# Victim's Dilemma: The Role of Toughness Promoting Interstate Cooperation

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

March 29, 2026

**Abstract**

Recent years have been marked by a surge in defections from cooperative arrangements, as states increasingly abrogate or violate international commitments. Yet existing theories of international cooperation do not adequately explain the incentives and motivations of states as they formulate responses to such behavior. We argue that reputational dynamics are central to understanding how states respond to defection, and that reputation operates through two distinct channels. First, violations damage the offending state's reputation for upholding commitments, reducing victims' and observers' appetite for future cooperation. Second, states that fail to punish violations risk damaging their own reputation for toughness, creating incentives to confront rather than tolerate defection. Two pre-registered survey experiments test these mechanisms. A conjoint experiment on the American public shows that a partner's past violations have a larger effect on support for cooperation than any other feature of a prospective agreement, including the gains at stake, ease of monitoring, and regime type. Leadership turnover in the offending state partially — but not fully — offsets this reputational damage. A vignette experiment on UK respondents demonstrates that tolerating violations generates significant reputational and material costs, including among observers not directly harmed, and that these costs persist even when the violator is much more powerful. Together, these results suggest noncompliance creates dual reputational pressures that simultaneously raise the cost of future cooperation for violators and discourages accommodation as a viable response.

*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. Authors thank Mark Crescenzi, Lisa Dellmuth and Stefanie Walter for helpful comments and Inken von Borzyskowski and Felicity Vabulas for sharing data related to exit from international organizations.

In a widely publicized speech at the 2026 World Economic Forum in Davos, Switzerland, Canadian Prime Minister Mark Carney declared that a recent pattern of coercive and unilateral behavior by the United States marked the end of the rules-based international order. In Carney's telling, Washington's imposition of tariffs, withdrawal from multilateral agreements, and broadly excessive use of economic and military pressure, had greatly undermined U.S. credibility in the world. According to the Canadian prime minister, the damage was done and would be long-lasting: "This bargain no longer works... we are in the midst of a rupture, not a transition."

Carney's speech raises a critical question that is not squarely addressed by the literature on international cooperation: *how does one state's defection—sometimes called bad behavior or "non-cooperation"—shape the prospects for future cooperation?* Existing theories largely treat widespread violations as off the equilibrium path; the benefits of cooperation and the threat of punishment are, in many accounts, sufficient to deter this level of conflict (Keohane, 1984; Axelrod, 1984; Martin, 1992; Simmons, 2000). As a result, we know little about whether and on what terms states can find their way back to cooperation following defections. And because existing models consider the prospect of punishment as a mechanism that sustains cooperation by discouraging defection, they have little to say about the incentives and choices of trade partners and third party observers to defection, leaving open the question of why some might pursue highly public confrontations of the sort witnessed at Davos.

We argue that reputational dynamics are central to understanding how states respond to defection, and that reputation operates through two distinct channels. The first and more familiar channel occurs when violations damage the offending state's reputation for upholding commitments, raising the cost of future cooperation (e.g., Guzman, 2008; Tomz, 2007; Keohane, 1984; Morse and Pratt, 2025). This mechanism helps explain why states victimized by U.S. violations of trade commitments might demand harsher terms in future agreements or decline to cooperate altogether.

More notably, we also highlight a second channel focused on the reputational concerns of two other actors: the victimized party and third-party observers. While the violator's damaged reputation for compliance hurts the odds of cooperation, so too do the concerns of these other actors, who worry that accommodating the violator (and being seen to do so) will harm their *reputation for toughness*. While there has been a great deal of work on such reputations in the domain of

international security (e.g., Schelling, 1966; Weisiger and Yarhi-Milo, 2015; Lupton, 2020; Yarhi-Milo, 2018; Renshon, Dafoe and Huth, 2018; Jervis, Yarhi-Milo and Casler, 2021), we show that these theories have oft-overlooked implications for understanding the consequences of bad behavior in international cooperation, law, and political economy. Concerns about appearing tough help explain why targets of defection respond with the kind of vocal, high-profile resistance on display at Davos — even when doing so risks foreclosing a return to cooperation. Our argument is that both reputational dynamics are important mechanisms through which defection affects future patterns of cooperation for all parties: prospective entrants to agreements, states facing possible or realized defections, and third party observers.

In service of our argument, we present two novel pre-registered survey experiments, fielded as part of a "sequential, non-harmonized meta-design" in which each study addresses different but overlapping features of our argument (Kertzer, Renshon and Xu, 2025). First, we use a conjoint experiment—in which non-compliance is randomized alongside other features—to explore the reputational effect of past violations of international agreements on targets' and observers' appetite for future cooperation. The conjoint experiment addresses the substantive questions, *how large are the reputational effects of violations, compared to other salient factors?*, *do past violations increase the future costs of cooperation for violators?*, *how much do observers infer from non-compliance targeting other states (not their own)?* and *do concerns about toughness inhibit cooperation with past violators?*

We find that past violations have a very large effect compared to other features of future cooperative deals (e.g., cooperative gains, ease of monitoring), suggesting that even profitable and ostensibly enforceable future bargains may not compensate for past breaches of trust. This is noteworthy given the lack of evidence on the relative importance of reputations in this domain compared to other salient aspects of the cooperative arrangement and helps us—for the first time—properly calibrate the importance of reputation. We also show that the reputational effect of past violations increases the costs of future cooperation for violators, and that past violations have a similar effect on the beliefs and preferences of both victims and observers. Consistent with work on the leader-specific reputations in the context of militarized crises (Wolford, 2007; Renshon, Dafoe and Huth, 2018), we report the optimistic result that leadership turnover in the offending state can substantially, though not entirely, offset the reputational effect of past actions. Finally, we find

evidence consistent with our "toughness" argument: cooperation with past violators is more likely in situations in which states do not have to fear being seen as "weak" by observers.

Second, we use a vignette experiment—in which we randomize state responses to violations by others—to examine the reputational dynamics that shape the choice to accommodate or confront violators (Dellmuth and Walter, 2025). Existing work on reputation in cooperation theory overwhelmingly focuses on potential defectors' reputation for compliance as a break on cheating (Guzman, 2008; Tomz, 2007; Keohane, 1984; Morse and Pratt, 2025), encouraging states to remain in bargains over the long term despite short term incentives to exploit partners. Again drawing on the international security literature, we argue that another reputational dynamic is highly salient for the politics of non-cooperation: targets' and observers' reputation for toughness (Schelling, 1966; Weisiger and Yarhi-Milo, 2015). We argue that targets and observers of violations have strong incentives to cultivate a reputation for toughness in showdowns over compliance.

This second study addresses the substantive questions, *does accommodation of violators hurt states' reputations for toughness?*, *do states pay fewer costs when accommodating major powers (when they arguably have less freedom to stand their ground)?*, and *does accommodation rehabilitate violators' reputation for fulfilling commitments?* We find that confronting—rather than accommodating—violations increases reputation for toughness, increases the generosity of proposed terms in new agreements, and reduces the odds of being exploited in the future. We find further that the costs of accommodation are no smaller when confronting a major power. We also show that accommodation does not lessen the reputational damage done to the violator; they are not rehabilitated. Last, and perhaps most surprisingly, states that confront violations by punishing defectors are seen as more attractive partners for third parties and accommodating bad behavior does not sanitize the reputation of the defecting state.

This paper makes a number of contributions. First, we offer evidence on the comparative importance of many factors theorized to affect the attractiveness of cooperative bargains, including the gains at stake, ease of monitoring, costs of being cheated, time horizon, past compliance behavior, and regime type. We show that reputational concerns—operationalized by past actions—have a particularly large effect relative to these other factors. Further, we break new ground by showing that past actions and features of the current agreement interact in theoretically coherent ways. Second, we provide evidence of a reputational mechanism that provides states with clear and mean-

ingful incentives to punish non-cooperative behavior. The perception that accommodation signals weakness functions as a what Axelrod (1986) called a "metanorm" (or "secondary" rule; Hart, 1961) about how states *ought* to respond to breaches of shared norms. These processes are crucial to sustaining cooperation in environments that lack external enforcement (Ostrom, 1990; Brunnée and Toope, 2010). Finally, and related, we build a bridge between research on reputation in the domains of international security and international cooperation and law. We show that reputation for toughness and leader-specific reputations, long of principal interest to scholars of international security, are also important features of compliance interactions in the sort of non-securitized policy domains often of interest to scholars of international cooperation and law.

# 1 How Dual Reputational Concerns Shape Responses to Defection

A long tradition in international relations research examines what makes cooperation between states possible and sustainable. Scholars have identified a range of factors shaping the attractiveness of cooperative arrangements, including the presence of institutions (Keohane, 1984; Axelrod and Keohane, 1985), domestic political institutions and regime type (Mansfield, Milner and Rosendorff, 2002; Lipson, 2013), and the ease with which cheating can be detected (Jervis, 1978). Yet for many IR scholars, a dominant factor driving the choice to cooperate is the past behavior of a potential partner (Keohane, 1984; Axelrod, 1984; Tomz, 2008; Crescenzi et al., 2012; Weisiger and Yarhi-Milo, 2015; Morse and Pratt, 2025). Violations of international commitments signal that a partner may not honor future obligations, and this reputational damage is often compounded when commitments are embedded in international law (Guzman, 2008). As a result, breaking or abandoning international agreements frequently triggers costly backlash, exclusion, or material penalties (Martin, 1992; Schmidt, 2025).

Despite this consensus, the empirical record is mixed (Downs and Jones, 2002): some hostile behaviors trigger costly punishment, while others generate accommodation. Conventional accounts focus almost exclusively on the reputation of the *offending* state, treating reputational concerns as a mechanism that deters violations in the first place (Guzman, 2008; Tomz, 2007; Keohane, 1984; **?**). Recent work has begun to more seriously consider the incentives of victims and observers as they react to "non-cooperation" (Dellmuth and Walter, 2025). Among other concerns, accommodating

a violator may preserve cooperative gains in the short run, but risks signaling weakness and inviting further predatory behavior.

In this paper, we seek to assess how dual reputational concerns drive foreign audience reactions to breaches of cooperative agreements. We focus specifically on behaviors that represents defection from a pre-existing cooperative relationship. Such behavior includes two key elements. First, it must represent a shock to expectations (i.e., a departure from the state's prior behavior), thereby conveying new information about the state's type (Tomz, 2007; Mattes and Weeks, 2019). Second, it reflects a clear hostility toward international commitments (Keohane, 1984; Guzman, 2008; Dellmuth and Walter, 2025). The 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. North Korea's 2003 decision to withdraw from the NPT, expel international inspectors, and restart reprocessing facilities clearly qualifies as defection, for example. So too does a U.S. leader publicly casting doubt on whether the United States would come to defense of NATO countries.

To fix ideas, consider a scenario in which countries $A$ and $B$ interact in view of an audience, $C$, over some initial period of time $t_1$. If $A$ engages in defection at some point in $t_1$—for example, by brazenly violating a bilateral agreement with $B$—we expect this action to shape the behavior of both the harmed actor ($B$) as well as foreign audiences ($C$). In this environment, the reaction of these audiences determine the intensity of costs that the offending state $A$ will face. If $B$ and $C$ choose to accommodate country $A$'s non-cooperation, $A$'s costs will be minimal. If instead they decide to exclude $A$ from cooperative endeavors, or impose more demanding terms on it in the future, the costs of defection may be quite steep.

As these actors formulate a response to hostile behavior, we argue that their motivations will be influenced by two distinct reputational dynamics. First, the act of defection should damage country $A$'s reputation for upholding international commitments. We call this first mechanism the "reputational cost of defection." Second, foreign audiences must consider their own reputational concerns as they formulate a response to country $A$: an accommodative reaction from $B$ and $C$ is likely to damage their reputation for toughness, potentially inviting future predatory behavior. We call this the "reputational cost of accommodation."

