Reputations for Resolve and Higher-Order Beliefs in Crisis Bargaining

Allan Dafoe¹,², Remco Zwetsloot¹,², and Matthew Cebul³

Abstract
Reputations for resolve are said to be one of the few things worth fighting for, yet they remain inadequately understood. Discussions of reputation focus almost exclusively on first-order belief change—A stands firm, B updates its beliefs about A’s resolve. Such first-order reputational effects are important, but they are not the whole story. Higher-order beliefs—what A believes about B’s beliefs, and so on—matter a great deal as well. When A comes to believe that B is more resolved, this may decrease A’s resolve, and this in turn may increase B’s resolve, and so on. In other words, resolve is interdependent. We offer a framework for estimating higher-order effects, and find evidence of such reasoning in a survey experiment on quasi-elites. Our findings indicate both that states and leaders can develop potent reputations for resolve, and that higher-order beliefs are often responsible for a large proportion of these effects (40 percent to 70 percent in our experimental setting). We conclude by complementing the survey with qualitative evidence and laying the groundwork for future research.

Keywords
bargaining, belief structure, conflict, game theory, survey experiment

¹Centre for the Governance of AI, University of Oxford, Oxford, United Kingdom
²Department of Politics and International Relations, University of Oxford, Oxford, United Kingdom
³Weiser Center for Emerging Democracies, University of Michigan, Ann Arbor, MI, USA

Corresponding Author:
Remco Zwetsloot, Manor Rd, Oxford OX1 3UQ, United Kingdom.
Email: remcozwetsloot@gmail.com
[In bargaining situations], each party’s strategy is guided mainly by what he expects the other to accept or insist on; yet each knows that the other is guided by reciprocal thoughts. The final outcome must be a point from which neither expects the other to retreat; yet the main ingredient of this expectation is what one thinks the other expects the first to expect, and so on. Somehow, out of this fluid and indeterminate situation that seemingly provides no logical reason for anybody to expect anything except what he expects to be expected to expect, a decision is reached.

—Thomas Schelling (Schelling 1960, 70)

Can both sides simultaneously have images of high resolve or is there a zero-sum element involved?

—Robert Jervis (Jervis 1978, 199)

Introduction

Navigating an international crisis can require incredibly complex inferences. Even seemingly straightforward strategies can backfire dramatically. Suppose that persons $A$ and $B$ are playing Chicken, the game in which two cars race toward each other and the goal is to get the opponent to swerve first. One popular strategem says that $A$ should throw her steering wheel out of the window in order to convince $B$ that she cannot swerve, thereby compelling $B$ to do so. Yet suppose that $B$ incorrectly believes that $A$ hid a spare steering wheel under her seat. If $A$ does not know that $B$ believes this, she may surrender her ability to swerve without actually becoming committed in $B$’s eyes, raising the chances of a tragic crash. Or suppose that $A$ incorrectly believes that $B$ believes that $A$ has a spare steering wheel. This belief might prevent $A$ from using a tactic that would have improved her odds of victory.

Much of the complexity in international crisis bargaining is attributable to higher-order beliefs—beliefs about beliefs (Schelling 1960; O’Neill 1999). While these complications are perhaps most apparent in hypothetical games like Chicken, examples abound in real-world international crises as well. For instance, in the years prior to the Berlin crisis, Premier Khrushchev had deliberately exaggerated the Soviet Union’s capabilities, and widely-discussed American concerns about a “missile gap” had convinced him that the Americans had fallen for the ruse. “[O]n the assumption that the Americans believed the Soviets were ahead in the arms race,” Khrushchev chose to escalate tensions over Berlin (Kaplan 1983, 304). Khrushchev was forced to back down, however, when he learned that US policymakers did not believe this at all—a fact that the Americans purposely signaled via intelligence leaks to NATO units that the US knew to be compromised by Soviet agents (Schlosser 2014, 284). Higher-order beliefs thus played a critical role in both the onset and resolution of one of the most significant crises of the Cold War.
In this article, we examine the theoretical and empirical import of higher-order belief dynamics for two crucial phenomena in international relations: resolve and reputations for resolve. Reputations, especially for resolve, have been a long-standing topic of scholarship, but most studies consider only first-order beliefs: whether and how $B$ updates his beliefs about $A$ on the basis of $A$’s behavior. We argue that taking higher-order beliefs into account substantially amplifies the importance of reputations for resolve. This is because resolve is interdependent: one side’s (perceived) resolve can affect the other side’s (perceived) resolve, meaning that small initial changes in beliefs can spiral into large overall changes, with potentially decisive consequences.

We evaluate our theory using a survey experiment conducted on a sample of quasi-elites recruited through Stephen Walt’s *Foreign Policy* blog. We begin by corroborating reputation scholars’ core assertion: states that behaved resolutely in the past are perceived to be more highly resolved in the present. We find that the effect of reputation on perceived resolve is larger than any other experimental manipulation, including the effect of material capabilities, and persists across a range of informational conditions and subgroups.

We then go a step further, advancing scholarship on reputations for resolve by demonstrating that respondents also form higher-order beliefs in line with our argument. Exogenous increases in $A$’s perceived resolve, induced by providing information about $A$’s behavior in disputes with states other than $B$, produce decreases in $B$’s perceived resolve, even though respondents learned nothing about the usual factors said to affect $B$’s resolve. The size of the parameter estimates we obtain suggest that higher-order beliefs can account for as much as 70 percent of the reputational manipulation’s effect on the balance of resolve, with a lower bound of roughly 40 percent. We thus answer Robert Jervis’s longstanding but still open question: there is indeed a strong zero-sum element to resolve.

Our design also permits a direct test of a competing hypothesis linking higher-order beliefs to resolve, which Jervis has referred to as the “domino theory paradox” and Daryl Press as the “never again theory”: if $A$ backed down in a recent dispute, $B$ will believe that $A$ will expend extra effort to restore its tarnished reputation for resolve, and therefore that $A$ will be more rather than less resolved in subsequent disputes (Jervis 1991; Press 2005). Our evidence is contrary to these hypotheses.

Our results reinforce recent studies demonstrating the importance of reputation in the domains of conflict, finance, and alliance politics, among others (see Crescenzi 2017, for a review). These studies all focus on the first-order effects of reputation. We show, however, that the story does not stop there. In doing so, we join a number of others who place higher-order beliefs at the center of international relations specifically (Schelling 1960; O’Neill 1999; Morrow 2014) and social, economic, and political life more broadly (Kuran 1995; Chwe 2001). These beliefs are what make resolve interdependent, and are therefore crucial to the onset and resolution of crises and wars.
Resolve, Reputation, and Higher-Order Beliefs

Much of international politics pivots around perceptions. A state’s ability to persuade, deter, and compel depends in large part on how that state is perceived—whether others consider its military capable, its threats credible, its promises reliable. As a result, states and leaders are motivated to act in ways that favorably shape others’ perceptions and expectations, i.e. to cultivate reputations for desirable qualities or behavioral tendencies.

One of the most important kinds of reputation in international politics is a reputation for resolve. We define resolve as follows:

Resolve: the probability that an actor is willing to stand firm in a crisis, given its beliefs about the world at that time.

We briefly explain how this behavioral definition relates to other common conceptualizations of resolve. Formalized mathematical models of crisis bargaining often define resolve as a state’s “value for war,” which is said to be a function of material factors such as military capabilities, the issue under dispute, and domestic audiences (see, e.g. Morrow 1989; Powell 1990; Fearon 1994; Sartori 2016). Psychological accounts of resolve instead emphasize internal traits such as dispositional determination, willpower, or “sticktoitiveness” (Kertzer 2016, 8-9). Here, we define resolve in terms of states’ behavior in a crisis, in which states may either back down or stand firm. In this sense, a state’s “value for war” and its dispositional traits are important factors that affect resolve—the probability that an actor stands firm—but they may not be the only ones, so they do not define resolve.

