Are Red Lines Red Herrings?

Abstract: If an older conventional wisdom in scholarly and policy-making circles held that reputation was "one of the few things worth fighting for," a more recent argument holds that past actions are relatively costless. Using archival evidence, reputation critics have argued that a country's credibility is rarely at stake, as foreign observers discount an actor's actions in the past when calculating her credibility in the present. We introduce a new type of evidence into this debate, presenting the results from original surveys fielded on four samples (from foreign decision-makers, foreign publics, and American IR scholars) to study the reputation costs incurred by various actors as a result of the Russian invasion of Crimea and the ongoing Syrian civil war. Our results suggest that reputations costs do indeed exist, that they are worse in cases of failed threats, and that they attach to both leaders and the countries they represent. Finally, our studies revealed the existence of a "home-away gap": compared to foreign observers, American IR scholars seem to underestimate the magnitude of reputation costs the US has incurred by backing down on threats.
American foreign policy has faced no shortage of debates over the past half century, but one of the most central concerns is the role of reputation. An older tradition, developed by early deterrence theorists, saw a reputation for resolve as “one of the few things worth fighting for,” and thus as one of the most critical objectives in American foreign policy. In a world where actors have incentives to misrepresent their resolve, a key indicator observers use to assess whether an actor will stand firm in a dispute in the present is how it behaved in crises in the past. According to this line of argument, the United States should therefore be willing to fight even in circumstances of limited material importance in order to affect others’ beliefs about America’s credibility and resolve. Fighting for face was thus seen as essential both for deterring your adversaries and reassuring your allies.

In many ways, this line of argument has congealed into a conventional wisdom in the foreign policy establishment. Politicians and pundits routinely rely on reputational arguments when seeking to sell military interventions to wary domestic audiences, and it is not at all unusual to hear claims that accommodation, compromise or “inaction will jeopardize America’s credibility.” At the same time, however, the past two decades have also witnessed the rise of a prominent dissenting tradition in International Relations (IR), which argues that foreign observers pay much less attention to past actions than the early deterrence theorists believed. Relying on a series of thoughtful case studies, these reputation cynics offer a variety of counterarguments indicating that foreign observers tend to not draw the same lessons from previous crises that policymakers alleged. It may be putting it too strongly to suggest that this reputation cynical argument constitutes a new “conventional wisdom” among scholars, but it is now sufficiently well-entrenched that when the New York Times recently published an article summarizing the state of IR scholarship on reputation

and credibility, it characterized arguments about the importance of reputation as “baseless” and as having “been extensively debunked.”

In this article, we argue for a middle ground between these two positions. Given the tendency for hawkish voices in Washington to use reputational logics to justify an endless supply of military misadventures, the reputation critics are correct to be cynical about arguments that treat credibility as the most important consideration shaping American foreign policy. Observers assessing resolve will rely on a large number of indicators to do so; often times these indicators point in competing directions, and the effect of past actions can be overridden by other considerations. Yet, even if signals rarely speak for themselves, and messages routinely get lost in translation, it is incorrect to assert that past actions send no messages whatsoever: actors pay reputation costs in the eyes of international audiences when they fail to demonstrate resolve on the world stage. As James Fearon notes, “credibility is not everything, but it’s not nothing either.”

We make this claim while offering two contributions, the first conceptual and the second empirical. Conceptually, we decompose reputational inferences into two phases: the evaluation phase (do observers draw inferences about reputation in response to an actor’s past actions in a manner that classical theories of reputation assume?), and the diagnostic phase (how much weight do observers attach to reputational inferences when planning present behavior?). One reason why there has been so much disagreement about reputation in IR is because different strands of scholarship have been asking subtly different questions, from broad-stroke questions such as whether reputations “matter,” to investigations that largely focus on the diagnostic phase and are thus amenable to empirical study using observational data on behavior in crises. We focus our efforts on the evaluations phase using surveys that directly elicit beliefs in order to examine the foundational question of whether actors draw inferences about reputation in the manner predicted by our theories.