Notably, both reputational processes are moderated by the same underlying inferential chal-

lenge: audiences observe behavior but must assess the disposition behind it (Mercer, 2010). Whether defection damages a violator's reputation, or accommodation signals weakness, depends on whether observers attribute the behavior to stable underlying preferences or to situational pressures beyond the actor's control (Tomz, 2008; Renshon, Dafoe and Huth, 2018). When behavior is clearly voluntary and conspicuous, reputational consequences will be most severe; when extenuating circumstances muddy the waters, the same behavior may generate more muted responses. As we discuss in the subsections that follow, this attribution logic shapes the magnitude of both reputational costs and points to specific moderating variables that we test in our experiments.

### Reputational Costs of Defection

The first reputational mechanism we examine occurs when defection from a cooperative relationship 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 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. Defection will thus cause the cost of cooperation to go up.

This 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 an act of defection by an offending state. Because this learning process is not limited to the victims of defection, we argue that reputational effects should emerge among *both* direct victims 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, 2025).

A scope condition for this expectation regards the publicity of the behavior and is taken from work on the mechanics of reputations: since reputations require publicity (Dafoe, Renshon and Huth, 2014), how far the reputational damage extends—and how costly it is for the perpetrator—depends on the *visibility* of the non-cooperative behavior. In a completely private interaction between $A$ and $B$, only $A$'s reputation *in the eyes of* $B$ can be affected. Therefore, defection can

only affect a state's broader reputation to the extent that the behavior is made public.[1]

While the expectation that bad behavior will be met with costly consequences is not new, adopting a reputational lens focuses attention on an important moderating variable that helps to determine the magnitude of costs that transgressor states will face: how their behavior is explained or attributed by audiences. Both the victim ($B$) and the audience ($C$) face the well-studied dilemma of observing only behavior while desiring to assess the state's underlying cooperative type. This complex inferential task requires making judgments about whether the state's non-cooperative posture will endure. Does the non-cooperation signal a change in a state's preferences or type, or is the behavior as an idiosyncratic departure driven by exigent circumstances? Audiences thus face the problem of observing a single leader's policy choice, but having their optimal response depend on the long-run policy preferences—unobserved, for the most part—that emerge from the state's political system.

To resolve this uncertainty, we argue that observers look for signals of domestic political support for the leader's actions. In other words, the effect of $A$'s non-cooperative behavior will depend on whether observers believe that $A$'s behavior reflects the underlying preferences of its domestic populace. We argue that one strong signal of underlying domestic support for a policy is the visible removal of a leader following a particular foreign policy action: when a non-cooperative leader is removed, observers are more likely to discount the bad behavior of that leader. Conversely, the retention of the leader reinforces the idea that the behavior reflects structural preferences, compounding reputational damage to the state. Leader turnover—long thought to help "reset" reputations (Renshon, Dafoe and Huth, 2018)—does so, we argue, partly through the process of changing how observers integrate behavior into reputational beliefs.

### Reputational Costs of Accommodation

In addition to the damage to the perpetrator's reputation, acts of defection trigger a second reputational mechanism. As victims and observers consider how to respond to the perpetrator, they will have their own reputational concerns about appearing weak in the face of antagonistic behav-

[1]Of course, this in turn means that states that anticipate backlash have an incentive to minimize visibility by concealing their demands or conduct. For example, the Trump Administration reportedly required trade partners to sign non-disclosure agreements, partially insulating the U.S. from reputational damage. "Govt Signs Secrecy Pact ahead of US Talks," June 22, 2025, *Bangkok Post* `https://www.bangkokpost.com/business/general/3054975/govt-signs-secrecy-pact-ahead-of-us-trade-talks`.

ior. Standing strong in opposition to non-cooperation, by condemning or retaliating against the perpetrator, enhances a state's reputation for toughness in international politics. Accommodating defection, on the other hand, signals weakness and diminishes a country's perceived resolve (Bloch and McManus, 2024). The reputational damage associated with accommodation can create political costs. Among other things, such an effect could encourage further predatory behavior by the perpetrator or other states (Dellmuth and Walter, 2025).

This mechanism provides a potential second motivation for actors to respond harshly to defection. States that accommodate violations are likely to be viewed as irresolute or weak. This reputational effect may be most intense among direct victims. However, even observer states often want to ensure that defection does not set a damaging precedent that could undermine beneficial patterns of cooperation or make themselves look weak. Accordingly, we argue that third parties will also pay a reputation cost for cooperating with states that have a clear and recent record of hostile behavior.

As with defection, the reputational effects arising from accommodation are often complicated by a recurring attribution problem. When audiences observe a country accommodating hostile behavior, they may attribute that act to a lack of toughness or place blame on contextual circumstances. In particular, audiences may have trouble distinguishing between cases where a state voluntary accommodates defection and instances when they are coerced by a more powerful state. Accordingly, we argue that the negative effect of accommodation on reputation for toughness should be smaller when the perpetrator state is relatively powerful. When the offending state is very powerful, even highly resolute victims and observers may lack the ability to respond with meaningful punitive action. As a result, the choice to accommodate sends a less informative signal compared to the case in which the offender is relatively less powerful. Put differently, absent the capacity to punish, it is not obvious that accommodation reflects limited resolve or toughness. However, when a state has the capacity to punish a non-cooperative partner yet chooses not to, this may be more easily chalked up to the former being irresolute or a pushover.

### Empirical Implications

Our argument generates several empirical implications. Our first set of predictions concerns the effects of defection: we predict that non-cooperation by country $A$ will increase the cost of cooper-

ation that the violated state and observers will demand from $A$, that these costs will be mediated by reputational damage to $A$'s reputation and moderated by the belief about whether the "bad behavior" in question reflects $A$'s underlying preferences or is an aberration. Our argument also generates implications about the costs of accommodation: we predict that accommodation of the violator state—by either the victim or an observer—will harm the accommodator's reputation for toughness and that this will be moderated by beliefs about whether accommodators had the choice about whether to comply (i.e., by $A$'s relative power).

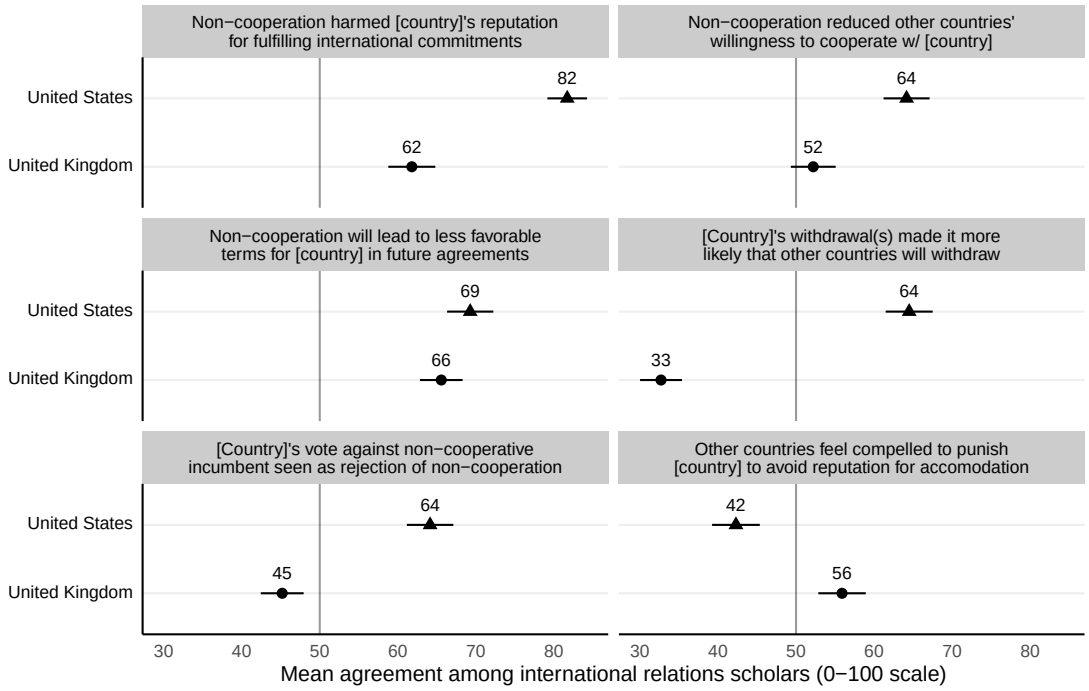
## 2 Scholars Perceive the Importance of Reputation

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 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.

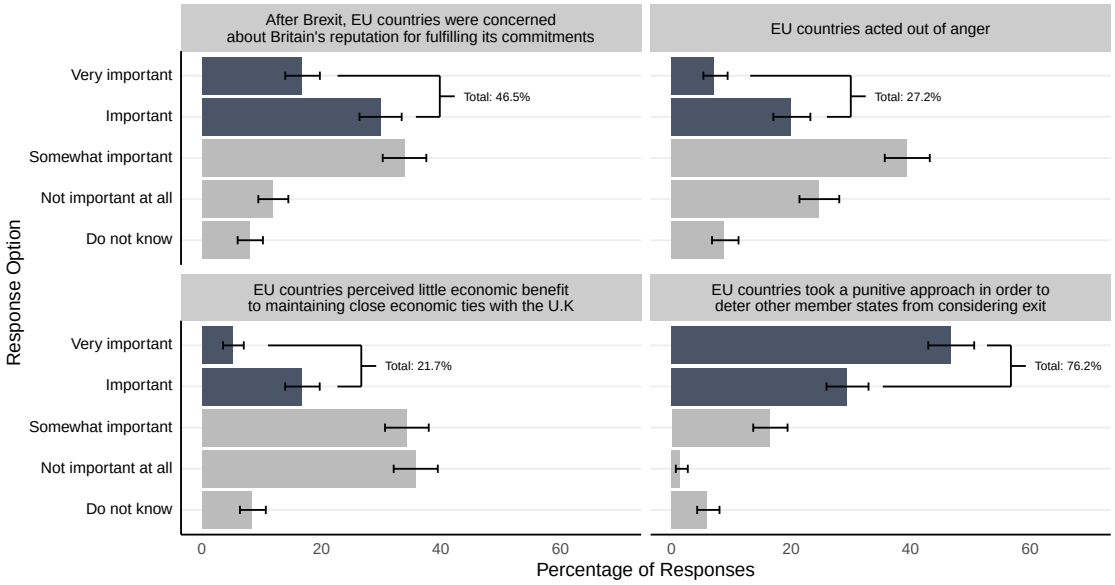
We take three lessons from our descriptive survey questions. First, our respondents believed that leaving agreements stifles future cooperation, both by reducing the willingness of others to cooperate and by downgrading the terms under which agreements are offered. Roughly two-thirds of our sample (64%; 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

[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.

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 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 responses to Brexit. There was less support, however, for potential alternatives, such as 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 Overview of 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 generated from our reputation-based theory. Because the theory carries implications for different stages of interactions and for different actors, we adopt a "sequential, non-harmonized" meta-design (Kertzer, Renshon and Xu, 2025) where "meta-design" describes the structure of the overall research design rather than the identification/randomization within a particular experiment. The key feature of such a meta-design is that the experiments test

overlapping but differing empirical predictions and build on one another to test more aspects of a theory than a traditional "harmonized" design focused exclusively on replication.

To clarify the objectives of our experimental studies, Table 1 presents a summary of our research questions linked to the *empirical estimands*: the observed quantity or contrast generated by our experimental designs. Appendix B captures our estimands more broadly, using a streamlined version of the framework suggested by Lundberg, Johnson and Stewart (2021). 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.

We field two pre-registered experiments. In our conjoint experiment, American respondents ($N = 1,800$) rate the attractiveness of a cooperative agreement between their country (the U.S.) and an un-named *Country A*, where $A$'s past behavior (i.e., whether it has violated a cooperative agreement) is randomized alongside a number of other attributes. In a separate vignette experiment, U.K. respondents ($N = 3,314$) assess the reputation for toughness of a country that either accommodates a violation of a cooperative agreement or not. The country whose behavior is randomized is either the UK itself, a country that has been directly violated or a bystander country. As Table 1 shows, we design our experiments to test overlapping implications of our theory, though the conjoint experiment is more informative of Research Questions 1 & 2 and the vignette experiment more directly addresses Research Question 3.

Research Question 1 lays the groundwork for our original efforts by investigating the effect of past non-cooperation on the costs of future cooperation. In accordance with a great deal of previous work (e.g., Axelrod, 1984; Keohane, 1984; Tomz, 2007; Guzman, 2008), we expect that a reputation for noncompliance will dampen enthusiasm for cooperation with the violator state.