Understanding resolve in behavioral terms has several benefits. By clearly specifying what kind of behavior is (in)consistent with being resolved, our definition makes resolve measurable and facilitates empirical analysis. More importantly, our definition allows scholars to directly incorporate another essential determinant of whether A will stand firm in a crisis with B: A’s beliefs about B’s resolve. As exemplified by Chicken, that A’s crisis behavior depends on her beliefs about B’s resolve is what makes interstate crises complex strategic interactions. This relationship also speaks to the literature on reputations for resolve, which we briefly review below.

Building Reputations for Resolve

International crises can often be understood as “contests of resolve” (Morrow 1989, 941-42) in which the outcome is determined by one side’s ability to convince the other that it will not back down. In such contests, a history of resolute past behavior can be a valuable asset. We say that actors have a reputation for resolve when observers have formed beliefs about that actor’s tendency to stand firm in a certain class of disputes on the basis of that actor’s past behavior (Dafoe, Renshon, and Huth 2014).
The effect of a reputation for resolve remains a subject of debate. Proponents contend that reputations offer a path to successful commitment, deterrence, and compellence (Huth 1997), as costly past actions can corroborate threats and promises that might otherwise be dismissed as “cheap talk” (Fearon 1995; Trager 2017). These scholars argue that an “undesired image can involve costs for which almost no amount of the usual kinds of power can compensate” (Jervis 1970, 6), and so reputations are “one of the few things worth fighting over” (Schelling 1966, 124). In support of this position, a number of recent studies find evidence that states’ past behavior shape how other states act toward it.7

Others, however, dispute the link between past actions and present behavior. For instance, Mercer argues that the desirability of an action conditions how people interpret the disposition of its initiator, concluding that “people do not consistently use past behavior to predict similar behavior in the future” (Mercer 1996, 45-47, 212). Similarly, Press claims that “blood and wealth spent to maintain a country’s record for keeping commitments are wasted,” since opponents form assessments of credibility largely based on their perceptions of states’ interests and capabilities, not their past behavior (Press 2005, 10). Detractors also aver that evidence of reputational effects is largely observational and often indirect, and is therefore inconclusive.

Thus, one contribution of our article is to bring experimental methods to bear on a basic question: do observers use a state’s past behavior to predict what it is likely to do today?8

**HREP**: Perceptions of a state’s resolve will (a) increase if it stood firm in past crises and (b) decrease if it backed down in past crises.9

In addition, we also seek to advance the reputation debate beyond HREP toward a discussion of higher-order beliefs. Note that HREP focuses purely on how first-order beliefs—beliefs about an actor’s traits or behavioral tendencies—change in response to past behavior. Such arguments are at the center of most theoretical accounts of reputational dynamics. First-order beliefs are, however, not the only kinds of beliefs that matter. In the next section, we contend that reputations can also affect higher-order beliefs—beliefs about beliefs (about beliefs, etc.)—with important consequences for state behavior.

### Reputational Effects and Higher-Order Beliefs

Consider a hypothetical dispute between states \( A \) and \( B \), in which \( B \) is considering whether to escalate. Initially, \( B \) has some sense of how likely \( A \) would be to stand firm (\( A \)'s resolve), though he cannot be sure. \( B \) then observes \( A \) behaving resolutely in some other dispute, and after considering that \( A \) might, say, be more militarily capable than \( B \) originally thought, \( B \) updates his beliefs about \( A \)'s resolve accordingly. We can call this a first-order reputational effect, in which one side updates its
beliefs about its opponent’s characteristics or behavioral tendencies solely on the basis of that opponent’s past behavior.

In many cases, however, higher-order reasoning further complicates the story. For example, A might know that B was paying careful attention to her behavior in the other dispute. If so, A might conclude that, since she stood firm, B has become less resolved, and this could further embolden A to also stand firm against B. If B expects A to be thinking along these lines, he might conclude that there is little he can do to make A back down, further reducing B’s resolve. If actors make inferences in this way, the first-order reputational effect will be joined by higher-order effects. These two effects together make up the total reputational effect, i.e. the full shift in the balance of resolve between two actors due to a single initial reputational shock (see Section “Do Higher-Order Beliefs Contribute to Reputational Effects?” for a formalization). Indeed, higher-order beliefs can theoretically produce large substantive effects, as relatively small first-order reputational effects cascade through higher-order belief chains into large changes in the balance of resolve.

Put another way, our argument is that resolve is a function of not just the material factors emphasized in the modeling literature or the internal traits emphasized in psychological accounts, but also states’ beliefs about their opponents’ resolve—A’s resolve is decreasing in (her beliefs about) B’s resolve, and vice versa. In short, resolve is interdependent. Furthermore, that resolve is interdependent implies that higher-order beliefs may play a central, albeit under-appreciated, role in international crises—A’s beliefs about B’s resolve (which dictate A’s own resolve) are themselves a product of A’s higher-order beliefs about B’s beliefs about A’s resolve, and vice versa. To express our arguments in terms of reputations, we contend not just that states and leaders can possess reputations for resolve, but also that these actors are aware that such a reputation may decrease the likelihood that their opponents stand firm during crises, and so update their own resolve accordingly. In short, states may expect that their reputations precede them.

\[H_{INT}: \text{Perceptions of a state’s resolve will (a) increase when there is a decrease in perceptions of its opponent’s resolve, and (b) decrease when there is an increase in perceptions of its opponent’s resolve.}\]

There are two principal reasons why we may not expect this hypothesis to hold. First, states may not form higher-order beliefs if they lack the ability to do so. Kertzer, for example, argues that in many situations with higher-order uncertainty about resolve “the complex nature of the decision-making environment actors face stretches far beyond the limits of human cognition” (Kertzer 2016, 149; see also Mercer 2012). And even if they have the ability, states may be insufficiently confident in their higher-order beliefs to incorporate them into their assessments of resolve. If this were the case, we would expect to see no evidence of interdependence.

Second, \[H_{INT}\] is not the only possible relationship between higher-order reasoning and reputational effects. \[H_{INT}\] implies that higher-order reasoning magnifies
first-order effects—states that stand firm earn larger reputational bonuses, whereas states that back down suffer larger penalties. In what Jervis labels the “domino theory paradox” and Press “never again theory,” however, higher-order effects run counter to the first-order effect (Jervis 1991; Press 2005). Domino theory, which undergirded much of US foreign policy during the Cold War, holds that if a state backs down in one situation, observers will infer that it is more likely to do so in other situations as well. The paradox is that the corresponding higher-order beliefs may be “self-defeating” (Jervis 1991, 36-37):

An actor who has had to back down once will feel especially strong incentives to prevail the next time in order to show that the domino theory is not correct, or at least does not apply to him. In other words, an actor who believes the domino theory—or who believes that others accept it—will have incentives to act contrary to it. Indeed, statesmen sometimes respond to a defeat by warning others that this history will make them extremely resistant to retreating again. Furthermore, if others foresee this, they will expect a defeated actor to be particularly unyielding in the next confrontation.

In short, the domino theory paradox predicts that, in certain circumstances, higher-order reasoning can mitigate or even reverse first-order reputational effects.

**H\textsubscript{DTP}:** Perceptions of a state’s resolve will increase when it backed down recently and has a prior reputation to recover.

**Research Design**

The scientific study of reputations is limited by their inaccessibility. Reputations consist of perceptions, which exist in people’s minds and are not directly observable. Even records of private deliberations, free of any incentives to dissemble, may lead us astray. To the extent that an opponent’s reputation is a constant during a crisis and is common knowledge among decisionmakers, discussions are likely to focus on new information such as troop movements even if reputation matters (Weisiger and Yarhi-Milo 2015). Reputational inferences, given their evolutionary importance, may also be so automatic that they are made subconsciously (Bowles and Gintis 2011). These difficulties were acknowledged by Snyder and Diesing, who noted that their inability to find much evidence of reputational inferences in case histories “may be only an artifact of the record” (Snyder and Diesing 1977, 496).