Empirically, we introduce a new type of evidence into these reputation debates. Previous scholarship on reputation has typically taken one of three forms. Much of the earlier deterrence literature was largely deductive and theoretical, seeking less to demonstrate empirically that reputation mattered than to demonstrate logically the conditions in which this would be the case. More recent work has often been divided between large-N quantitative studies — which have tended to find...
support for the effects of past actions, but bracketed the perceptual quality of calculations of credibility — and in-depth case studies by reputation cynics, which capture perceptions of credibility in particular cases, but raise questions about generalizability and selection effects. Yet, if reputations are ultimately beliefs about the tendencies of other actors, it is worthwhile to broaden our evidentiary toolkit by studying them using some of the same techniques political scientists routinely use to study beliefs more generally: surveys.

We explore the evaluation phase of reputation using four surveys fielded in 2015-16, innovating in two key respects. First, rather than ask respondents about fictional or hypothetical events, we focus on two highly salient, real-world cases: the conflict in Crimea, where Russian forces “reconnected” with the (formerly Ukrainian) territory in late 2014 over the objections of the United States, and the conflict in Syria, where the Obama administration repeatedly declared the use of chemical weapons by the Assad regime to be a “red line” that would bring the United States into the conflict, only to subsequently back down. Since many of our models of reputation would predict that this type of behavior is unlikely (if reputation costs are real, strategic leaders should be less likely to put themselves in situations that would incur them), these provide important cases in which to study reputational inferences in the wild, without asking respondents to think through unrealistic or unwieldy hypotheticals. Second, we field our surveys on a sample of elite decision-makers — current and former members of the Israeli Knesset — alongside a sample of IR scholars in the United States, and two nationally representative samples of the Israeli public. By fielding our surveys simultaneously along three very different types of samples, we can compare potential gaps between how American IR scholars think about reputation, with how foreign leaders and publics do.

We find that both President Obama and the United States suffered substantial reputation costs in both crises, while Russia’s reputation appears to have been enhanced following its actions in Syria. We also find that the reputational damage for the United States was more stark in the case of Syria than Crimea — consistent with the predictions of audience cost theory, which argues that backing down on public threats produces particularly negative consequences. In contrast to traditional audience cost models, however, our results focus our attention on the international ramifi-


cations of failed threats rather than their electoral consequences. Finally, we find that IR scholars at home perceive significantly smaller reputation costs when compared to observers abroad. Since reputations are ultimately beliefs about you in the eyes of others, our findings thus suggest that at least in some circumstances, IR scholars may perhaps be underestimating reputational inferences.

The Rise and Fall of Credibility and Reputations in IR

Does reputation for resolve matter in international relations? A “first wave” of scholars, led by Thomas Schelling, coalesced around an affirmative answer to this question. Schelling saw reputations as critical to IR, famously arguing that the United States had sacrificed thirty-thousand dead in the Korean War, but that doing so was “undoubtedly worth it” since “Soviet expectations about the behavior of the United States are one of the most valuable assets we possess in world affairs.” He made the case that commitments in world politics were interdependent (both across time and space) so that if a country consistently kept its promises and carried out its threats, its commitments would be perceived as more credible and its ability to deter would be enhanced (“a potent means of commitment, and sometimes the only means, is the pledge of one’s reputation…”).

Conversely, a failure to follow through on threats or keep promises would reduce the credibility of a country’s commitments in future interactions. Like much of his other work, Schelling’s clear, direct and powerful insights gained widespread acceptance within IR, setting a research agenda for several decades following his original work.