Research Questions 2 & 3 address the reputational mechanisms that are the focus of our argument. Research Question 2 focuses on the impact of past bad behavior on cooperation that occurs through damage to the violator's reputation for upholding commitments. While such a result is anticipated by many theories, there is as yet scant evidence on the *relative* importance of reputational factors compared to other salient aspects of the cooperative arrangement: 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

| Question | Empirical Estimands |
|---|---|
| 1. What is the effect of past non-cooperation on costs of future cooperation for the violator? | [*conjoint exp:*] 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. |
| 2. Does non-cooperation decrease support for cooperation through *actual damage to the violator's reputation for fulfilling commitments*? | [*conjoint exp:*] conditional AMCE: interaction of *past behavior* and *identity of harmed country* attribute (levels: U.S./Country B) on support for cooperation with hypothetical Country **A** in Prolific sample. |
| | [*conjoint exp:*] conditional AMCE: interaction between *past behavior* attribute and *leadership turnover* attribute in online conjoint study using Prolific sample. |
| ...and how does the effect of reputation compare to other salient aspects of the cooperative arrangement? | *[conjoint exp:]* comparison of AMCE of past behavior to other attributes of agreement |
| ...and can that reputation be rehabilitated? | *[vignette exp:]* ATE of accommodation of violator on violator's reputation and approval, in online convenience sample of UK residents via Prolific |
| 3. Does non-cooperation by **A** decrease support for cooperation with **A** through potential partners' *concern for their reputation for toughness?* | *[conjoint exp:]* conditional AMCE: interaction between *past behavior* attribute and *secret agreement* attribute in online conjoint study using Prolific sample. |
| | *[vignette exp:]* ATE of accommodation on accommodator's reputation for toughness in online convenience sample of UK public, via Prolific |
| ...and does the power of the violator moderate the reputational harm of accommodation | *[vignette exp:]* interaction of accommodation and violator power level on reputation for toughness |

Table 1: **Research Questions and Empirical Estimands**

other features in a highly powered design. Instead, some vary framing of, or responses to, violations (e.g., Chilton, 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 *relative* 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). We also investigate the related question of whether a the reputational damage done by violating agreements can be "laundered" through other states' acquiescing to the violation.

Research Question 3 builds upon the work in international security on reputations for toughness to test a core part of our argument. We argue that one way that violations dampen cooperation is through the reputational concerns of the violated parties and observers, who worry that being seen to accommodate bad behavior will make them look weak. This dynamic is tested indirectly in our conjoint experiment and more directly in our the vignette experiment, where it is the primary outcome.

## 4 The Reputational Consequences of Bad Behavior

In our pre-registered (single profile) 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).[4] Our goal is to estimate 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. Figure 2 visualizes the survey flow, Table 3

[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.

lists our pre-registered outcomes and Appendix C contains details of information presented to respondents and the randomization scheme.

**Logistics and Design** We measured attentiveness both before, during and after the conjoint experiment (see Appendix C.2 for details). The survey centers on a foreign policy scenario (see Appendix C.4) involving a potential cooperative agreement between the U.S. and a hypothetical Country A.[5] We manipulate 15 factors/attributes related to Country $A$, its previous behavior, the potential cooperative agreement and the international system.[6] 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.[7] Table 2 presents the attributes and levels of the conjoint.

Our main outcome variable is support for the cooperative agreement with Country $A$ (scaled $0-100$). 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 $A$. This alternate DV addresses our exploratory question of how reputational effects shape the *cost* of cooperation: is plausible that the main effect of a bad reputation is—rather than *availability* of cooperative partners—in having to agree to more onerous terms in order to secure a cooperative agreement in the first place. 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 in order to facilitate legibility and comprehension.[8]

**Recruitment and Statistical Power** We recruited a general population sample of 1,800 adults—motivated by power calculations for our interaction hypotheses and detailed in our pre-analysis

[5]Single profile designs do have trade-offs (see Hainmueller, Hangartner and Yamamoto, 2015), but are 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 (one is rarely presented with two entirely different international agreements and asked to choose).

[6]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).

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

[8]See Appendix C.4 for details on how text is presented and randomized.

Figure 2: Consort Diagram for Pre-registered Conjoint Experiment

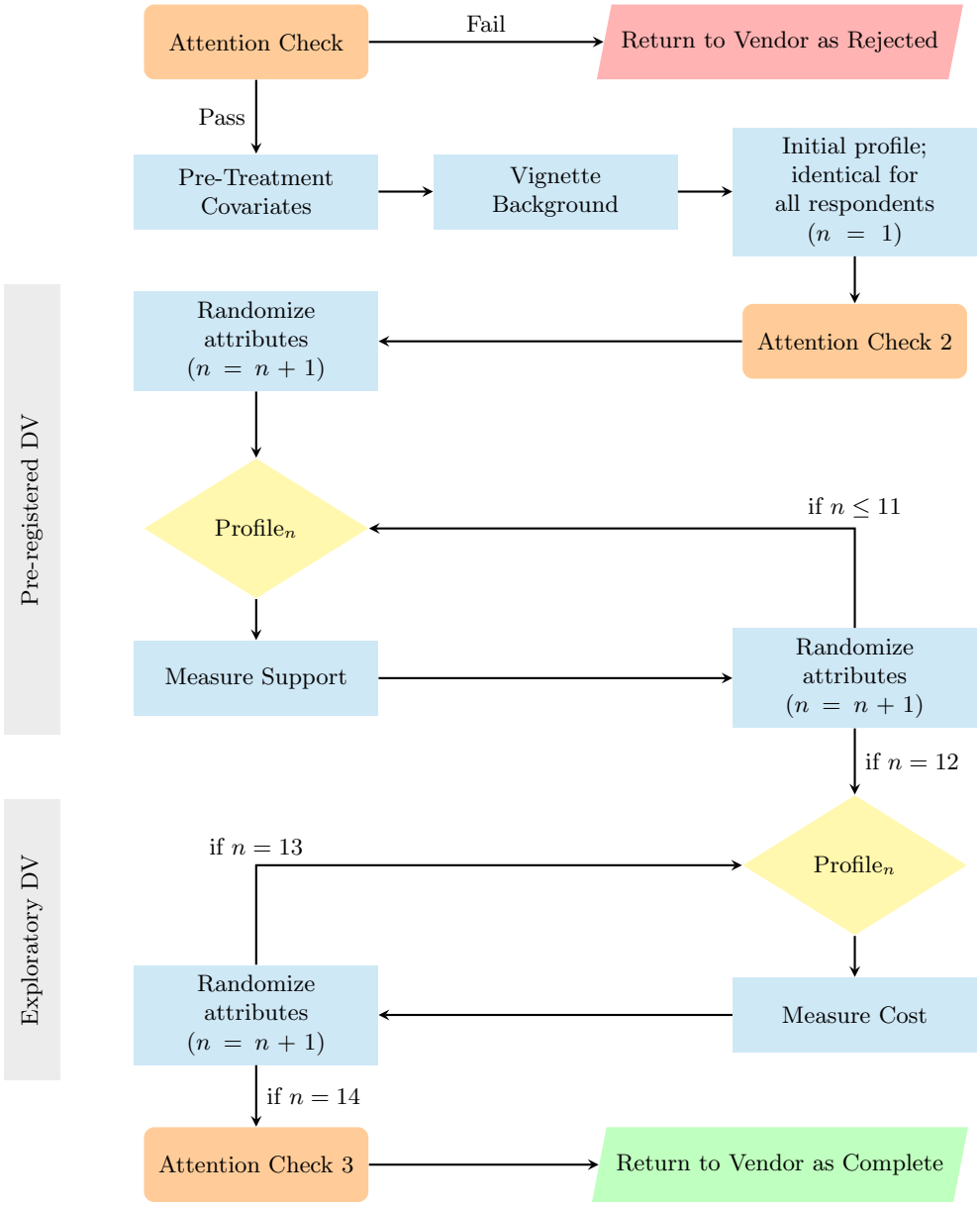


Table 2: Conjoint Attributes and Levels

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

*Note*: All dimensions randomized independently across profiles.

plan—based in the United States via Prolific.[9] We fielded the experiment from November 22-26[10], 2024.

As a first cut at estimating the quality of our sample, we consider our attention checks (Ap-

[9]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).

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

pendix C.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, and that attentiveness did decline moderately between the $1^{st}$ and $13^{th}$ profile (from 79% to 66%).[11]

Our main AMCE results (Figure 3) are generated by regressing support (pre-registered) and reservation price (exploratory) on a complete battery of attribute level indicators using OLS. As per the suggestion in Liu and Shiraito (2023) and as pre-registered, we present AMCEs both with and without adjustment for multiple hypotheses (Benjamini and Hochberg, 1995).

### 4.1 The Direct and Indirect Consequences of Non-Cooperation

**The critical—and far-reaching—importance of past behavior** We summarize the substantive results of our conjoint in Table 3. Our first research question focused on the link between non-cooperative behavior and future cooperation. We find strong support for $H_1$, which predicted that past violations will make it harder and more costly for states to secure future cooperative deals. 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 $0-100$ scale and raised the reservation price by 6 points.[12] 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.

Our second hypothesis implicates reputations in a broad sense by investigating whether past violations affect future bargains even when the bad behavior is directed against a third-party ($H_2$). As we show in Figure 4a, when the past violation targeted a third-party country, support for the agreement declines by 19.4 points.[13] That bad behavior has roughly similar effects whether directed

[11] 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.

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

[13] 95% CI:18.3, 20.6; $p < .001$; BH adj. $p < .001$. Estimates are from OLS models that regress outcomes on a complete battery of attribute treatment level indicators and the interaction of past partner ("another country" or

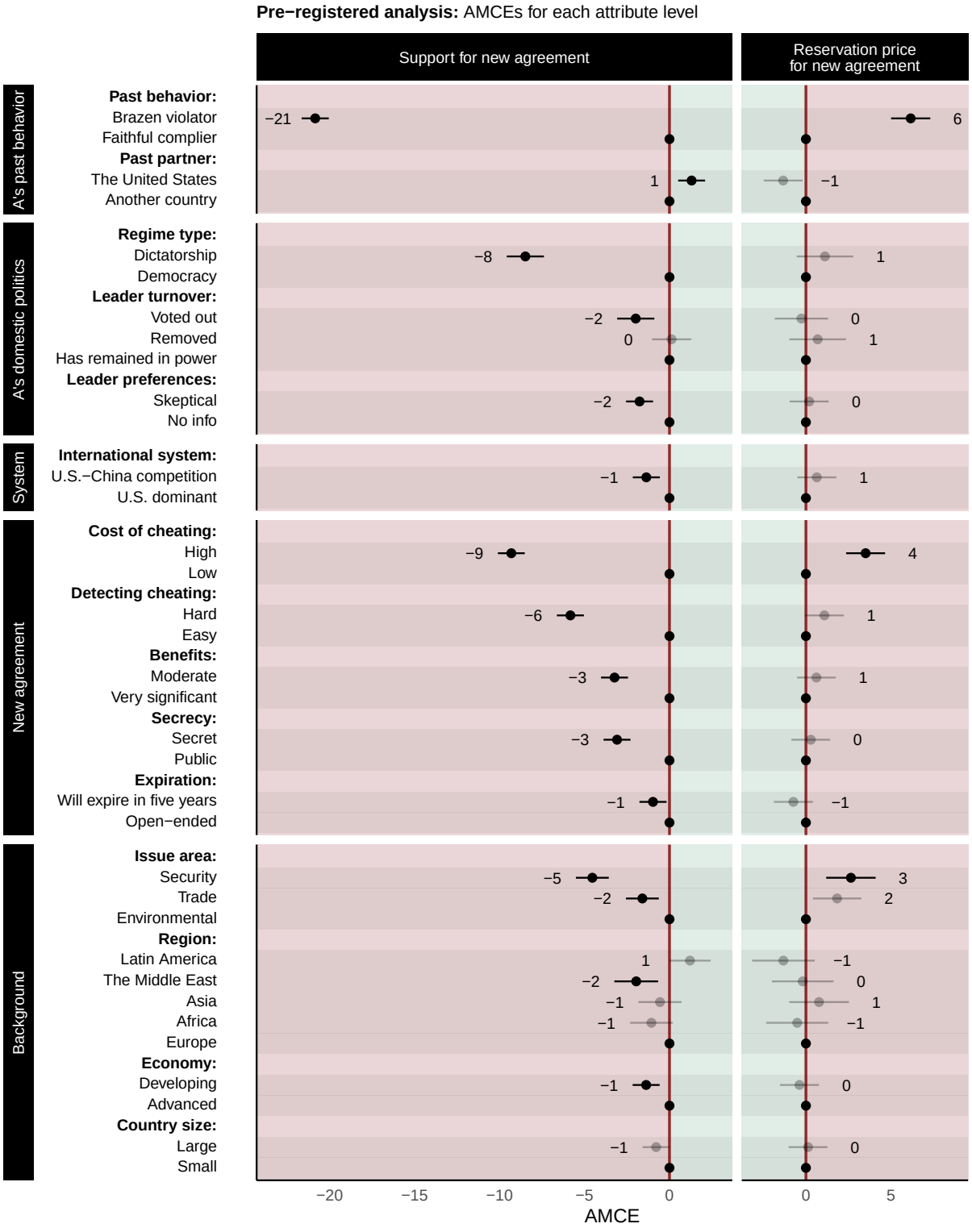



Figure 3: **Average Marginal Causal Effect (AMCE):** a summary measure of the overall effect of an attribute (relative to reference category), averaging over the effect of all other attributes. Black circles (•) indicate that the AMCE remains statistically significant after BH correction for multiple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