To overcome these challenges, we employ a scenario-based survey experiment. Respondents are told about a scenario, features of which are randomly varied or omitted, and then asked about their opinions or beliefs. Survey experiments are especially suitable for research questions about how people incorporate, interpret, and act on particular types of information (Mutz 2011), and these are precisely the questions in which scholars of reputation are interested. Below, we review the survey design and the sample in more detail.
Survey Vignette

Respondents read an abstract scenario about two countries (A and B) engaged in a serious territorial dispute. The features of the scenario are presented in a bullet list format, each assigned with some probability independent of each other (a full factorial design). Each feature has several different levels, including its omission. For the full text of the survey and treatment allocations, see Online Appendix B.

Respondents were first informed of each country’s regime type, either democracy or dictatorship (some received no information on regime type). Respondents then learned about the military capabilities of each country, with A having either “substantially stronger military forces” than B, being “about equally strong,” or B being substantially stronger (again, some received no information about capabilities). In all conditions except no information, respondents are also told that neither country has nuclear weapons.

The respondent then reads the reputational feature of the scenario, State A’s history of crisis interactions. This bullet involves two variables: whether A stood firm or backed down in past crises, and whether A’s previous conflicts were against B or other countries.10

According to most impartial observers, of the three most recent major international crises that Country A has faced against [other countries/Country B], Country A [did not back down and Country A achieved its aims/backed down in each crisis and failed to achieve its aims].

There is also a special Domino Theory Paradox version of this manipulation, in which A stood firm in three past crises, but then backed down in a fourth, most recent crisis. In all conditions except no information, respondents are also informed whether A’s past crises occurred under the current or a previous leader.

Lastly, the vignette includes several other experimental features related to the history of the dispute, as well as recent threats and promises.11 The scenario concludes by stating that the crisis is serious and that many observers consider major military action in the near future likely. Respondents are then asked the primary outcome question for both countries:

What is your best estimate, given the information available, about whether Country A/B will back down in this dispute?

Respondents have five options, ranging from “very unlikely” (0 percent to 20 percent chance) to “very likely” (80 percent to 100 percent chance). Note that this question exactly matches our definition of resolve in Section “Resolve, Reputation, and Higher-Order Beliefs.”12

Before detailing our sampling strategy, we briefly address the benefits and drawbacks of our abstract survey design. Our survey vignette describes a crisis between
two abstract states, $A$ and $B$, and the scenarios respondents read were typically no more than 150 words. The benefit of this approach is its flexible simplicity; short vignettes are less taxing for respondents, and abstracting away from real-world states may permit cleaner manipulation of the concepts of interest (though not necessarily; see Dafoe, Zhang, and Caughey 2018). The attendant drawback is a decreased level of realism in abstract survey scenarios, relative to vignettes that feature actual countries (Tingley 2017). Reassuringly, however, recent research employing both abstract and real-world crises scenarios recovers reputational effects of a very similar magnitude to those presented below (Renshon, Dafoe, and Huth 2018).

Regardless of the design, it is of course not possible to create environments akin to those faced by real-life decision-makers through surveys alone, and it is therefore best to complement survey experiments with other kinds of evidence. Though more systematic observational work will have to await future research, we offer some preliminary supportive evidence for our claims in Section “Discussion: Interdependent Resolve in Real Life,” and discuss the conditions under which evidence of interdependent resolve is most and least likely to be found.

Sample

We administered the survey to a convenience sample of respondents recruited through Stephen Walt’s Foreign Policy blog. On August 1st, 2011 Walt posted a blog entry inviting readers to “Become a data point!” (see Online Appendix B.1). From this advertisement, over 1,000 respondents took the survey.

This sampling strategy was intended to mitigate potential biases that may arise from using regular citizens to proxy for elite policymakers. The general concern is that elites are better informed about and more experienced with foreign policy decision-making than the average citizen—as a result, they are more likely to employ higher-order strategic reasoning and consider their opponents’ perspective (Hafner-Burton, Hughes, and Victor 2013; Hafner-Burton et al. 2014). Ideally, then, one would run our experiment on key decision-makers, such as military leaders, foreign policy advisors, and politicians. Unfortunately, such subjects are rarely accessible, and even if they are it is usually as part of a small convenience sample.

To approximate our elite population of interest, we instead sought to recruit a sample of quasi-elite respondents who are abnormally well-informed about and interested in foreign policy, whose backgrounds and world views more closely parallel those of high-level elites. We expected a sample drawn from Walt’s Foreign Policy readership to be highly educated and knowledgeable of foreign policy, and that is indeed the case. Of those 87 percent who answered the demographic questions, 83 percent reported having a college degree or higher, and 50 percent a postgraduate degree. Moreover, 60 percent claimed to have “particular expertise with foreign affairs, military affairs, or international relations.” Politically, the group leans Democratic, as 89 percent claimed to agree more with the policies of Democrats (11 percent more with those of Republicans). The respondents were 88 percent
male, which is far from representative of the general population but not obviously unrepresentative of foreign policy elites. More details about the sample are provided in Online Appendix A.

To be sure, then, our sample remains an imperfect approximation of foreign policy elites. Still, we argue that these imperfections will likely lead us to underestimate the effect of reputation and higher-order beliefs on perceived resolve. To start, the Democratic bias should cut against our proposed hypothesis, as liberals are generally less likely than conservatives to invoke concerns over credibility, reputation, and honor in international affairs (Trager and Vavreck 2011). And while respondents may also hew closer to Stephen Walt’s particular foreign policy views, Walt has repeatedly argued against the importance of reputation. Lastly, our sample likely remains less experienced with actual foreign policy decision-making than true elites. But as mentioned above, experience is linked to higher-order strategic reasoning, and Tingley and Walter (2011) show that experienced players care more about reputation than inexperienced ones. If anything, then, our results likely underestimate the reputational effects that would be found among actual elites.

In short, given the cost and difficulty in obtaining elite samples, our sampling design is a reasonable first step toward the empirical study of higher-order beliefs in international crises. Importantly, we think it highly unlikely that our sample would produce an opposite effect of reputation on perceived resolve relative to true elites—our respondents may be less attentive to reputation, but this should only depress the magnitude of reputational effects, not reverse their direction.

**Results**

To preview our results, we reach two main findings. First, reputations for resolve matter; when respondents learn that a state has stood firm in past crises, they consider it much more likely to stand firm today. Second, resolve is interdependent; we find that increases in A’s perceived resolve are associated with decreases in B’s perceived resolve, and estimate that higher-order belief updating is responsible for a large proportion of the total observed effect of past behavior on the balance of resolve.

**Can States Build Reputations for Resolve?**

We begin with $H_{REP}$, which considers whether a country’s past behavior affects perceptions of its current resolve. We find strong evidence that it does. Respondents who learned that “Country A did not back down and Country A achieved its aims” in past crises thought that A was more likely to stand firm than those that received no information about past behavior (a 10 percentage point increase, from 60 percent to 70 percent). Similarly, when A “backed down in each crisis and failed to achieve its
aims,” respondents thought A was roughly 10 percentage points more likely to back down. These effects are both highly significant (Figure 1).

Moreover, this reputational treatment has the largest effect of all manipulations in the survey. The effect of going from a history of backing down to a history of standing firm is about 20 percentage points, which represents a quarter of the resolve variable’s total range.\(^{16}\) This reputational effect is roughly twice the size of the second-largest effect, information about power: shifting from B having “substantially stronger military forces” to A having “substantially stronger military forces” increases perceptions of A’s resolve by roughly 10 percentage points.\(^{17}\)

Readers may wonder whether this potent reputational effect is driven only by a subset of respondents. Reputation skeptics contend that even if past behavior matters, it does so only to the extent that “a decision maker uses an adversary’s history of keeping commitments to assess the adversary’s interests or military power” (Press 2005, 21). An observable implication of this view is that we should only see effects of past behavior on respondents that lack information about the balance of power. We find, however, that the reputational effect persists among respondents who were told about military capabilities (Figure 12, Online Appendix C.1).\(^{18}\) And while the reputational effect decreased slightly as scenario complexity increased,\(^{19}\) it remained above 10 percentage points even among respondents who received the maximum amount of information about the scenario (Figure 11, Online Appendix C.1). Reputational effects also persist across state leaders and across demographic subgroups, including gender, education, political affiliation, and cultural background. In sum, we find strong support for HREP.

