powerful country and has a large influence on the behavior of other countries around the world.] Country A is a [small/large] [dictatorship/democracy] with [an advanced/a developing] economy that is located in [Europe/Asia/the Middle East/Latin America/Africa].

(2) A previous agreement between the United States and Country A: In the past, Country A and [Country B/the United States] were members of an international treaty focused on [security/economic/environmental] issues. The treaty lasted for many years. That treaty has now expired and so is no longer in force. An independent watchdog group charged with monitoring compliance documented how Country A [repeatedly and brazenly violated the terms of the agreement even when it would have been relatively easy to honor them/faithfully fulfilled the terms of the agreement even when it was quite difficult to honor them]. Since then, the leader of Country A [has remained in power. OR was removed from power/voted out of office ...after being rejected by the public and elites and has been replaced by a new leader with different views on most issues]. [no info/ That [same/new] leader of Country A has recently expressed their skepticism of international cooperation and agreement.]

(3) A newly proposed agreement between the United States and Country A Under the proposed agreement, the United States and Country A would commit to [increase defense spending/reduce carbon emissions/reduce tariffs on imports from each country]. The agreement and its terms will be [highly-publicized; other countries would see that the U.S. is cooperating with Country A/secret; other countries would not see that the U.S. is cooperating with Country A]. The agreement is designed such that the United States and Country A share all benefits of the agreement equally. It would be [easy/difficult] to detect if Country A were not upholding their end of the agreement. If Country A violated the terms of the deal it would be [minimally/extremely] harmful to the United States. Experts believe the agreement would produce [moderate/very significant] benefits to the United States.

For each profile, respondents see both the narrative version as well as a summary in table form. One representative summary is depicted below:

Introduction	
The United States is considering negotiating a new trade agreement with another country that we will call ”Country A.”	
About Country A and the International System	
Country A...	is a democracy. is in Europe. is a small country. has an advanced economy.
The United States...	has a large influence on the behavior of other countries around the world.
New agreement between Country A and the United States	
The new agreement would be...	secret; other countries would not see that the U.S. is cooperating with Country A.
All benefits are...	shared equally between the United States and Country A.
New treaty has...	moderate benefits to the United States.
Country A cheating would be...	extremely harmful to the U.S. easy to detect.
The treaty would......	reduce tariffs on imports from each country.
Previous agreement between Country A and another country	
Old treaty was with...	another country.
Old treaty covered	trade issues.
Country A...	faithfully fulfilled the terms of the agreement even when it was difficult.
Country A’s leader during that time	has remained in power.
Country A’s leader...	is skeptical of international cooperation.

Table 9: Example of table format that respondents see

B.5 Conjoint Randomization

Randomization:

- Paragraph randomization:
  - "Scenario Introduction" always comes first.
  - Order of paragraphs 1-3 is randomized across respondents and then held constant for all profiles a respondent sees.
- Sentence randomization:
  - Within paragraph 1, order of sentences are randomized across respondents.
  - Within paragraph 3, order of sentences are randomized across respondents.

C Parallel Encouragement Design

C.1 Assumptions

Required assumptions would be:

1. Manipulation exclusion restriction: in the encouragement arm, the encouragement only affects the outcome through its influence on the value of the mediator. In other words, manipulating/encouraging the mediator does not activate any other mechanisms that affect Y.

2. Parallel randomization: "treatment alone is randomized in the natural-mediator arm and that both the treatment and the mediator are randomized in the manipulated-mediator arm" (Acharya, Blackwell and Sen, 2018, 363).

3. No interactions: The direct effect of X on Y for any subject cannot depend on the value of M. In other words, in the encouraged mediator arm, the effect of abrogation on support for cooperation cannot depend on beliefs about U.S. reputation for toughness.

4. Monotonicity: No perverse effects of manipulation of mediator. In other words, providing information that the agreement is secret can't increase beliefs that it is public.