| | | **Expectation** | **Test of...** | | **Findings** |
|---|---|---|---|---|---|
| *Conjoint* $H_1$ | main effect (AMCE) | Violating previous agreement reduces support for cooperation with Country A | Core theoretical prediction (IV $\rightarrow$ DV) | ✓ | Past violations reduce support for new agreement by 20.8 points on our 0–100 scale.<br><br>(95% CI: 20.1, 21.6; $p < .001$; BH adj. $p < .001$). See Figure 3. |
| *Conjoint* $H_2$ | 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.<br><br>(95% CI:18.3, 20.6; $p < .001$; BH adj. $p < .001$). See Figure 4a. |
| *Conjoint* $H_3$ | 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).<br><br>Democracies: (95% CI:17.8, 21.6; $p < .001$; BH adj. $p < .001$)<br>Dictatorship: (95% CI:20.2, 24.1; $p < .001$; BH adj. $p < .001$). See Figure 4b. |
| *Conjoint* $H_4$ | 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.<br><br>(95% CI:0.1, 3.2; $p = 0.042$; BH adj. $p = 0.056$) See Figure 4c. |

Table 3: **Outcomes for pre-registered hypotheses in Conjoint Design**

at the respondent's country or a third party suggests a key role for reputation (as opposed to other factors, such as anger at your own state having been harmed).[14]

**Two distinct reputational mechanisms** Our two other core hypotheses—$H_3$ and $H_4$—indirectly test 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 interaction effects with 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 $\times$ leader turnover—as a bundle of both mechanisms, and the second—past violation $\times$ secret agreement—as implicating only respondents' concerns over their reputation for toughness. Motivated by Leeper, Hobolt and Tilley (2020), we generate our ACIEs in two different ways, the first ("restricted models") requiring more assumptions and the second ("unrestricted") requiring fewer.[15]

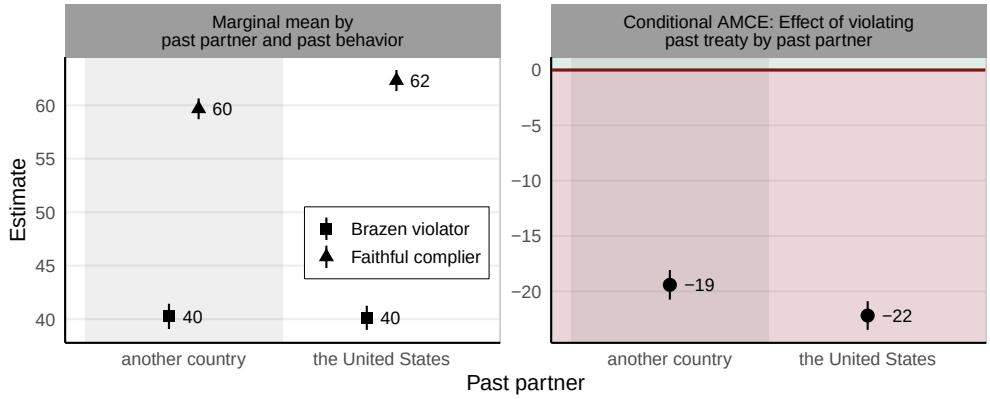
With respect to leader turnover, our expectation was that the negative effect of a past violation would be *smaller* when followed by leader replacement in the partner state ($H_3$). The middle panel of Figures 4b depicts the effect of non-cooperative behavior conditional on leader turnover and the right panel displays the pre-registered interactive quantities of interest. Consistent with our expectations, leader turnover *significantly* attenuates the effect of past violations: whether the cooperative partner is a democracy or an autocracy, leader replacement increases support for cooperation with the state in question by between $20 - 24$ points.[16] These results are broadly consistent with both of our mechanisms without being able to distinguish between the two, since the replacement of such a leader could be interpreted in two ways: first, it may undermine the

"the United States") and past behavior ("brazen violator" or "faithful complier"). Conditional AMCE estimates in Figure 4a.

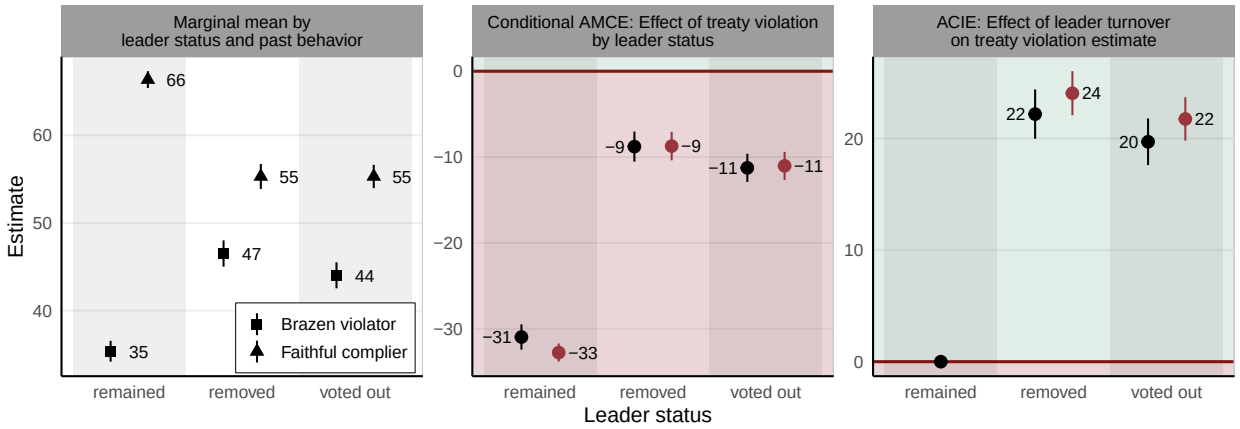
[14]Similar results are obtained using our exploratory DV: past violations increase the reservation price by a similar magnitude when it is directed at the respondent's country (6 points) as when it is directed against a third country (7 points; see Appendix C).

[15]Leeper, Hobolt and Tilley (2020) focused on sub-groups defined by respondent-level characteristics; our sub-groups are defined by exposure to either secrecy or leader turnover. 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"). Our "restricted model" implicitly assumes that there are no relevant interactions other than that between past behavior and the other attribute of interest (leader turnover or secrecy). "Unrestricted" estimates relax the "no other relevant interactions" assumption. Both are both broadly consistent with how we articulated the relevant contrasts of interest in our pre-analysis plan.

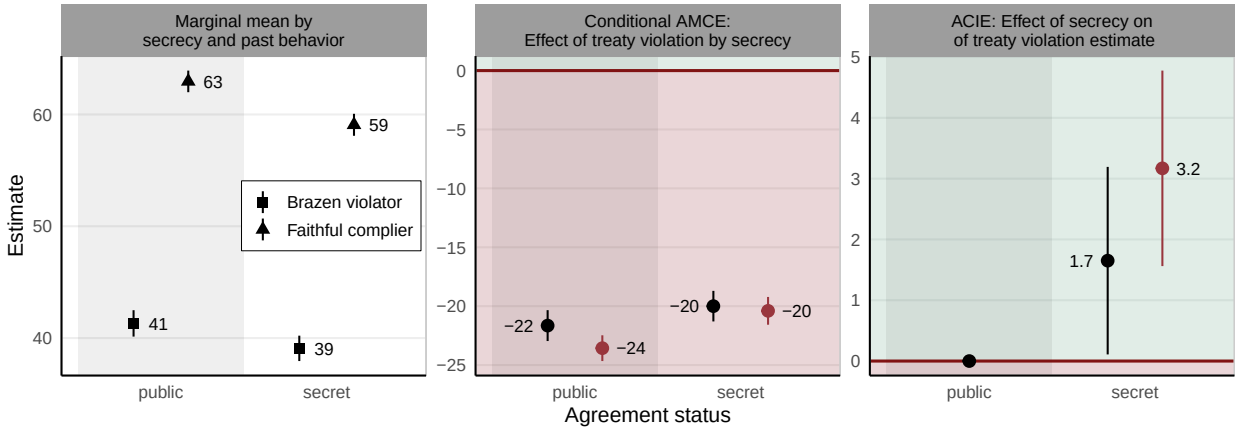
[16]Autocracies: 95% CI:20.2, 24.1; $p < .001$; BH adj. $p < .001$. Democracies: 95% CI:17.8, 21.6; $p < .001$; BH adj. $p < .001$. Leader replacement also reduced the average reservation price by 7 percentage points in the autocratic case and 5 percentage points in the democratic case (Appendix C)




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




(b) **Conjoint 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) **Conjoint 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

inference that non-cooperative preferences are an enduring feature of the partner state; second, it might provide plausible cover for respondents to cooperate with the state without damaging their own reputation for toughness.

In fact, because of how our treatments are designed, our pre-registered contrasts are likely *over-estimating* the extent to which leader turnover attenuates the effect of past violations. In fact, there is reason to believe that the negative effect of non-cooperation is even more durable than the estimates based on our pre-registered contrasts suggest. Using exploratory contrasts—see Appendix C.6—we find that leader replacement eliminates only about $\frac{1}{3}$ of the effect of past non-cooperative behavior.

While the leader turnover treatment may be interpreted as a bundle of our two mechanisms, the secrecy treatment implicates only our argument concerning states' 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. We thus 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. Our results provide some evidence consistent with this expectation: the effect of past non-cooperation is mitigated to a small degree by making the agreement secret. 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.

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

# 5 The Reputational Effects of Accommodation

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 also provided initial, indirect support in favor of both reputations broadly and the two specific reputational mechanisms from our argument. Finally, the conjoint study also offers a guide to build a factorial experimental design, highlighting salient features that provide verisimilitude and realism to our vignette without influencing the effects of our treatments.

While the conjoint study leveraged randomization of non-cooperative behavior, our second study fixes noncooperation and randomly assigns whether such violations are accommodated or not. This allows us to home in on our second reputational mechanism: states' concern for their own reputation for toughness if they should be seen accommodating a state that previously defected. To that end, we fielded a pre-registered vignette experiment on a sample ($N = 3,314$) of the public in the United Kingdom.[17]

**Logistics and Design** In our vignette experiment, one state ($A$) engages in cooperative defection targeted toward another country ($B$). Respondents are randomized into one of three different IDENTITY conditions: they either learn about the behavior of $B$ (the state that has been targeted), $C$ (an observer to $A$ and $B$'s interaction) or their own country (the *United Kingdom*; also an observer). We randomize the BEHAVIOR in question, such that $B/C/UK$ either decides to accommodate Country $A$ or does not. The vignette includes an additional treatment arm, the POWER of Country $A$ relative to whichever other country is in the vignette ($B/C/UK$). We use this additional treatment arm to assess whether the reputational effect of accommodation is moderated by the target's ability to resist. We also randomly vary the REGIME TYPE (democracy/dictatorship) of Country $A$ and its GEOGRAPHIC REGION (Europe/Asia), and average over these arms in order to attempt to fix respondents' background beliefs about the scenario. After the experiment, respondents answer questions related to the reputation of the state described in the vignette as well as exploratory questions about State $A$'s reputation. Our consort diagram is depicted in Figure 5.