In contrast, we find no evidence for HDTP. According to HDTP, a state that backs down once after a history of standing firm will be perceived to be more resolved to stand firm in the present, as observers expect it to try to re-establish its lost reputation for resolve. Yet as Figure 1 shows, backing down once after a history of always standing firm reduces perceived resolve by about 8 percentage points.

**Figure 1.** Effect of A’s past behavior on perceptions of A’s resolve, relative to respondents that received no information about past behavior (baseline probability of standing firm ≈ 60 percent). X-axis displays percentage point change; horizontal lines represent 90 percent and 95 percent CIs.
compared to a history of standing firm \( (p < 0.0001) \), nearly returning perceived resolve to its baseline probability in the no-information condition. In other words, backing down once almost entirely eliminates the reputational gains that \( A \) achieves by standing firm in the initial three crises.

**Do Higher-Order Beliefs Contribute to Reputational Effects?**

Typically, reputational effects like the ones reported above are interpreted in terms of first-order updating: \( A \) takes an action, and observers update their beliefs about \( A \). Yet as discussed in Section “Reputational Effects and Higher-Order Beliefs,” the total reputational effect of \( A \)’s past behavior on \( A \) and \( B \)’s resolve may also consist of higher-order effects. To test this idea, we need a treatment that affects beliefs about \( A \)’s resolve, but cannot affect \( B \)’s resolve through any channel except (beliefs about) \( A \)’s resolve. If such a treatment affects \( B \)’s resolve, then we can conclude that perceptions of \( B \)’s resolve are influenced by perceptions of \( A \)’s resolve—in other words, that respondents form higher-order beliefs and that resolve is interdependent (\( H_{INT} \)).

It is useful to define the relevant estimand and identifying assumptions more formally. Let \( R_i \) denote the resolve of agent \( i \), \( k \) denote a level of belief updating, and \( \hat{R}_i^k \) denote an observer’s beliefs about actor \( i \)’s resolve after the \( k \)th level of updating. Next, define \( \Delta \hat{R}_i^k \) as the change in beliefs about actor \( i \)’s resolve after the \( k \)th level of updating, i.e. \( \hat{R}_i^k - \hat{R}_i^{k-1} \). In these terms, the “direct” or “immediate” effect of our reputation treatment \( T \) (information about \( A \)’s past behavior) is given by \( \Delta \hat{R}_i^1 = \hat{R}_i^1 - \hat{R}_i^0 \), where \( \hat{R}_i^0 \) refers to perceptions of \( i \)’s resolve before observing \( T \) (i.e. among respondents in the “no information” condition).

Now, suppose that observers go through two levels of belief updating following reputation treatment \( T \) (this updating process is shown graphically in Figure 2). Our argument suggests that we should see \( \Delta \hat{R}_A^1 > 0 \), and, by interdependence, \( \Delta \hat{R}_B^2 < 0 \). In this situation, our estimand, which we will call the *interdependence coefficient* and label \( \theta \), is given by \( \frac{\Delta \hat{R}_B^2}{\Delta \hat{R}_A^1} \). This ratio tells us by how much an immediate one-unit change in \( A \)’s perceived resolve subsequently changes \( B \)’s perceived resolve. For example, if a 10 percentage point increase in \( A \)’s resolve decreases \( B \)’s resolve by 5 percentage points, \( \theta = \frac{-5}{10} = -0.5 \). More generally, if we assume that the interdependence of resolve is symmetric across actors and constant across levels of updating, we have \( \theta = \frac{\Delta \hat{R}_B^k}{\Delta \hat{R}_A^k} = \frac{\Delta \hat{R}_B^1}{\Delta \hat{R}_A^1} \) for all \( k \in \mathbb{N} \). We can then empirically estimate \( \theta \) by taking the ratio of the observed total effect of treatment on \( B \)’s and \( A \)’s resolve, \( \frac{\Delta \hat{R}_B}{\Delta \hat{R}_A} \), and state our interdependent resolve hypothesis more precisely as \( H_{INT} : \theta < 0.20 \).

As mentioned above, our treatment must fulfill two conditions for this estimation strategy to succeed, analogous to the identifying assumptions in instrumental...
variable analysis: (C1) \( \Delta \hat{R}_A^1 > 0 \) (instrument strength) and (C2) \( \Delta \hat{R}_B^1 = 0 \) (exclusion restriction). Notably, most treatments do not satisfy C2. For example, the fact that A stood firm against B in the past could lead to inferences not just about A’s resolve, but also about B and the A-B dyad, such as B’s dispositional determination or domestic audiences, violating our identifying assumption. Thus, we cannot use information about A’s past behavior against B to estimate the interdependence of A and B’s resolve.

However, information about A’s past actions against another country is a treatment that likely satisfies C2. When told about A’s extra-dyadic behavior, observers can make inferences about how much A values its reputation, whether A’s domestic public is liable to punish leaders for backing down, or any other factor that shapes A’s resolve. But the treatment is uninformative about the A-B relationship, the territory under dispute, or other factors that could reasonably shape perceptions of B’s resolve—except, that is, perceptions of A’s resolve. This treatment, then, allows for a clean test of our interdependence hypothesis.

The results of this test are displayed in Figure 3. We find that A’s behavior against other countries significantly affects perceptions of B’s resolve as expected—respondents who were told that A stood firm against other countries in the past assess B’s resolve to be roughly 7 percentage points lower relative to baseline, and A backing down against other countries leads to a similar increase in perceptions of B’s resolve. For A standing firm in extra-dyadic crises, \( \hat{\theta} = \frac{-0.065}{0.091} = -0.71 \), and for A backing down in extra-dyadic crises, \( \hat{\theta} = \frac{-0.074}{0.098} = -0.75 \). Both of these effects are statistically significant, lending strong support to \( H_{INT} \).
These results offer compelling evidence that reputational effects exist, and that they are compounded by higher-order beliefs. Still, we have yet to specify precisely how important are higher-order beliefs. Can we estimate the proportion of the total reputational effect—the effect of past behavior on the overall difference in resolve between A and B—that is attributable to higher-order belief updating? To help us answer this question, define the interdependence multiplier (IM) as the factor by which a first-order reputational effect should be multiplied in order to obtain the total reputational effect. In our simple formalization, the magnitude of the interdependence multiplier depends on two parameters: the interdependence coefficient, θ, and the number of levels of belief updating, which we label n. Given our model (and assuming |θ| < 1), Online Appendix D.2 derives the IM to be, for any n,

\[ IM = \left( \frac{1 - |\theta|^n}{1 - |\theta|} \right) \]

which converges to \(1/(1 - |\theta|)\) as \(n \to \infty\), a situation with “common knowledge” in which everybody knows about a fact or event, knows that everybody knows it, and so on.

Figure 4 plots the magnitude of the IM across hypothetical values of θ and n. At one extreme, if we assume that actors engage only in first-order reasoning (the light blue line), the IM is always 1, and the total reputational effect is equal to the first-order reputation effect. This, implicitly, has been the assumption in most past discussions of reputation. At the other extreme, if it is plausible to assume common knowledge (the red line), the total reputational effect is more than twice as large as the first-order effect for any \(|\theta| > 0.5\), and nearly four times as large at \(|\theta| = 0.73\) (the average of our \(|\theta|\) estimates calculated from Figure 3 above). Between these extremes, the IM’s magnitude is substantial across a range of parameter values,
underscoring that higher-order beliefs can be responsible for a large proportion of any given total reputational effect.