It was not until decades later that cracks began to emerge in this theoretical facade. As the Cold War drew to a close, a group of “reputation critics” began to challenge the claim that reputations for resolve critically affected international politics and interstate interactions, or even existed at all. In a series of papers, Ted Hopf showed that, as intuitively compelling as Schelling’s model of deterrence and reputation was, there was little evidence of it at work in the Cold War: “even when the United States engaged in absolutely no behavior designed to bolster its deterrent credibility, it still gained a reputation for resolve…” Conversely, no matter what they did, the United States seemed unable to lose that reputation:

if the United States did not lose its reputation for defending Europe, Northeast Asia and the Middle East after losing in a country of as much strategic importance as Iran, then where else in the Third World could such a loss of credibility occur?\textsuperscript{16}

This work, along with the research that followed, challenged key assumptions of Schelling’s, such as the notion that commitments were interdependent, reputations could form and persist, and so on.\textsuperscript{17} Jonathan Mercer, for example, argued that leaders attribute decisions to back down to situational characteristics of the crisis rather than to any permanent or dispositional feature of the opponent, preventing reputations for resolve from forming between adversaries.\textsuperscript{18} Perhaps the strongest critique of reputations came from the work of Daryl Press. Using case studies of European politics in interwar period and Soviet-American relations during the Cold War, Press found no evidence that either Nazi leaders or American policymakers made predictions about the likely behavior of their opponents based on their record of backing down in previous crises. Press argued that traditional “realist” variables — power and interest — explained crisis behavior and outcomes far more convincingly than expectations derived from classic theories of reputation.\textsuperscript{19} Shiping Tang gives us the strongest statement of the “reputation critics”: there is a “cult” of reputation among political leaders, who expend resources to chase something that doesn’t exist in the first place, since the anarchic nature of the international system prevents reputations from forming in the first place.\textsuperscript{20}

By the late 1990s, the second wave had already succeeded in throwing much of what we thought we knew about reputations into question. In an oft-cited review of the literature from that period, Paul Huth notes the “substantial gap between the intuitive belief that reputations are an important cause of international conflict and the development of a compelling logical argument and empirical evidence to support such a conclusion.”\textsuperscript{21} In other words: by 1997, we knew very little with any degree of certainty. This understanding of reputation — as a foolish illusion chased by policymakers — has by and large won the day. It is now commonplace to see op-eds and policy-oriented articles dismissing concerns about reputation as “baseless.” This conception of reputation

\textsuperscript{17} Hopf put it more strongly: “…deterrence theory assumptions are prima facie empirically invalidated by the evidence of Soviet lessons from these [events].” See ibid., 117.
\textsuperscript{18} Mercer, Reputation and international politics.
\textsuperscript{19} Press, Calculating credibility: How leaders assess military threats.
\textsuperscript{20} Tang, “Reputation, Cult of Reputation, and International Conflict.”
has reached the highest levels: Barack Obama, in discussing notions of credibility and reputation (and the interdependence of commitments), argued: “Look, this theory is so easily disposed of that I’m always puzzled by how people make the argument.” Jeffrey Goldberg, who conducted that interview, summarized the President’s views by noting that Obama “generally believes that the Washington foreign-policy establishment, which he secretly disdains, makes a fetish of ‘credibility.’” This was summarized more pithily by another commentator writing for Vox, who declared: “there’s at least one person who thinks the concept of credibility is total bullshit: President Barack Obama.”

While it seems clear that the reputation “cynics” have won the day among a certain popular understanding of world politics, they have spurred pushback within political science and international affairs research. One recent strand of empirical work on reputations has pushed back against the claim that reputations do not matter in international politics. Alex Weisiger and Keren Yarhi-Milo, using large-N observational data, have shown the consequences of reputations: countries that backed down in previous disputes are significantly more likely to be challenged again, while states that stood firm are less likely to be challenged in the future. Applying the reputation logic to protracted conflicts, Joe Clare and Vesna Danilovic find that, at least among the set of countries involved in enduring rivalries, countries that backed down in a past crisis are more likely to initiate a subsequent dispute. In a recent study of coercive diplomacy between the United States and smaller powers after the Cold War, Frank Harvey and John Milton find that reputation for resolve played an important role in how US adversaries assessed the credibility of American threats. Finally, using a conjoint experimental design, Joshua Kertzer, Jonathan Renshon, and Keren Yarhi-Milo provide evidence on how observers weight the myriad factors might affect judgments of resolve, finding that, while capabilities and stakes do affect calculations of credibility, so too do past actions, and that there are no levels of the former where the latter ceases to matter.