[17] PAP: `https://osf.io/mzrh2`.

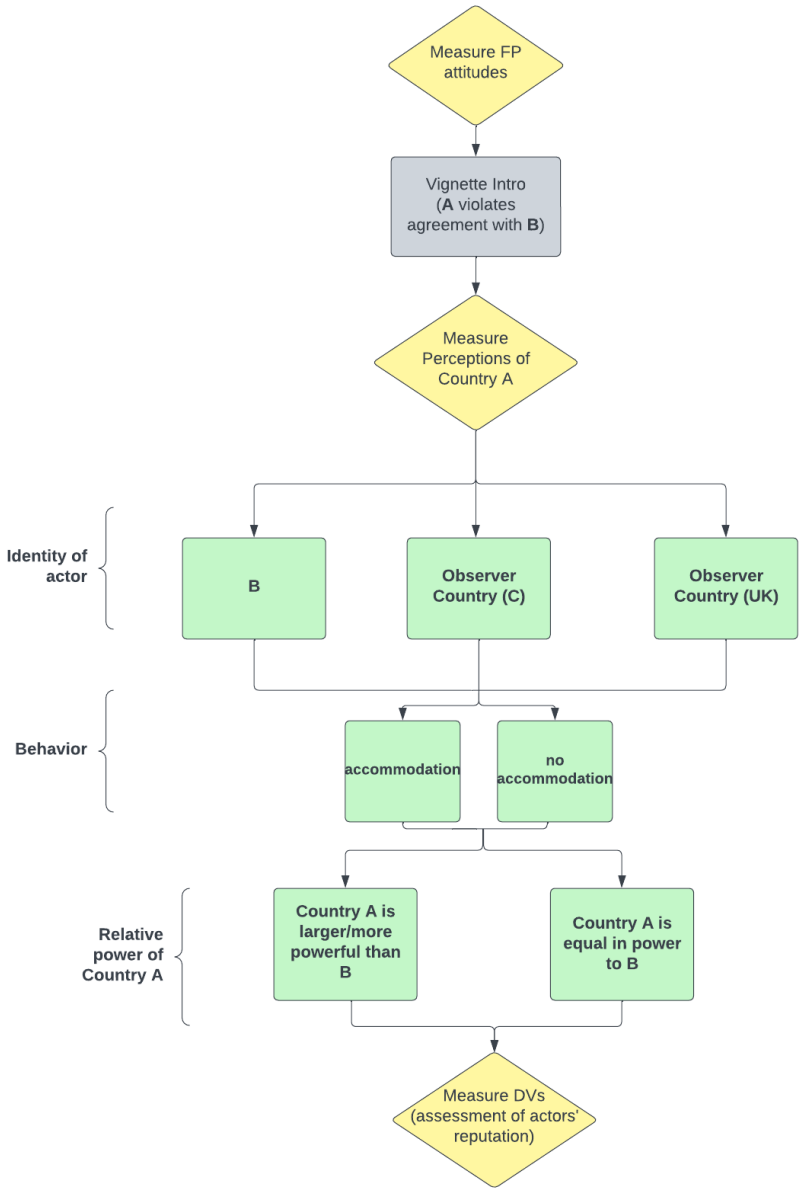



Figure 5: Consort Diagram for Study 2 (vignette experiment). **Green** boxes represent orthogonally randomized elements (randomized with equal probability in all cases).

Our main interest is the average treatment effect (ATE) of accommodation—by Country $B/C/UK$—on reputations for toughness, as well as the potential that this main effect is moderated by the relative power of Country $A$. In Table 4 we list our main hypotheses, and in Table 5 we list a number of exploratory questions suggested by our theory. We rely on two-tailed hypothesis tests, controlling the false discovery rate at $q = .05$ using the Benjamini–Hochberg procedure across our four primary hypotheses (*Vignette* $H_{1a-c}$ and *Vignette* $H_2$). As per our PAP, we determined our

$N$ for the study based on attaining 0.8 power to detect small effects (Cohen's $d = .2$) for our first set of main contrasts (*Vignette* $H_1 a - c$).[18] This size $N$ allows us to detect interaction effects for *Vignette* $H_2$ only if it they exceed seven points on the 0-100 scale.

## 5.1 The Cost of Accommodation

### The reputational costs of accommodation

We begin with our main pre-registered hypothesis—*Vignette* $H_1$—which anticipated that accommodation would negatively affect a state's reputation for toughness, operationalized as standing firm in foreign policy disputes. Our experimental design allows us to estimate the cost of accommodation by two different classes of actors: the victim of the dispute (Country $B$) or one of two third party observers (Country $C$ and the $UK$).

In accord with our argument, we find that accommodation substantially damages a country's reputation for toughness, regardless of whether the accommodator is the victim or an observer, or whether or not that observer is another country or respondents' home country (Figure 6). Averaging over the other features of the experiment, accommodation leads to a 34 point drop on our $0 - 100$ point measure of a country's reputation for standing firm.[19] This is a *large* effect—representing a shift of about 1.24 standard deviations—and suggests that states that accommodate non-cooperative behavior pay significant reputational costs for doing so.

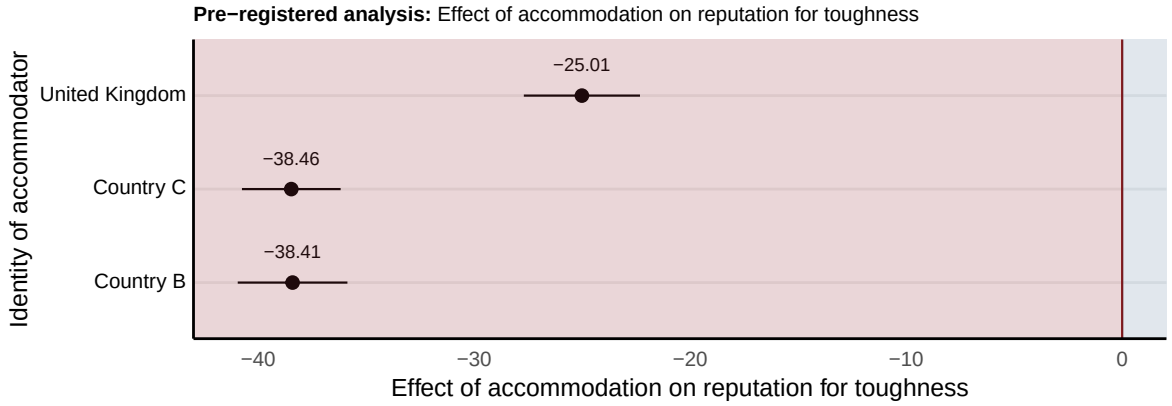
Importantly, we find that whether accommodation comes with reputational costs does not depend on whether the state in question was directly harmed or an innocent bystander. When the *victim* of defection (Country $B$) accommodates the transgressor, its reputation for toughness declines by 38.4 points.[20] This cost is roughly equivalent (38.5 points) when Country $C$, the third party observer, is identified as the accommodator.[21] Notably, the cost of accommodation is smaller (25 points) when the accommodator is the respondent's home country, though that difference represents the combined bundle of the UK being a real country (compared to hypothetical; see Brutger et al. 2022) as well any "hometown bias" that lead to respondents "going easy" on their

[18] As per PAP, we used a rough (and conservative) approximation to anticipate how the BH correction might affect our detectable effect size: we applied a Bonferroni adjustment to the per-test $\alpha$ level.

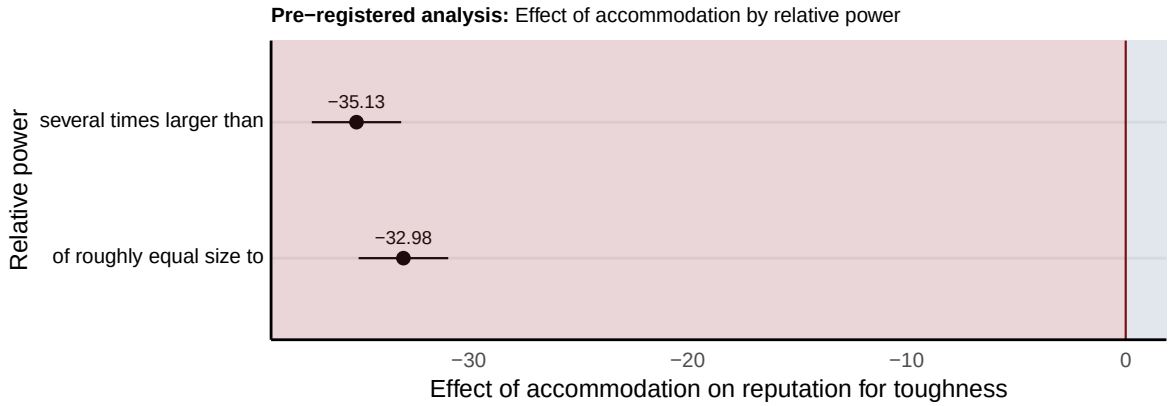
[19] 95% CI: 32.6, 35.5; $p < .0001$

[20] 95% CI: 35.9, 40.9; $p < .001$

[21] 95% CI: 36.2, 40.8; $p < .001$




(a) *Vignette* $H_1$




(b) *Vignette* $H_2$

Figure 6: Treatment effects of accommodation on reputation for toughness. The top panel displays average treatment effects by identity of the observed state. The lower panel shows treatment effects by relative power of the violating state.

own country (Kertzer, Renshon and Yarhi-Milo, 2018).[22]

In accordance with the main thrust of our argument, we also find that the reputational damage done to accommodators carries material costs. Not only do respondents attribute damage to accommodators' reputation for toughness, they are also more reluctant to engage in future cooperation with them: when Country $B$ or $C$ accommodates bad behavior, it reduces support for new cooperative deals with Country $B$ or $C$ by an average of 17.3 points.[23] We summarize this and other exploratory results in Table 5. Accommodation has more modest—but still precisely estimated—effects on the share of the cooperative benefits that the respondent would demand from

[22] 95% CI: 22.3, 22.7; $p < .001$
[23] 95% CI: 15.1, 19.1; $p < .001$

| | | Expectation | Test of... | | Findings |
|---|---|---|---|---|---|
| *Vignette* $H_{1a-c}$ | main effect (ATE) | Accommodation—relative to non-accommodation—decreases reputation for toughness when identity of accommodator is Country A ($H_1a$), Country B ($H_1b$), and the UK ($H_1c$) | second mechanism in theory (concern over reputation for toughness) | ✓ | Accommodation reduces perceptions of resolve by an average of 34.1 (95% CI: 32.6, 35.5) points on our 0–100 point scale.<br>• B: -38.4 (95% CI: -40.9, -35.9; $p < .001$, BH adj. $p < .001$).<br>• C: -38.5 (95% CI: -40.75, -36.18; $p < .001$, BH adj. $p < .001$).<br>• UK: -25 (95% CI: -27.7, -22.3; $p < .001$, BH adj. $p < .001$). |
| *Vignette* $H_2$ | interaction effect (CATE) | Accommodation—relative to non-accommodation—generates smaller effects on the state's reputation for toughness when Country A is stronger (rather than of equal power), pooling across *identity* treatments. | moderating effect of attribution | ✗ | Relative power does not moderate the effect of accommodation.<br>Accommodating a powerful state reduces perceptions of resolve by 35.1 points on our 0–100 scale, while accommodating a state of equal power reduces perceptions of resolve by 33.0 points. The difference in these effects is 2.14 (95% CI: 5.0, -.75) points is not statistically signifcant ($p = .146$, Bh adj. $p = .146$) |

Table 4: **Outcomes for pre-registered hypotheses in Vignette Experiment (Study 2)**. $H_1$ and $H_2$ control for false discovery rate using B-H adjustment.