We can now use the $IM$ to estimate the effects of higher-order beliefs on perceived resolve in our survey experiment. As stated above, our average estimated $\theta = -0.73$, and the average total effect of the reputation treatments presented in Figure 3 is roughly 16 percent. To estimate how much of that 16 percent change is attributable to higher-order beliefs, let us assume that $n = 3$. This implies $IM = \frac{1 - (-0.73)^3}{1 - 0.73} = 2.26$. The estimated first-order effect is then simply the total effect divided by the $IM$, $\frac{16}{2.26} = 7.08$. This leaves roughly 9 percentage points attributable to higher-order reasoning—a sizable 55 percent of the total effect.

We can also more intuitively verify this result by beginning with the first-order effect, and then building out via $n = 3$ reasoning to the total effect. Suppose that, as we just estimated, respondents on average perceive $A$ to initially be 7.08 percentage points more likely to stand firm when $A$ has stood firm repeatedly in past cases ($\Delta R^1_A = 7.08$). The respondent reasons, however, that $B$ will become less resolved after inferring this change. Specifically, $B$’s resolve decreases by $7.08 \times 0$ percentage points ($\Delta R^2_B = -5.17$). This, in turn, will increase $A$’s perceived resolve by $-5.17 \times \theta$ ($\Delta R^3_A = 3.77$). The total reputational effect is therefore $|\Delta R^1_A| + |\Delta R^2_B| + |\Delta R^3_A| = 16.02 \approx 16$ percentage points, which is indeed the total reputational effect that we observe in the data (and is also $7.08 \times IM$).

This estimation strategy is limited by our inability to directly observe $n$, respondents’ level of higher-order reasoning—we assumed $n = 3$ above, but the true value could be higher or lower. While we therefore cannot definitively identify the proportion of the total reputational effect attributable to higher-order beliefs, we remain
confident that these beliefs drive a substantial portion of the effect in this case, for several reasons. First, we can quantify our uncertainty by deriving bounds. As our survey recovers reputational effects that are inconsistent with mere first-order reasoning, we use $n = 2$ as a lower bound. In this case, the $IM = 1.73$, and higher-order effects are responsible for about 42 percent of the total reputational effect. The upper bound is represented by common knowledge ($n = \infty$). In this case, the $IM = 3.7$, and the higher-order effects constitute about 73 percent of the total reputational effect (see Online Appendix D.2). Note that our reputation treatment had a relatively large effect on $B$’s perceived resolve ($|\hat{\theta}| = 0.73$ is large), which results in correspondingly large higher-order effect estimates even assuming low levels of higher-order reasoning (Figure 4 is instructive). Of course, the interdependence coefficient $\theta$ might vary in size in different settings or contexts—we discuss this possibility extensively in Section “The Degree of Interdependence.”

Second, numerous other studies and examples discussed in Section “Higher-Order Beliefs” suggest that at least some degree of higher-order reasoning, be it conscious or intuitive, is a regular feature of human cognition in political, economic, and social settings—though higher-order belief chains can become prohibitively complex, $n = 2$ reasoning is relatively common. This reassures us that $n = 2$ is a reasonable lower bound to estimate higher-order effects.

In sum, we find clear evidence that higher-order beliefs are responsible for a large portion of our observed reputational effects—even restricting the calculations to a minimal level of higher-order reasoning, the proportion approaches 50 percent. In the next section, we move beyond our survey data to discuss real-world applications and implications of our argument.

**Discussion: Interdependent Resolve in Real Life**

Above, we provided evidence of higher-order reputational dynamics, and argued that the magnitude of these effects depends on the order to which people form beliefs ($n$) and the degree of interdependence ($\theta$). Our abstract survey vignette allows us to observe the effect of higher-order belief updating in a way that is difficult to do in the real world. Still, some uncertainty remains as to whether and how the interdependence of resolve manifests in the messier context of real-world bargaining situations. This section discusses the circumstances under which we expect higher-order beliefs and interdependent resolve to matter most, and the attendant implications of our argument for real-world crisis bargaining.

**Higher-Order Beliefs**

As discussed in Section “Reputational Effects and Higher-Order Beliefs,” one broad issue is whether decision-makers engage in higher-order reasoning (Kertzer 2016; Mercer 2012). If they do not, our arguments have little relevance to the real world. Fortunately, evidence from a variety of contexts suggests that explicit and implicit
higher-order reasoning is common, and does not require especially sophisticated agents.

First, experimental evidence from behavioral economics suggests that most people often form at least second-order beliefs, and many reach higher orders as well (e.g. Nagel 1995; Camerer, Ho, and Chong 2004). In psychology, researchers have found that the magnitude of the famous “bystander effect” depends on whether the context allows bystanders to form beliefs about other bystanders’ beliefs (Thomas et al. 2014, 2016). A third example comes from a recent field experiment in Benin, where voters who learned information about politicians’ behavior punished and rewarded performance only when they thought that others knew this information as well, suggesting a higher-order understanding of coordination dynamics (Adida et al. 2020). Many other studies also find significant effects from interventions targeted at higher-order beliefs (Bicchieri 2016; Mildenberger and Tingley 2019).

Moreover, scholars have produced abundant evidence of higher-order reasoning among policymaking elites in high-stakes situations. At the domestic level, autocrats faced with potentially unsatisfied publics go to great lengths to create impressions of widespread support for their rule, or, if that fails, at least keep everyone guessing about each other’s beliefs (Kuran 1995; Chwe 2001, 20-21). Internationally, as McManus (2014, 726) illustrates with the case of Israel-US bargaining over Iran’s nuclear program, states often attempt to stake their allies’ reputation on supporting them, believing that their allies will then believe themselves bound to act (see, e.g. Jervis 1978, 180; Trager 2017, Ch. 1, for more examples). And conflicts are often said to end only when “opponents succeed in coordinating their expectations” (Slantchev 2003, 621; see also Carter 2017). In these and many other examples, beliefs about beliefs are the driver and focus of significant strategic contention.

To be clear, our argument does not require that actors always run through every step of the inferential process in a deliberate or conscious way. Higher-order beliefs can be incorporated into decision-making heuristically, implicitly, and subconsciously. As Chwe (2001, 78) argues for the case of driving, “I stop at a red traffic light out of habit, but a fully specified argument for doing so would involve an infinite regress: I stop because I think that other people are going, and I think that other people are going because I think that they think that I am stopping, and so on.” Developmental psychologists have found that children display a great degree of higher-order understanding early on in life, implicitly engaging in complex theorizing about the mental states of others long before they can explicitly articulate their reasoning (Wellman 2014). Higher-order belief dynamics also often become embedded in cultural and legal norms (O’Neill 1999; Ridgeway 2011; Morrow 2014).

The Degree of Interdependence

While it is therefore clear that real-world actors do engage in higher-order reasoning, the extent to which first-order reputational effect are amplified also depends on the degree to which resolve is interdependent, i.e. the extent to which an initial change in
A’s resolve affects B’s resolve. Here, contextual factors are likely to play a large role. We identify and discuss four such factors.

First, the degree of interdependence depends on the extent to which the context resembles a prototypical contest of resolve. The central features of such contests are (1) the costs of conflict are so great that losing is preferred to both sides standing firm, but (2) the issue under dispute is sufficiently valuable for a coercive victory to be preferred to a peaceful compromise (Morrow 1989, 941). We expect resolve to be most interdependent in conflicts that most closely approximate these conditions. Nuclear crises are often cited as the quintessential examples of such contests (Schelling 1966; Powell 1990), but they are by no means the only ones. In the 1898 Fashoda Crisis, for example, France dispatched a mission to Egypt in an attempt to force Britain to make concessions, but withdrew when it became convinced that Britain was more resolved than it initially believed (Trachtenberg 2012, 13-16).26

More contemporary examples can be found in proxy conflicts like the ongoing Syrian civil war. Direct conflict between the U.S. and Russia over Syria seems prohibitively costly, yet Damascus remains a valuable prize. In these circumstances, resolve can be highly interdependent. Indeed, this interdependence featured regularly in debates about U.S. intervention in Syria. Critics argue that Obama’s failure to enforce the infamous chemical weapons red-line in 2013 undermined U.S. deterrence, paving the way for Russian intervention in 2015—Putin might have been compelled to stay out, were it not clear (to both parties) that the U.S. was irresolute. At the same time, Obama’s reluctance to engage in Syria resulted in part from his belief that Russia would counter-escalate in response to limited U.S. intervention, especially after Russia stepped up its involvement in 2015.27 In other words, U.S. lack of resolve fueled Russian resolve, further depressing U.S. resolve.