Much as first-generation work on reputations spurred innovations in deterrence theory, recent

27. Kertzer, Renshon, and Yarhi-Milo, “How Do Observers Assess Resolve?”
scholarship has branched off into a vibrant strain of work on audience costs. Reputations are central to this body of theory, since this type of argument assumes that reputations for resolve exist, that publics care about maintaining them, and that irresolute behavior harms them. While national publics seem to respond broadly in line with audience cost theory, recent work has extended and added nuance to these predictions. For example, the public seems to disapprove of leaders who back down not only because they see the inconsistency between words and deeds as damaging the country’s reputation for resolve, but also because they disapprove of the initial belligerence involved in making the threat in the first place. In another paper, Ryan Brutger and Joshua Kertzer show that the public is not of one mind in how it perceives reputational costs: while hawks are concerned about the negative reputational consequences of inconsistency, doves are equally concerned with the negative reputational consequences of belligerence and interventionism. Another strand of literature suggests that backing into a war a leader promised not to fight also invites disapproval from the public, mainly because it damages the credibility of the country’s future promises. And, outside of the narrower question of reputation for resolve, a large body of literature shows that reputations matter in IR across multiple domains.

Two features of this third wave of reputation research bear mentioning. One is a renewed focus


29. Kertzer and Brutger, “Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory.”


on methodological challenges that may have slowed the search for evidence on the importance of reputations. For example, the lack of evidence of decision-makers referring to past actions during crises might also stem from the simple fact that lessons from past actions are common knowledge and as such are less likely to appear in the historical record: in the Ex-Comm meetings of the Cuban Missile Crisis, nobody needed to be told what had happened at Munich in 1938.\(^{33}\) Additionally, references to past actions might be absent because of strategic selection, which lead observed adverse reputational effects to be biased towards zero.\(^{34}\) Finally, and echoing a growing movement within political science, we have seen a call for both more experimental work (to aid in causal inference) and better and more direct measurement of the beliefs of leaders.\(^{35}\)

A second feature of this third wave of literature on reputations is the disaggregating of reputations. This occurred as a direct response to the challenge of Robert Jervis to answer the question, “If one president acts boldly, will other states’ leaders draw inferences only about him or will they expect his successors to display similar resolve?”\(^{36}\) As it happens, the vast majority of work on reputations (69%, according to one recent review essay) has assumed that reputations are “owned” by countries.\(^{37}\) As more of IR has incorporated elites and leaders into their theories, reputation research have followed along, with some recent work emphasizing “leader-specific” reputations.\(^{38}\) Evidence has appeared on both side of the debate. For example, Alex Weisiger and Keren Yarhi-Milo show (using observational data) that leadership turnover does not produce a clean reputational slate, implying that countries do in fact “own” some portion of their reputation for resolve.\(^{39}\)

Danielle Lupton finds that leaders can develop a reputation for resolve over time, and that early

---

33. Dafoe, Renshon, and Huth, “Reputation and Status as Motives for War.”
35. Ibid., 389.
37. Dafoe, Renshon, and Huth, “Reputation and Status as Motives for War.”
perceptions of a leader’s resolve significantly influence later perceptions. They show that this process is heavily dependent on the influence the actor is perceived to have: when leaders are perceived to have significant influence over a policy, they get more of the credit or blame for whatever reputational consequences ensue.

Of course, there are exceptions to the trends we have highlighted. Recent research on reputations, for example — our “third wave” — are not universally bullish on reputations. Some recent work has argued against some of the foundations of reputational theories of world politics, and in so doing, fits better with earlier “reputation critics.” However, in general we see this as an accurate depiction of the landscape of research on reputation. We see the third-wave of reputation research as promising, not least because it diverges from previous generations of work by triangulating reputation from multiple methodological vantage points, including the use of both observational (qualitative and quantitative work) and experimental (including elite experiments) data.

While the third-wave of reputation research has helped push back against some of the more extreme of the reputation critics, there are three fundamental questions that have remained unanswered. First, do “reputational costs” exist in the first place? What kind of situations generate these costs? Second, who pays these costs; that is, which actors should we focus our collective attention on, leaders or states? Finally, how are these costs distributed? Audience cost research, for example, has mostly focused on the domestic consequences of backing down, whereas an earlier theoretical literature emphasized the international ramifications. We assess each in our study.