Country $B$ or $C$ in a potential future cooperative deal (2.46 points).[24] Moreover, respondents anticipate that other states would make similar judgments of them: when the $UK$ accommodates defection by continuing to negotiate a new agreement with the violating state, our respondents judged that the $UK$ would enjoy less support for cooperation from other actors and anticipated that other cooperative partners would be more willing to engage in non-cooperative acts against the $UK$ (See EQ5 in Table 5).

Finally, we find no evidence of a reputational tradeoff to accommodation. It's possible that, for example, accommodation might signal weakness but also a strong willingness to cooperate that observers might associate with a commitment to fulfilling international obligations. However, our results instead indicate the opposite: accommodation does nothing but harm a state's reputation across multiple dimensions. Pooling across the identity of the accommodator, accommodation *harmed* the accommodator's reputation for compliance by 16.4 points.[25] This effect was detectable but relatively small for Country $B$, the direct victim of A's non-cooperative behavior. For third parties (Country $C$ and the $UK$), the effect was considerably larger.[26]

[24]Share of cooperative benefits: 95% CI: 1.61, 3.3; $p < .001$.

[25](95% CI: 15,22.2; $p < .001$)

[26]Country B ATE = -3.8 points, 95% CI: 1.47, 6.03; $p = 0.001$. Country C ATE = -28 points, 95% CI: CI: 25.74,

| *Exploratory Question* | | Expectation | | Findings |
|---|---|---|---|---|
| *EQ 1* | interaction effect (CATE) | Does effect of accommodation on reputation for toughness depends on the identity of the accommodator? | ✓ | Accommodation by B and C reduces perceptions of resolve by similar amounts (-38.4 vs. -38.5), but accommodation by the UK has a smaller effect (-25.0). The difference between the UK effect and the B and C effects are -13.3 ($p < .001$) and -13.4 ($p < .001$) respectively. |
| *EQ* 2 | main effect (ATE), alternate DV | Does accommodation increase Country [B/C/UK]'s reputation for compliance, relative to non-accommodation? | ✗ | States that accommodate non-cooperation harm their reputation for compliance by an average of 16.4 points (95% CI: 15, 17.9, < .001) |
| *EQ* 3 | main effect (ATE), alternate DV, subset to *identity*=B or C | Does accommodation (by B or C) shapes respondent preferences for behavior towards B/C? | ✓ | Accommodation by B and C reduces support for cooperation, makes exploitation more likely, and increases the cost of future cooperation. Effect of accommodation on…<br>… support for future co-op: -17.3 (95% CI: -15.5, -19.1, $p < .001$)<br>… share of future benefits demanded by observer: 2.46 (95% CI: 1.61, 3.33; $p < .001$)<br>… willingness of observer to violate agreement: 4.36 (95% CI: 2.53, 6.2; $p < .001$) |
| *EQ* 4 | conditional effect (CATE), alternate DV, subset to *identity*=B or C | Does the relative power of Country A moderate the effect of accommodation (by B or C) on respondent preferences for behavior towards B/C? | ✗ | Relative power does not moderate the effect of accommodation.<br>Change in cost of accommodation when moving from "several times more powerful" to "of roughly equal size"…<br>… support for future co-op: 2.55 (95% CI: 2.55, 6.17; $p = .167$)<br>… share of future benefits demanded by observer: -1.02 (95% CI: -2.73, .69 ; $p = .242$)<br>… willingness of observer to violate agreement: -1.71 (95% CI: -5.39, 1.96; $p = .359$) |
| *EQ* 5 | main effect (ATE), alternate DV, subset to *identity*=UK | Does accommodation (relative to non-accommodation) by UK affect respondents' beliefs about other actors' likely behavior towards them? | ✓/✗ | Accommodation by UK leads respondents to anticipate less support for cooperation and future violations, but does not affect anticipated cost of cooperation. Effect of accommodation on…<br>… anticipated support for future co-op: -11.4 (95% CI: -13.7, -9.15; $p < .001$)<br>… anticipated share of future benefits demanded: .327 (95% CI: -1.05, 1.71; $p = .46$)<br>… anticipated willingness of others to violate agreement: 3.56 (95% CI: 1.1,6.03; $p = .005$) |
| *EQ* 7 | main effect (ATE), alternate DV | Does accommodation (relative to non-accommodation) affect respondent approval of Country B/C/UK's 's behavior (pooling across identity treatments)? | ✓ | Accommodation reduces approval of Country B/C/UK's behavior by 34 points (95% CI: 33.2, 36.4; $p < .001$) |
| *EQ* 8 | conditional effect (CATE), alternate DV | Does Country A's relative power moderate the effect of accommodation (relative to non-accommodation) on Country B/C/UK's approval (pooling across identity treatments)? | ✗ | The effect of accommodation on approval is not moderated by the relative power treatment. When Country A is "several times larger," the cost of accommodation is 34.5 points. This quantity is 35.1 when Country A is "of roughly equal size." This difference of .64 (95% CI: -3.81, 2.53; $p = .691$) is not statistically significant. |
| *EQ* 9 | main effect (ATE), alternate DV | Does accommodation (relative to non-accommodation) affect respondent perceptions of Country A's reputation for compliance and toughness (pooling across identity treatments)? | ✓/✗ | Accommodation of Country A improves Country A's reputation. Effect of accommodation on…<br>… perceived resolve: 3.57 points (95% CI: 1.51.5.63; $p = .0007$)<br>… perceived compliance: .69 points (95% CI: -.72, 2.1; $p = .956$)<br>… approval: -.13 points (95% CI: -1.52,1.26; $p = .180$) |

Table 5: **Outcomes for pre-registered exploratory questions in Vignette Experiment (Study 2).**

We speculate that respondents understand accommodation, particularly by third parties, as a violation of the implicit rules of enforcement or the secondary rules of international law. By failing to punish Country $A$'s non-cooperative behavior, respondents may understand Country $C$ and the $UK$ to be—to some extent—complicit in the violation. Such a dynamic was visible in Western criticism of India's policy toward Moscow in the wake of Russia's 2022 invasion of Ukraine. By refusing to criticize Moscow, abstaining from votes to condemn the invasion in multilateral fora, and continuing to buy Russian oil and other goods, some observers came to see India as abetting Moscow's unlawful behavior. Our exploratory finding suggests that two dimensions of reputation—for toughness and for compliance—may both push in the direction of enforcement and ultimately support cooperative outcomes in equilibrium. At the same time, as consensus around international laws and norms erode, so too may reputational incentives participate in their enforcement.

Overall, our results suggest that states who are victims of non-cooperative behavior have good reason to be concerned that accommodating such behavior will be costly in reputational terms. Accomodating violators significantly damages states' reputations for toughness and reduces support for future cooperation with the accommodator, whoever they are. This reputational damage occurs even for states that merely continue to cooperate with states that have engaged in non-cooperation against other actors.

**Relative power and the costs of accommodation**

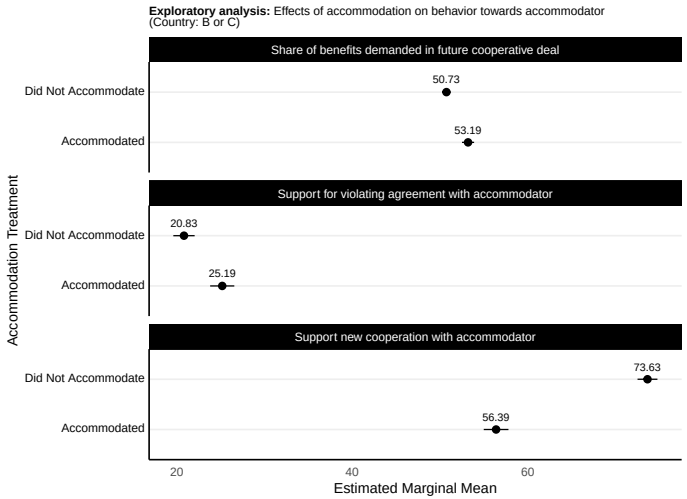
In *Vignette* $H_2$, we anticipated that the reputational costs of accommodation would be smaller in cases when the violation came from a particularly powerful state. Our intuition was that when states have little choice but to acquiesce—because of overwhelming power differentials or dependence on the transgressor—accommodation would provide less information about the state's underlying preferences or type. However, we find that relative power does not moderate the effect of accommodation: whether we described the non-cooperative behavior as coming from a state that is "several times larger than" or "of roughly equal size to" the victim, the effect of accommodation on the state's reputation for toughness was about the same. The small difference (2.14 points) between these two estimates—−32.98 points for equally sized countries and −35.13 for countries that were "several times larger"—-was not statistically significant, though recall that we were only

30.71; $p < .001$. UK ATE = -17.5 points, 95% CI: 14.74, 20.24; $p < .001$.

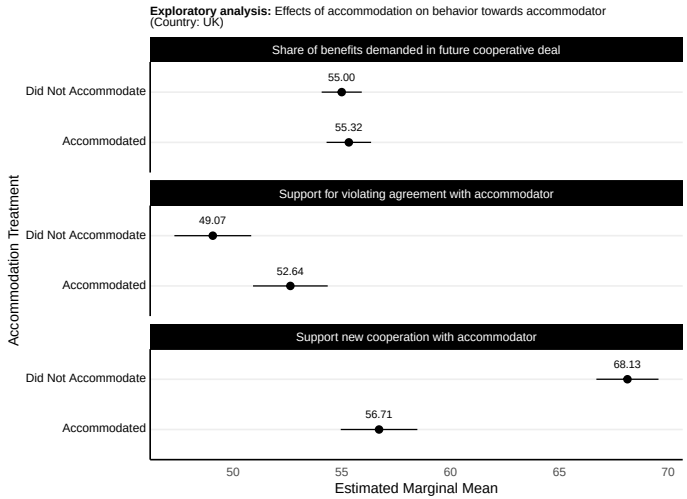
powered to detect an interaction effect of seven points.[27] Without over-interpreting a null, it is worth mentioning that the sign is in the opposite direction from our prediction: rather than getting a break for accommodating more powerful countries, our respondents punish weaker countries *more* for accommodation.

We take this as evidence that, at least for the subjects we recruited, power differentials in the way we operationalized them were not highly relevant to the reputational inferences respondents draw. It is possible that our hypothesis is correct but that for the situations we described it was not immediately obvious that the non-cooperative state might use their power advantage to exploit the victim in the absence of accommodation. Future work might address this by more directly linking the welfare of the victim to the gains from cooperation. Rather than power per se, then, the relevant moderating variable may be *dependence*.

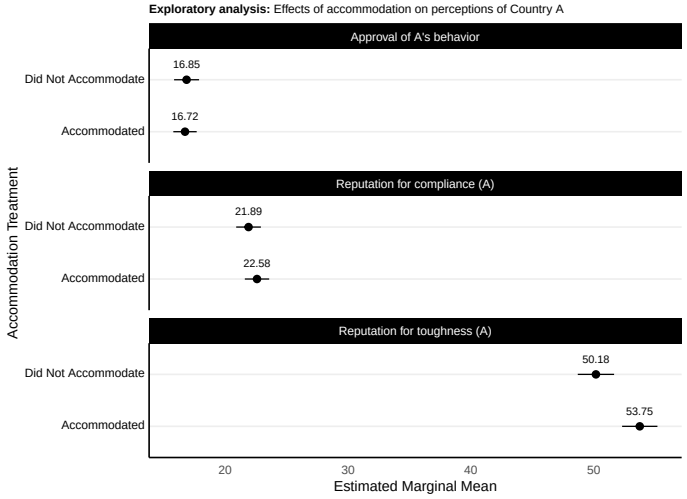
[27] Equal size: 95% CI: -35.03, -30.94; $p < .001$. Larger: 95% CI: -37.17, -33.09; $p < .001$. Difference: 95% CI: -.75, 5.04; $p = .147$




(a)




(b)




(c)

Figure 7: Exploratory analyses

# 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 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.

## References


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. 1986. "An Evolutionary Approach to Norms." *American Political Science Review* 80(4):1095–1111.