Second, interdependence is also affected by the extent to which actors are able to act strategically, conditioning their behavior on what they think others are likely to do. Actors may fail or be unable to do so for various reasons. A prominent example is found in the expansive literature on credible commitments, which discusses commitment devices that may leave actors unconditionally resolved in a crisis. The moment actors truly commit to a strategy—when they throw their only steering wheel out of the window—they have effectively set their interdependence coefficient to 0: even if they subsequently change their beliefs about their opponent’s resolve, their own course of action is already set. That said, absolute commitments are exceedingly difficult and risky to make.28 In the real world, then, resolve is almost always interdependent at least to some extent.

A third and related factor is the concentration of decision-making authority: when leaders have the ability to change course quickly, their resolve is likely to depend more on the other side’s resolve. If, on the other hand, authority is diffuse, it will be difficult for a country to change course in the face of new information. One example is the delegation of military decision-making to local commanders, who can then choose to stand firm or back down in their theater of operations. Such delegation may be necessary from a practical perspective, but it also means central decision-
makers have less flexibility. Another obstacle to short-term policy change in response to belief updating is the number of veto players in a political system (Tsebelis 2002). To the extent that this number is correlated with regime type, there is reason to think that the resolve of democracies is less interdependent than that of their autocratic counterparts.

Lastly, a fourth factor affecting the degree of interdependence is the observability of resolve. An actor’s resolve is more likely to influence its opponents' calculations when it is readily observable. Resolve is likely to be more observable the more an actor’s deliberative processes are public, or when opponents share a cultural understanding of the meaning of certain behaviors or events (O’Neill 1999, 153-54). Strategic intelligence can also play an important role. The interdependence of resolve was on clear display, for example, during the Washington Disarmament Conference in 1921, when the United States was able to break Japan’s ciphers and read Tokyo’s private diplomatic communication. One such message stated the absolute minimum naval tonnage ratio that the Japanese government would be willing to accept. “Knowing how far Japan could be pushed... allowed the United States to do so with full confidence, merely waiting for the Japanese to give in” (Bauer 2013, 211). When Japan’s (lack of) resolve became perfectly observable to the US, American resolve dramatically increased in response.

In sum, we expect actors’ resolve to be most interdependent in real-world contexts that resemble prototypical contests of resolve, where actors are strategic and have centralized decision-making structures, and where resolve can be inferred with confidence. When interdependence is high, higher-order amplification of first-order reputational effects can produce especially large swings in the balance of resolve. It is in these circumstances, then, that beliefs about beliefs should be most consequential for crisis outcomes. These intuitive sources of variation in the interdependence of resolve could themselves be the object of empirical study, but for now, we leave this task for future work.

The Power of Beliefs

Each party is the prisoner or the beneficiary of their mutual expectations.

—Thomas Schelling (Schelling 1960, 60)

Reputations for resolve have long been the subject of debate: some see them as indispensable assets, while others dismiss past actions as irrelevant to current crises. Using an experimental approach, this article strongly reinforces the former view—states and leaders can form reputations for resolve and leverage them to their advantage during crises. Moreover, we emphasize that higher-order beliefs play an under-appreciated yet crucial role in this process, as first-order reputational effects are amplified by actors’ beliefs about their opponents’ beliefs. In this sense, international contests of resolve hinge not merely on past actions, but on actors’ combined expectations about the implications of past actions for present behavior.
In conclusion, we make several suggestions for further empirical research on higher-order beliefs and interdependent resolve in international politics. As mentioned earlier, one limitation of our study is that we lack direct knowledge of the level of higher-order reasoning at which respondents analyzed the crisis scenario. A number of studies in behavioral economics measures “k-level reasoning” directly in laboratory game settings (see Hafner-Burton, Hughes, and Victor 2013, for a review), but these measures are rarely found in survey research on international politics. Future studies could adapt such techniques to shorter survey formats, where they could serve as useful individual-level measures of strategic competence.

We also highlight several interesting avenues for future research. One promising idea is that opponents’ perceptions of a leader’s competence could influence their higher-order inferences about that leader’s likely beliefs, consequently shaping their own bargaining behavior. Some leaders are understood to be especially experienced, calculating, or wise, whereas others may be seen as inexperienced, impulsive, or even wholly incompetent. Not only could variation in leader competence directly affect a state’s behavior, but the perception of competence or ineptitude might also shape others’ higher-order expectations about that leader’s beliefs and world view, with implications for their own behavior during crises.

Moreover, there are also strong reasons to expect that beliefs about beliefs matter in contexts far beyond crisis bargaining, including collective action and coordination on many important international issues, such as climate change and international law (Mildenberger and Tingley 2019; Morrow 2014). Schelling’s argument that sets of beliefs can act as prisons (or paradises) matters as much to problems like gender inequality as it does to international conflict (Ridgeway 2011). Along these lines, the study of higher-order belief dynamics also presents an exciting opportunity for collaboration across research fields. Such dynamics can only be understood by combining the insights and methods of psychological, cultural, and strategic approaches—whether and how one actor forms and acts on beliefs about another actor’s beliefs depends on cognitive processes, systems of social meaning, and the anticipated consequences of different courses of action (O’Neill 1999). And once they are understood, insights on higher-order beliefs are likely to travel across many domains and issue areas.

In short, higher-order beliefs are an incredibly rich area for future studies of broad applicability and substantive importance. This article offered some theory and a novel methodological framework for diagnosing the effects of higher-order beliefs, which we hope will contribute to a vibrant sub-literature on this topic in international relations scholarship.

**Acknowledgments**

We would like to thank Stephen Walt for generously agreeing to advertise the survey, Tyler Bowen and Mojmir Stehlik for research assistance, and for help on this and related work, Sophia Hatz, Maria Hjelm, Susan Hyde, Josh Kertzer, James Morrow, Barry O’Neill, Jonathan Renshon, Todd Sechser, Steve Weber, Alex Weisiger, Keren Yarhi-Milo, the
participants of the Department of Peace and Conflict Research workshop at Uppsala University and the Division of Social Science seminar at Hong Kong University of Science & Technology, and especially Robert Trager.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Remco Zwetsloot https://orcid.org/0000-0002-2788-6257

Supplemental Material
Supplemental material for this article is available online.

Notes
1. For clarity, we use gendered pronouns to refer to A (female) and B (male) throughout the article.
2. McManus (2014, 728) briefly alludes to higher-order belief dynamics, but does not explore them empirically.
3. In real-world crises, the “game” is obviously more complicated. We collapse the policy space to “back down” or “stand firm” simply for ease of exposition.
4. For more on the meaning of resolve in international relations, see Fearon (1992, 66); O’Neill (1999, 107); Kertzer (2016, Ch. 1).
5. Our behavioral definition is not optimal for all research objectives. For instance, it is often productive to focus on the “value for war” because bargaining models can be solved with respect to this exogenous parameter. Such models tend to assume common knowledge (or common beliefs) about many aspects of the world (for exceptions, see Acharya and Grillo 2015; Chadefaux 2020; Smith and Stam 2004). Our broader definition, in contrast, is helpful for those seeking to understand “pre-equilibrium” dynamics in a world where common beliefs have not yet been established, as is often true when crises and wars happen.
6. A state’s reputation for resolve encompasses a number of more specific reputations—such as those for military competence and cost tolerance—concerning characteristics that affect its probability for standing firm. The extent to which behavior in one dispute affects how a state is expected to act in another dispute depends on a number of factors; for an in-depth discussion see Dafoe, Renshon, and Huth (2014).
7. Within conflict studies, past behavior has been found to affect dispute initiation (e.g. Weisiger and Yarhi-Milo 2015), coercive success (Cebul, Dafoe, and Monteiro Forthcoming), war onset (e.g. Crescenzi, Kathman, and Long 2007), and military alliances (e.
g. Mattes 2012; LeVeck and Narang 2017). Others examine the importance of reputation for cooperation; see Crescenzi (2017).