Decomposing reputational inferences: the diagnostic and evaluation phases

Some of the dissensus in the reputation literature can be chalked up to methodological differences. Much as scholars have noted a methodological divide in the human rights literature — in which the quantitative literature is generally much more pessimistic about the efficacy of human rights


treaties than the qualitative literature—a similar methodological divide existed in the second wave of reputation literature, with the quantitative and formal scholarship generally more convinced about the impact of reputations than the qualitative scholarship.\textsuperscript{43}

However, that methodological divide cannot explain everything, and, of course, exceptions to those tendencies exist on both sides. We argue that a deeper cause of the disagreement about reputations in IR stems from IR scholars working on reputation asking subtly different questions. For international reputation costs to “matter” in IR, two conditions have to be met. First, observers must change beliefs about an actor in response to that actor’s behavior; in the discussion below, we refer to this as the evaluation phase. Second, observers must then draw upon reputational considerations when formulating or analyzing present behavior; we refer to this as the diagnostic phase, since it focuses on the question of how much diagnostic weight in an analysis is placed on reputation compared to other variables like the intrinsic interests at stake.\textsuperscript{44} Different reputation critiques in the second wave of reputation literature emphasize different phases of this causal chain. Tang largely focuses on the evaluation phase, questioning whether actors draw reputational inferences the way our theories say they should. In his line of argument, the need for actors to assume the worst means that “reputation becomes impossible to develop under anarchy... A state is assigned its baseline image by its adversaries and allies at the beginning of a crisis, and no past behavior can change that image ex ante.”\textsuperscript{45} Press, on the other hand, largely focuses on challenging the diagnostic phase, arguing that “when assessing credibility during crises, leaders focus on the ‘here and


\textsuperscript{44} A similar distinction between the evaluation and diagnostic phase in reputational inference is made in the interdisciplinary literature on indirect reciprocity, which distinguishes between social norms (how should observers update reputations in response to past actions?), and action rules (how should we behave as a result of this reputation? – usually couched in the form of punishment decisions). See Hannelore Brandt and Karl Sigmund, “The logic of reprobation: assessment and action rules for indirect reciprocation,” \textit{Journal of Theoretical Biology} 231, no. 4 (2004): 475–486; Hisashi Ohtsuki and Yoh Iwasa, “The leading eight: Social norms that can maintain cooperation by indirect reciprocity,” \textit{Journal of Theoretical Biology} 239, no. 4 (2006): 435–444; David G. Rand and Martin A. Nowak, “Human cooperation,” \textit{Trends in Cognitive Sciences} 17, no. 8 (2013): 413–425. See also Sartori, \textit{Deterrence by Diplomacy}, 65-66 on situational reputation. At the same time, it is not simply that the evaluation phase is about beliefs, and the diagnostic phase about behavior — for example, credibility is a belief, yet calculating credibility is inherently about the diagnostic phase rather than the evaluation phase. On decision weights more generally, see Daniel Kahneman and Amos Tversky, \textit{Choices, Values and Frames} (Cambridge: Cambridge University Press, 2000). Techniques like conjoint analysis were developed precisely to study the use of decision weights; for an application to the study of reputation, see Kertzer, Renshon, and Yarhi-Milo, “How Do Observers Assess Resolve?”

\textsuperscript{45} Tang, “Reputation, Cult of Reputation, and International Conflict,” 50.
now,’ not on their adversary’s past behavior.”

Other critics target both phases. Hopf and Mercer, for instance, share some of Press’s interest in the diagnostic question of how much actors rely on past actions in one context when calculating credibility in another, but both are also interested in questioning the evaluation phase as well. Hopf shows that the evaluation phase functions “through irrational processes,” in a very different manner than deterrence theory suggests: he finds that, for example, America’s reputation for resolve improved even when the US “engaged in absolutely no behavior designed to bolster its deterrent credibility.” Similarly, Mercer offers an attribution-based theory of reputation that explicitly targets the evaluation phase of reputational models, arguing that “it is the desirability of the target’s immediate behavior rather than its past behavior” that determines whether we attribute it to dispositional factors — and thus, whether reputations can form, such that “there is not a simple correlation between behavior and reputation.”