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.

Brunnée, Jutta and Stephen J. Toope. 2010. *Legitimacy and Legality in International Law: An Interactional Account*. Cambridge: Cambridge University Press.

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.

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. 2025. "Responding to Non-Cooperation in Global Governance." *Unpublisehd manuscript* .

Donahue, Bailee and Mark JC Crescenzi. 2023. "Weathering the Storm: Discordant Learning about Reputations for Reliability." *Foreign Policy Analysis* 19(2):1–23.

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.

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.

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.

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.

Hart, H. L. A. 1961. *The Concept of Law*. Oxford: Oxford University Press.

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

Jervis, Robert, Keren Yarhi-Milo and Don Casler. 2021. "Redefining the debate over reputation and credibility in international security: Promises and limits of new scholarship." *World Politics* 73(1):167–203.

Jost, Tyler and Joshua D Kertzer. 2023. "Armies and influence: Elite experience and public opinion on foreign policy." *Journal of Conflict Resolution* 68(9):1769–1797.

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. 2018. "Are Red Lines Red Herrings?" *Working Paper* .

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

Kertzer, Joshua D, Jonathan Renshon and Weifang (Victor) Xu. 2025. "Cross-National Survey Experiments." *Working Paper* .

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.

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.

Lupton, Danielle L. 2020. *Reputation for resolve: How leaders signal determination in international politics.* Cornell University Press.

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.

Mattes, Michaela and Jessica L. P. Weeks. 2019. "Hawks, Doves, and Peace: An Experimental Approach." *American Journal of Political Science* 63(1):53–66.

Mercer, Jonathan. 2010. *Reputation and international politics*. Cornell University Press.

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* .

Morse, Julia C and Tyler Pratt. 2025. "Smoke and Mirrors: Strategic Messaging and the Politics of Noncompliance." *American Political Science Review* pp. 1–19.

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

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge 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.

Schelling, Thomas C. 1966. Arms and influence. Routledge.

Schmidt, Averell. 2025. "Damaged relations: How treaty withdrawal impacts international cooperation." *American Journal of Political Science* 69(1):223–239.

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. 2007. *Reputation and International Cooperation: Sovereign Debt Across Three Centuries*. Princeton, NJ: Princeton University Press.

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

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

Wolford, Scott. 2007. "The turnover trap: New leaders, reputation, and international conflict." *American Journal of Political Science* 51(4):772–788.

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

# Appendix

## Table of Contents

# A Descriptive Survey of TRIP IR Scholars

Below we provide the verbatim text of the questions we embedded in the September 2024 Teaching, Research, and International Policy (TRIP) survey.

## 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 6 presents a demographic breakdown of TRIP survey respondents.

Table 6: 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 Estimands

Table 7 captures our estimands, using a simple version of the framework suggested by Lundberg, Johnson and Stewart (2021). Theoretical estimands ($\tau$) 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 $\tau$ is in some cases—for example in row 2—a difference in potential outcomes, and thus inherently unobservable.

| Question | Theoretical Estimand ($\tau$) *unit-specific quantity* | *target pop.* | Empirical Estimands ($\theta$) |
|---|---|---|---|
| What is the effect of non-cooperation on costs of future cooperation for the violator? | causal: effect of a country violating an agreement on their prospects for future cooperation | all respondents (no restrictions) | [*conjoint exp:*] 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. |
| Does non-cooperation 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) | [*conjoint exp:*] conditional AMCE: interaction of *past behavior* and *identity of harmed country* attribute (levels: U.S./Country B) on support for cooperation with hypothetical Country **A** in Prolific sample.<br>[*conjoint exp:*] conditional AMCE: interaction between *past behavior* attribute and *leadership turnover* attribute in online conjoint study using Prolific sample. |
| ...and can that reputation be rehabilitated? | causal mechanism: effect of accommodation of violator on violator's reputation | all respondents (no restrictions) | *[vignette exp:]* ATE of accommodation of violator on violator's reputation and approval, in online convenience sample of UK residents via Prolific |
| Does non-cooperation by **A** decrease support for cooperation with **A** 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) | *[conjoint exp:]* conditional AMCE: interaction between *past behavior* attribute and *secret agreement* attribute in online conjoint study using Prolific sample.<br>*[vignette exp:]* ATE of accommodation on accommodator's reputation for toughness in online convenience sample of UK public, via Prolific |

Table 7: **Theoretical and Empirical Estimands**

The empirical estimand ($\theta$) takes into account real world constraints and focuses only on observed quantities. For example, in row 1, our empirical estimand ($\theta$) is informative of our theoretical estimand under the identification assumptions of the particular method. Here, that means randomization of Country $A$'s history of keeping or abrogating agreements. We list empirical estimands for two different experimental designs: a conjoint study and a vignette (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.

# C Conjoint Experiment

In this section we provide additional information on our conjoint experiment, including supplementary results, survey text, an example profile, and a description of the randomization procedure.

## C.1 Additional conjoint results

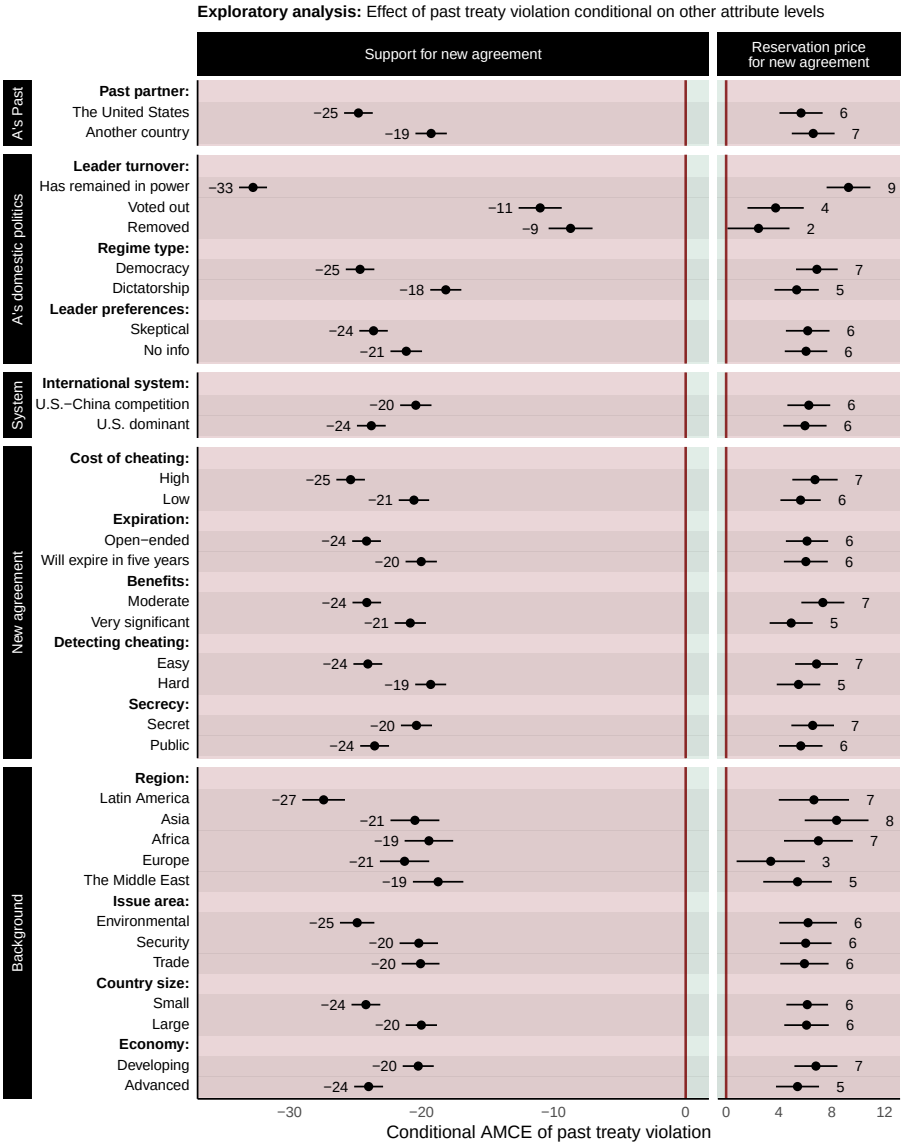



Figure 8: **Conditional Average Marginal Component Effect (cAMCE):** the effect of past treaty violation conditional on other attribute levels. Black circles (•) indicate that the AMCE remains statistically significant after BH correction for multiple comparisons (all estimates do remain significant) .

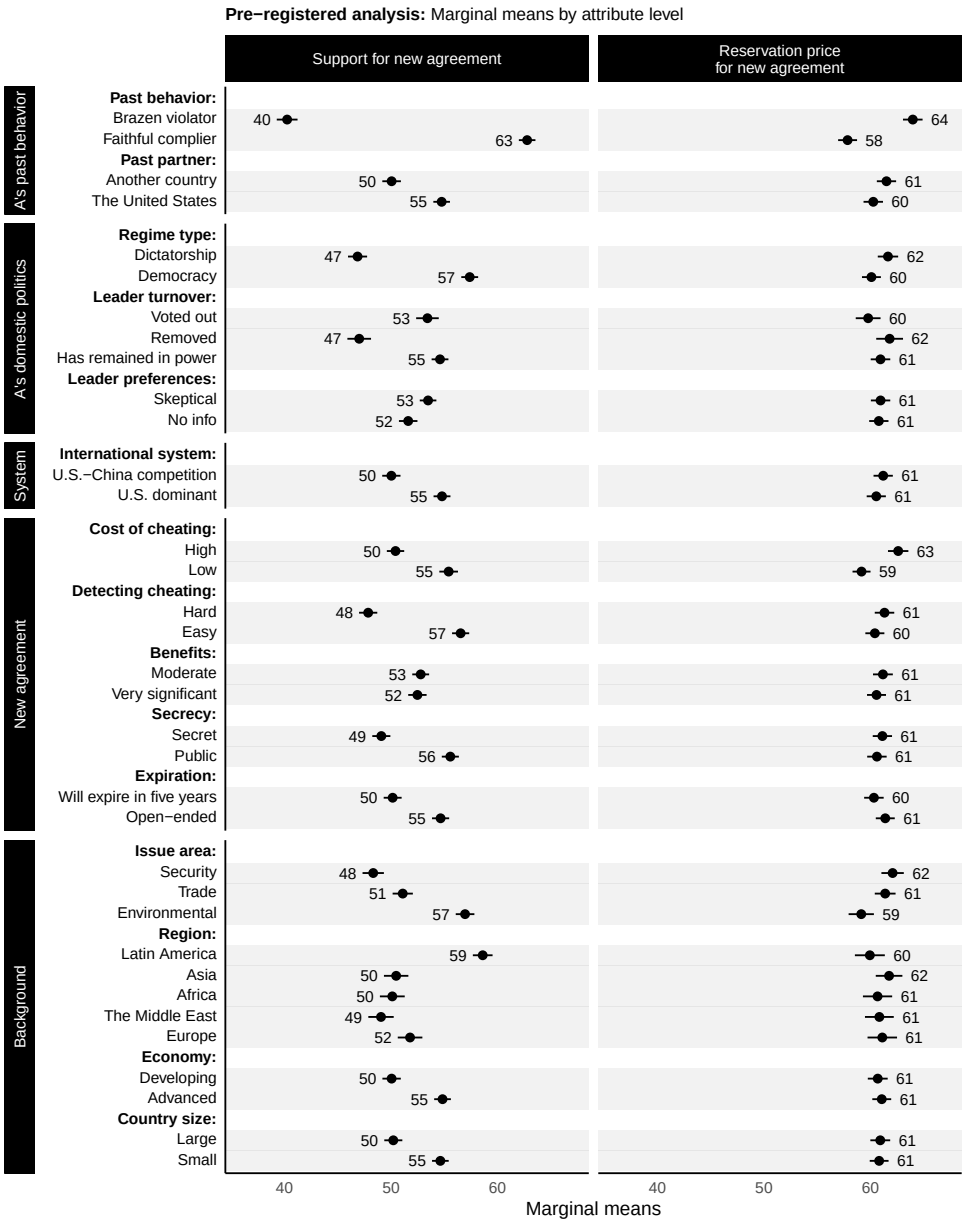
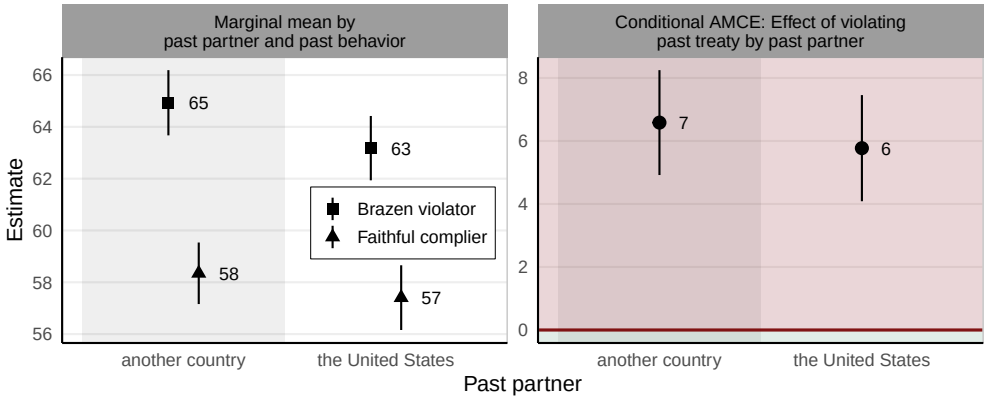



Figure 9: **Marginal Means:** describes the level of favorability toward profiles that have a particular feature level, ignoring all other features. Left column (pre-registered DV) uses data from the first 11 profiles, while the right column (exploratory DV) uses data from last 2 profiles.