8. For other recent experimental work, which focuses on first-order beliefs, see Kertzer, Renshon, and Yarhi-Milo (2021), Kertzer (2016), and Renshon, Dafoe, and Huth (2018).

9. The relevant “perceiver” of a state’s resolve could be any other international observer, be it another state involved in a dispute with the state in question, or an uninvolved third party—in either case, agents should make similar reputational inferences. For the purposes of empirical testing, however, the “perceivers” are our survey respondents, who assessed the resolve of two abstract states in crisis (see Section “Research Design”).

10. Some might wonder how much respondents can infer from State A’s disputes with other, unknown states (that are not B). Additional background details would allow respondents to form more precise beliefs, yet respondents can still make general reputational inferences from A’s past behavior even without these details. It is reasonable for respondents to infer that, if A stood firm three times in recent crises, then A is probably more likely to stand firm in a subsequent dispute, relative to the respondents’ prior beliefs. Empirically, we demonstrate that our respondents can and do draw reputational inferences even when given relatively limited information. And as noted below, our results accord with Renshon, Dafoe, and Huth (2018), who find similar reputational effects across abstract and real-world designs. Regardless, future work on higher-order beliefs could (and ideally would) employ diverse research designs in order to test the sensitivity of this or any other particular design feature.

11. The vignette manipulates a large number of features, yielding a total of 8,640 cells. In this, it is similar to related survey techniques such as conjoint analysis (Hainmueller, Hopkins, and Yamamoto 2014). We discuss the advantages and possible disadvantages of factorial vignette experiments with many manipulations in Online Appendix B.3.

12. Similar questions were used in Renshon, Dafoe, and Huth (2018) and Kertzer, Renshon, and Yarhi-Milo (2021), though neither define resolve precisely and both discuss only first-order effects.

13. For recent examples, see Walt (2015) and Walt (2016).

14. For other experimental work that uses non-elite samples to test theories of bargaining, see Renshon, Lee, and Tingley (2017), Kertzer, Renshon, and Yarhi-Milo (2021), and Cebul, Dafoe, and Monteiro (Forthcoming).

15. Main results are shown graphically below—full regression tables can be found in Online Appendix C.3.

16. The minimum likelihood respondents could give for A standing firm is 10 percent (0 percent to 20 percent), the maximum 90 percent (80 percent to 100 percent). This effect is approximately equal to one standard deviation of the resolve variable (μ = 62 percent, σ = 2 percent), which is equivalent to a Cohen’s d = 1. Cohen offered d = 0.8 as a “big” effect.

17. The capabilities results, as well as results from all other manipulations, can be found in Online Appendix C.

18. We do not manipulate interests, but Kertzer, Renshon, and Yarhi-Milo (2021, 22) find similar results when doing so: even when told that one side has high stakes and high relative power, the effect of past behavior on perceptions of resolve is large and significant.
19. Recall from Section “Survey Vignette” that all treatments had a “no information” option, meaning some respondents randomly received no information about several features of the scenario.
20. See Online Appendix D for a formal proof.
21. We say “likely” because we do not observe respondents’ belief updating processes directly and thus cannot prove that the exclusion restriction holds (as is the case for all instrumental variable-type identification). For more on identifying assumptions in survey experiments, see Dafoe, Zhang, and Caughey (2018).
22. $p \approx 0.006$ and $p \approx 0.01$, respectively, with bootstrapped 95 percent CIs of $(-1.22, -0.22)$ and $(-1.32, -0.18)$.
23. Generally, when scholars say an actor has nth order beliefs, they hold beliefs about the beliefs of others up to the $(n - 1)$th order. Actors that have only first-order beliefs, for example, have no informative beliefs about the others’ beliefs (see Camerer, Ho, and Chong 2004, for a more extensive discussion). In our case, $n = \max(k)$.
24. As $\theta \rightarrow 1$, $IM \rightarrow \infty$, which means that even an $\epsilon$ change in perceptions decides the contest of resolve. Here, the prior state of affairs will have been something of a knife-edge equilibrium.
25. The above adopts the simplifying assumption that a linear function is a meaningful parameterization for the interdependence of resolve (where $\theta$ is the slope). Future work could explore a more general functional form mapping a change in A’s perception of B’s resolve, and A’s “prior” resolve, to the marginal change in A’s resolve: $\tilde{f}_A(\Delta \hat{R}_B^{k-1}, \hat{R}_B^{k-1}, \hat{R}_A^k)$. For example, it’s plausible that the impact of $\Delta \hat{R}_B^{k-1}$ has a sigmoidal shape like the logistic function, where the slope is greatest at intermediate values of $\hat{R}_B^{k-1}$ (where A is most likely to be indifferent between escalating and backing down) and $\hat{R}_A^k$ (where $\hat{R}_A^k$ has the farthest to move in either direction). This conjecture can be grounded in analysis of Chicken using Quantal Response Equilibrium, where a player’s best response function is smooth in the other’s strategy and in fact logistically shaped; see Goeree, Holt, and Palfrey 2016.
26. Many of the other crises Trachtenberg discusses are similarly good examples.
27. For instance, see Hof (2016) and Yacoubian (2017).
28. One example of (imperfect) real-world efforts to signal unconditional resolve comes from the Cold War, during which many governments considered introducing an element of automaticity in retaliation in times of crisis. Schelling (1966, 38-39) calls this “irrational automaticity,” exemplified by Soviet Premier Khrushchev’s 1959 threat to Averell Harriman that Soviet rockets would fly “automatically” should the West send tanks into Berlin.

References

Acharya, Avidit, and Edoardo Grillo. 2015. “War with Crazy Types.” Political Science Research and Methods 3 (2): 281-307.
Adida, Claire, Jessica Gottlieb, Eric Kramon, and Gwyneth McClendon. 2020. “When Does Information Influence Voters: The Joint Importance of Salience and Coordination.” Comparative Political Studies 53 (6): 851-91.
Bauer, Craig P. 2013. *Secret History: The Story of Cryptology*. New York: CRC Press.

Bicchieri, Cristina. 2016. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford, UK: Oxford University Press.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton, NJ: Princeton University Press.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong. 2004. “A Cognitive Hierarchy Model of Games.” *Quarterly Journal of Economics* 119 (3): 861-98.

Carter, David B. 2017. “History as a Double-edged Sword: Historical Boundaries and Territorial Claims.” *Politics, Philosophy & Economics* 16 (4): 400-21.

Cebul, Matthew, Allan Dafoe, and Nuno Monteiro. Forthcoming. “Coercion and the Credibility of Assurances.” *Journal of Politics*.

Chadeeaux, Thomas. 2020. “What the Enemy Knows: Common Knowledge and the Rationality of War.” *British Journal of Political Science* 50 (4): 1593-1607.

Chwe, Michael Suk-Young. 2001. *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton, NJ: Princeton University Press.

Crescenzi, Mark J. C. 2017. *Of Friends and Foes: Reputation and Learning in World Politics*. Oxford, UK: Oxford University Press.

Crescenzi, Mark J. C., Jacob D. Kathman, and S. B. Long. 2007. “Reputation, History, and War.” *Journal of Peace Research* 44 (6): 651-67.