In this paper, we tackle this evaluation phase head-on, testing how the reputations of different types of actors are seen to have shifted in response to events in two of the defining foreign policy crises of the Obama administration: the Russian annexation of Crimea, and the humanitarian crisis in the Syrian Civil War. We focus on this evaluation phase for two reasons. First, it is the phase that has perhaps been the least directly tested in the third wave of literature on reputation. Much of the large-N literature on reputation, because it cannot directly study observers’ beliefs, focuses instead on observable implications like crisis initiation — rendering it a better test of the diagnostic phase (do past actions shape future behavior?) than the evaluation phase (do reputations form or change in the way our theories say they should?). And, as we show below, although existing survey data can partially speak to the diagnostic phase, it is ill-equipped to addressing the evaluation phase.

Second, in some ways, it perhaps offers the most daunting challenge to our theories of reputation. Although it would be irrational for observers to assume that past actions are perfectly predictive of future behavior given the extent to which situations often differ from one another in myriad ways, it would be similarly irrational for observers to assume that past actions have no diagnostic value whatsoever. Every other discipline in the social (and even some of the natural) sciences — from economics, to sociology, to evolutionary biology — emphasizes the extent to which actors un-

48. Mercer, Reputation and international politics, 10,36.
nder conditions of uncertainty often take information about past actions into account, which is why a vast literature in psychology shows that actors behave differently when they believe they are being observed.\textsuperscript{50} Children as young as five have been shown to engage in reputation management, and even infants have been found to evaluate observers based on their past actions.\textsuperscript{51} This is not to say that reputational considerations exert an unconditional effect, that they should always trump other indicators in IR, or that reputations are always worth fighting for — merely that it would be surprising if reputational diagnoses exerted effects in every other domain of human life apart from international politics.

In contrast, the evaluation phase — the process through which reputational assessments are made — is a more difficult test. Even outside of the reputational literature in IR reviewed above indicating that observers often do not update their reputational judgments in the manner in which our theories of reputation suggest, a broader body of work on judgment and decision-making has shown that individuals often update beliefs in manners deviating from canonical models.\textsuperscript{52} Learning, for example, is more difficult than many classic models in IR assume, as observers are often “prisoners of their preconceptions.”\textsuperscript{53} The images that we have of other actors are often sticky,

Crimea and Syria

Our interest here is studying reputational inferences drawn during two of the defining foreign policy crises of the Obama administration: the Russian annexation of Crimea in the spring of 2014, and the humanitarian crisis in Syria, particularly in the aftermath of Obama’s declaration that the use of chemical weapons by the Assad regime would constitute a “red line” that would lead to intervention by the international community. As an original automated content analysis of over 87,000 newspaper articles on the two crises in Figure 2 shows, reputational rhetoric loomed large in each crisis, though perhaps somewhat larger in the case of Syria.

Figure 1 plots the main developments in each crisis, though the broad outlines are by now well-known.\footnote{Timeline of events relating to Syria comes from the following sources: BBC News Syria Profile and POLITICO Timeline: U.S. Approach to the Syrian Civil War. Information on the Crimean conflict comes from the following sources: BBC News Ukraine Crisis Timeline and CNN Ukraine: Everything You Need to Know About How We Got Here.} In the Syrian crisis, Barack Obama issued his famous “red line” threat in August of 2012 following the tightening of U.S./EU sanctions and concerns that Assad might use non-traditional weapons to suppress the uprising:
Syria