**Analysis:** ($H_2$) AMCE of past behavior when past partner = another country.

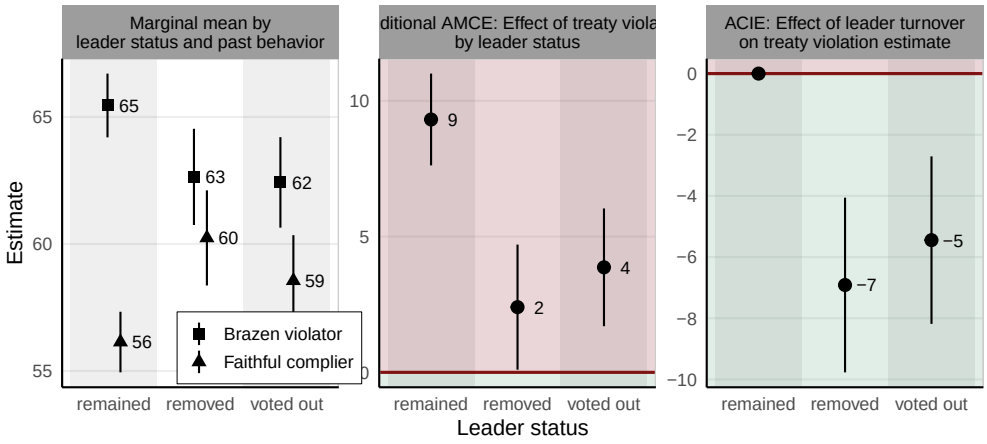
**Exploratory DV:** Reservation price for new agreement




(a) H2

**Analysis:** ($H_3$) ACIE of leader turnover * past behavior.

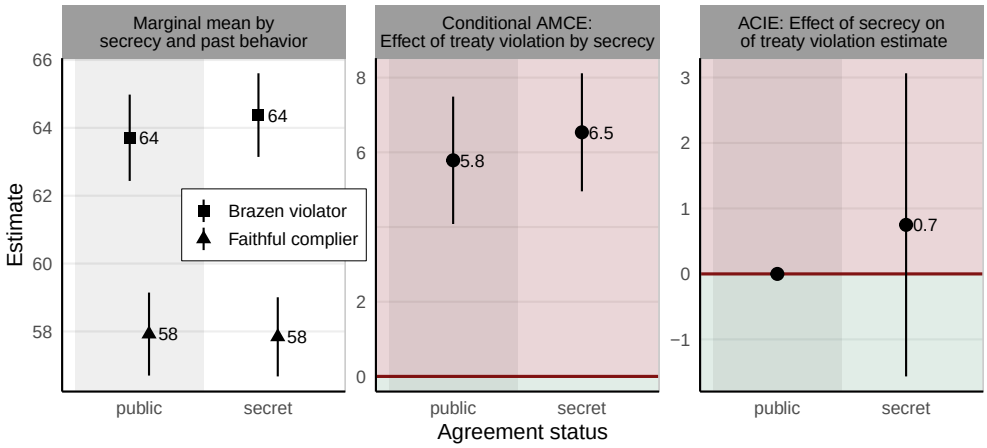
**Exploratory DV:** Reservation price new agreement




(b) H3

**Analysis:** (H4) ACIE of secrecy * past behavior.

**Exploratory DV:** Reservation price for new agreement




(c) H4

Figure 10: H2-H4

## C.2 Attention checks

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) 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.

Table 8 and Table 9 present results for our attention check questions.

Table 8: Pre-treatment attention check.

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

Table 9: 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) |

## C.3 Prolific Demographics

Table 10: 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) |

## C.4 Vignette Text

.

The vignette text in grey boxes below is copied verbatim from our survey. Respondents first encounter a transition page.

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.<br>is in Europe.<br>is a small country.<br>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.<br>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 11: Example of table format that respondents see

## C.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.6 Interpreting the Leader Turnover Interaction

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.[28]

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

[28] 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.

of non-cooperation is even more durable than the estimates based on our pre-registered contrasts suggest.

# D Vignette Experiment

## D.1 Attention and Manipulation Checks

### Pre-treatment attention screener

Table 12: Attention Check Pass/Fail Rates

| Attention Check Result | N | Proportion |
|---|---|---|
| Failed | 51 | 1.5% |
| Passed | 3316 | 98.5% |

### D.1.1 Post-treatment manipulation checks

Table 13: Manipulation Check Pass/Fail Rates

| Response | N | Proportion |
|---|---|---|
| **Power** | | |
| Incorrect | 152 | 4.5% |
| Correct | 2855 | 84.8% |
| **Accommodation** | | |
| Incorrect | 191 | 5.7% |
| Correct | 2920 | 86.7% |
| **Country** | | |
| Incorrect | 450 | 13.4% |
| Correct | 2808 | 83.4% |

## D.2 Prolific UK Demographics

Table 14: Sample Demographic Characteristics

| Category | N | Percent |
|---|---|---|
| **Income** | | |
| Less than £17,000 | 290 | 8.8% |
| £17,000 – £36,700 | 994 | 30.0% |
| £36,700 – £64,800 | 1004 | 30.3% |
| £64,800 – £81,400 | 417 | 12.6% |
| £81,400 – £199,000 | 432 | 13.0% |
| £199,000 or more | 19 | 0.6% |
| Prefer not to say | 158 | 4.8% |
| **Age** | | |
| 18-24 | 354 | 10.7% |
| 25-34 | 568 | 17.2% |
| 35-44 | 545 | 16.5% |
| 45-54 | 562 | 17.0% |
| 55-64 | 889 | 26.9% |
| 65+ | 392 | 11.8% |
| **Ideology** | | |
| Very left-wing | 161 | 4.9% |
| Left-wing | 710 | 21.4% |
| Slightly left-wing | 713 | 21.5% |
| Centre / middle of the road | 919 | 27.7% |
| Slightly right-wing | 526 | 15.9% |
| Right-wing | 245 | 7.4% |
| Very right-wing | 40 | 1.2% |
| **Gender** | | |
| Male | 1591 | 48.0% |
| Female | 1710 | 51.6% |
| Other | 13 | 0.4% |
| **Education** | | |
| Some secondary school or less (no qualifications) | 47 | 1.4% |
| GCSEs or equivalent (e.g., O-levels) | 397 | 12.0% |
| A-levels or equivalent (e.g., Scottish Highers) | 515 | 15.5% |
| Vocational / technical qualification (e.g., NVQ, BTEC, apprenticeship) | 415 | 12.5% |
| Undergraduate degree (e.g., BA, BSc) | 1328 | 40.1% |
| Postgraduate degree (e.g., MA, MSc, PhD, professional degrees) | 612 | 18.5% |

### D.3 Vignette Text and Randomization Procedure

After measuring foreign policy attitudes, the survey introduces our experimental vignette. All respondents first read the following passage.

> [vignette screen 1]
>
> On the following page, we will tell you about a hypothetical foreign policy interaction between Country **A** and other countries.
>
> Please read the scenario carefully and respond to the questions that follow to the best of your ability given the information provided.

On the next screen, they see:

> [vignette screen 2]
>
> **Scenario Background**
>
> Two foreign countries, Country **A** and Country **B**, have a history of close cooperation. Country **A** is a [democracy/dictatorship] in [Europe/Asia].
>
> The relationship between Country **A** and Country **B** is based on a long-standing international agreement that promotes economic, cultural, and security ties.
>
> **Country A has violated the agreement; wants to make it less favorable to Country B**
>
> In recent months, Country **A** has repeatedly and brazenly violated the terms of the agreement.
>
> Country **A** also publicly pressured Country **B** to renegotiate the agreement in Country **A**'s favor.

We then ask respondents to provide an initial assessment of Country A's reputation.

> We want to ask you about Country **A**, the state that demanded a renegotiation of the agreement. To what extent do you agree with the statements below? A score of 0 means that you do not agree at all and a score of 100 means that you agree completely.
>
> 1. Country A is the type of country that stands firm in foreign policy disputes (Toughness Reputation for A)

2. Country A is the type of country that complies with its international legal commitments (Compliance Reputation for A)

3. I approve of the behavior of Country A (Approval of A)

Respondents then view randomly varied information about how other countries respond to Country A in the wake of its non-cooperative behavior as well as randomly varied information about Country A's power relative to the country in question. They all read, "Please consider this additional information and then answer the questions that follow: ..."

Some respondents will be randomly assigned to evaluate the accommodation/non-accommodation choice of Country B, the victim of Country A's non-cooperative behavior. They read:

[vignette screen 3, *identity* = *B*]

**Country B [rejects/accommodates] Country A's demands; [ends/continues] cooperation**
The leaders of Country **B** deliberated about how best to respond to this pressure.

In the end, Country B [**condemned Country A's behavior and announced that it is terminating the agreement./expressed a willingness to accommodate Country A's demands and reaffirmed the importance of the partnership.**]

By way of background, Country A's economy and military are [**of roughly equal size to that of Country B./several times larger than that of Country B.**]

Other respondents will be randomly assigned to evaluate the accommodation/non-accommodation choice of a third country, the hypothetical Country C.

[vignette screen 3, *identity* = *C*]

**A third country [ends/continues] cooperation with Country A**
During the same period in which Country **A** has been violating its agreement with Country **B**, Country **A** has been negotiating an unrelated economic agreement with a different country, Country **C**.

The leaders of Country **C** deliberated about whether to continue these negotiations, given Country A's behavior.

In the end, Country **C** [**condemned Country A's behavior and announced that it is terminating the negotiations./reaffirmed the importance of its partnership with Country A and decided to proceed with the negotiations.**]

By way of background, Country A's economy and military are [**of roughly equal size to that of Country C./several times larger than that of Country C.**]

A final set of respondents will be randomly assigned to evaluate the accommodation/non-accommodation choice of the UK, where our respondents are based.

[vignette screen 3, *identity = UK*]

**The UK [ends/continues] cooperation with Country A**

During the same period in which Country **A** has been violating its agreement with Country **B**, Country A has been negotiating an unrelated economic agreement with the **United Kingdom (UK)**.

The leaders of the **UK** deliberated about whether to continue these negotiations, given Country **A's** behavior.

In the end, the **UK** [**condemned Country A's behavior and announced that it is terminating the negotiations./reaffirmed the importance of its partnership with Country A and decided to proceed with the negotiations.**]

By way of background, Country **A's** economy and military are [**of roughly equal size to that of the UK./several times larger than that of the UK.**]