Dafoe, Allan, Jonathan Renshon, and Paul K. Huth. 2014. “Reputation and Status as Motives for War.” *Annual Review of Political Science* 17:371-93.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. “Information Equivalence in Survey Experiments.” *Political Analysis* 26 (4): 399-416.

Fearon, James D. 1992. “Threats to Use Force: Costly Signals and Bargaining in International Crises.” PhD diss., University of California, Berkeley, CA.

Fearon, James D. 1994. “Domestic Political Audiences and the Escalation of International Disputes.” *American Political Science Review* 88 (3): 577-92.

Fearon, James D. 1995. “Rationalist Explanation for War.” *International Organization* 49 (3): 379-414.

Goeree, Jacob K., Charles A. Holt, and Thomas R. Palfrey. 2016. *Quantal Response Equilibria: A Stochastic Theory of Games*. Princeton, NJ: Princeton University Press.

Hafner-Burton, Emilie M., Alex Hughes, and David G. Victor. 2013. “The Cognitive Revolution and the Political Psychology of Elite Decision Making.” *Perspectives on Politics* 11 (2): 368-86.

Hafner-Burton, Emilie M., Brad L. LeVeck, David G. Victor, and James H. Fowler. 2014. “Decision Maker Preferences for International Legal Cooperation.” *International Organization* 68 (4): 845-76.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. “Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments.” *Political Analysis* 22 (1): 1-30.

Hof, Frederic C. 2016. *Syria: What Next?* Accessed March 2, 2021. https://www.atlanticcouncil.org/blogs/syriasource/syria-what-next/.

Huth, Paul K. 1997. “Reputations and Deterrence: A Theoretical and Empirical Assessment.” *Security Studies* 7 (1): 72-99.
Jervis, Robert. 1970. *The Logic of Images in International Relations*. Princeton, NJ: Princeton University Press.

Jervis, Robert. 1978. “Cooperation under the Security Dilemma.” *World Politics* 30 (2): 167-214.

Jervis, Robert. 1991. “Domino Beliefs and Strategic Behavior.” In *Dominoes and Bandwagons: Strategic Beliefs and Great Power Competition in the Eurasian Rimland*, edited by Robert Jervis and Jack Snyder, 20-50. Oxford, UK: Oxford University Press.

Kaplan, Fred. 1983. *The Wizards of Armageddon*. Stanford, CA: Stanford University Press.

Kertzer, Joshua D. 2016. *Resolve in International Politics*. Princeton, NJ: Princeton University Press.

Kertzer, Joshua D., Jonathan Renshon, and Keren Yarhi-Milo. 2021. “How Do Observers Assess Resolve?” *British Journal of Political Science* 51 (1): 308-30. doi: 10.1017/S0007123418000595.

Kuran, Timur. 1995. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge, MA: Harvard University Press.

LeVeck, Brad L., and Neil Narang. 2017. “How International Reputation Matters: Revisiting Alliance Violations in Context.” *International Interactions* 43 (5): 797-821.

Mattes, Michaela. 2012. “Reputation, Symmetry, and Alliance Design.” *International Organization* 66 (4): 679-707.

McManus, Roseanne W. 2014. “Fighting Words: The Effectiveness of Statements of Resolve in International Conflict.” *Journal of Peace Research* 51 (6): 726-40.

Mercer, Jonathan. 1996. *Reputation and International Politics*. Ithaca, NY: Cornell University Press.

Mercer, Jonathan. 2012. “Audience Costs Are Toys.” *Security Studies* 21 (3): 398-404.

Mildenberger, Matto, and Dustin H. Tingley. 2019. “Beliefs about Climate Beliefs: The Importance of Second-order Opinions for Climate Politics.” *British Journal of Political Science* 49 (4): 1279-1307. doi: 10.1017/S0007123417000321.

Morrow, James D. 1989. “Capabilities, Uncertainty, and Resolve: A Limited Information Model of Crisis Bargaining.” *American Journal of Political Science* 33 (4): 941-72.

Morrow, James D. 2014. *Order within Anarchy: The Laws of War as an International Institution*. Cambridge, MA: Cambridge University Press.

Mutz, Diana C. 2011. *Population-based Survey Experiments*. Princeton, NJ: Princeton University Press.

Nagel, Rosemarie. 1995. “Unraveling in Guessing Games: An Experimental Study.” *American Economic Review* 85 (5): 1313-26.

O’Neill, Barry. 1999. *Honor, Symbols, and War*. Ann Arbor, MI: University of Michigan Press.

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

Press, Daryl G. 2005. *Calculating Credibility: How Leaders Assess Military Threats*. Ithaca, NY: Cornell University Press.

Renshon, Jonathan, Allan Dafoe, and Paul K. Huth. 2018. “Leader Influence and Reputation Formation in World Politics.” *American Journal of Political Science* 62:325-39.
Renshon, Jonathan, Julia J. Lee, and Dustin Tingley. 2017. “Emotions and the Micro-foundations of Commitment Problems.” *International Organization* 71:S189-218. doi: 10.1017/S0020818316000473.

Ridgeway, Cecilia L. 2011. *Framed by Gender: How Gender Inequality Persists in the Modern World*. Oxford, UK: Oxford University Press.

Sartori, Anne E. 2016. “International Reputation with Changing Resolve.” Working Paper. Accessed March 2, 2021. https://asartori.mit.edu/sites/default/files/documents/Sartori%2ointernational%2oreputation%20Feb%202016%20.pdf.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schelling, Thomas C. 1966. *Arms and Influence*. New Haven, CT: Yale University Press.

Schlosser, Eric. 2014. *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*. New York: Penguin Books.

Slantchev, Branislav L. 2003. “The Principle of Convergence in Wartime Negotiations.” *American Political Science Review* 97 (4): 621-32.

Smith, Alistair, and Allan C. Stam. 2004. “Bargaining and the Nature of War.” *Journal of Conflict Resolution* 48 (6): 783-813.

Snyder, Glenn H., and Paul Diesing. 1977. *Conflict Among Nations: Bargaining, Decision Making, and System Structure in International Crises*. Princeton, NJ: Princeton University Press.

Tingley, Dustin H. 2017. “Rising Power on the Mind.” *International Organization* 71 (S): 165-88.

Tingley, Dustin H., and Barbara F. Walter. 2011. “The Effect of Repeated Play on Reputation Building: An Experimental Approach.” *International Organization* 65 (2): 343-65.

Trachter, Marc. 2012. “Audience Costs: An Historical Analysis.” *Security Studies* 21 (1): 3-42.

Trager, Robert F. 2017. *Diplomacy: Communication and the Origins of International Order*. Cambridge, UK: Cambridge University Press.

Trager, Robert F., and Lynn Vavreck. 2011. “The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party.” *American Journal of Political Science* 55 (3): 526-45.

Tsebelis, George. 2002. *Veto Players: How Political Institutions Work*. Princeton, NJ: Princeton University Press.

Walt, Stephen M. 2015. “The Credibility Addiction.” *Foreign Policy*. Accessed March 2, 2021. http://foreignpolicy.com/2015/01/06/the-credibility-addiction-us-iraq-afghanistan-unwinnable-war/.
Walt, Stephen M. 2016. “The Great Myth About U.S. Intervention in Syria.” *Foreign Policy*. Accessed March 2, 2021. http://foreignpolicy.com/2016/10/24/the-great-myth-about-u-s-intervention-in-syria-iraq-afghanistan-rwanda/.

Weisiger, Alex, and Keren Yarhi-Milo. 2015. “Revisiting Reputation: How Past Actions Matter in International Politics.” *International Organization* 69 (2): 473-95.

Wellman, Henry M. 2014. *Making Minds: How Theory of Mind Develops*. Oxford, UK: Oxford University Press.

Yacoubian, Mona. 2017. Critical Junctures in United States Policy Toward Syria.” Tech. rep. Washington, DC: United States Holocaust Memorial Museum.