- **March 2011**: Uprising begins; Assad announces conciliatory measures (e.g., lifts 48-year old state of emergency).
- **May 2011**: Tanks enter suburbs of Damascus in an effort to crush anti-regime protests. US and EU tighten sanctions.
- **August 2011**: President Barack Obama (joined by Germany, France and UK) formally calls on Assad to step down “for the sake of the Syrian people” and issues new sanctions on his regime.
- **June 2011**: IAEA reports Syria to UN Security Council over covert nuclear program
- **August 20, 2012**: In remarks to reporters at the White House, Obama warns Assad’s regime not to use chemical weapons, describing it as a “red line” for his administration that could prompt the United States to intervene militarily.
- **August 24, 2012**: The USS John C. Stennis and its strike group prepares to deploy to the Middle East.
- **August 29, 2012**: Senator John McCain says Obama’s inaction in Syria has led to a “savage and unfair fight” by the regime of Bashar Assad.
- **August 31, 2012**: A leading member of the Free Syrian Army says, contrary to media reports, they have received no operational support from the United States or any other state.
- **August 21, 2013**: Hundreds of people suffocate in rebel-held suburbs of the Syrian capital. UN investigators visit the sites and determine that ground-to-ground missiles loaded with sarin were fired on civilian areas while residents slept. The US and others blame the Syrian government, the only party known to have sarin gas.
- **September 27, 2013**: UN Security Council orders Syria to account for and destroy its chemical weapons stockpile; threatens to authorize the use of force in the event of non-compliance.
- **January-February 2014**: UN-brokered peace talks in Geneva fail, largely because Syrian authorities refuse to discuss a transitional government.
- **March 2014**: Syrian Army/Hezbollah forces recapture Yabroud, the last rebel stronghold near the Lebanese border.
- **September 2014**: US and five Arab countries launch air strikes against Islamic State around Aleppo and Raqqa.

Crimea

- **November 21, 2013**: President Yanukovych’s cabinet abandons an agreement on closer trade ties with EU, instead seeking closer co-operation with Russia. Small protests start.
- **Early December 2013**: Protesters occupy Kiev city hall and Independence Square in dramatic style. 800,000 people rally in Kiev.
- **December 17, 2013**: Vladimir Putin throws President Yanukovych an economic lifeline, agreeing to buy $15bn of Ukrainian debt and reduce the price of Russian gas supplies by about a third.
- **February 20, 2014**: Kiev sees its worst day of violence for almost 70 years. At least 88 people are killed in 48 hours. Video shows uniformed snipers firing at protesters holding makeshift shields.
- **February 21, 2014**: President Yanukovych signs compromise deal with opposition leaders.
- **February 22, 2014**: President Yanukovych disappears. Protesters take control of presidential administration buildings. Parliament votes to remove president from power and sets date for elections. Mr Yanukovych appears on TV to denounce “coup.” His arch-rival Yulia Tymoshenko is freed from jail.
- **March 1, 2014**: Russia’s parliament approves President Vladimir Putin’s request to use force in Ukraine to protect Russian interests.
- **March 16, 2014**: Crimea’s secession referendum on joining Russia is backed by 97% of voters. EU and US impose travel bans and asset freezes on officials from Russia and Ukraine in response.
- **March 18, 2014**: President Putin signs a bill to absorb Crimea into the Russian Federation.
- **March 28, 2014**: US President Barack Obama urges Moscow to “move back its troops” and lower tensions.
- **July 30, 2014**: The EU and US announce new sanctions against Russia.
- **September/October, 2014**: NATO reports a “significant” withdrawal of Russian troops from eastern Ukraine. Putin orders thousands of troops stationed near the Ukrainian border to return to their bases. Russia agrees to resume gas supplies to Ukraine over the winter in a deal brokered by the EU.
- **February 15, 2015**: Ceasefire goes into effect (but violations quickly follow).

Figure 1: Crisis Timelines
We have been very clear to the Assad regime, but also to other players on the ground, that a red line for us is we start seeing a whole bunch of chemical weapons moving around or being utilized. That would change my calculus.

Despite the attention paid to the so-called “red-line” statement (and there was quite a bit of attention focused on it), the U.S. ultimately did not intervene significantly in Syria, Assad remained in power and the uprising erupted into a full-blown civil war that has left an estimated 400,000 dead at the time of this writing. In stark contrast, the Crimea timeline shows a very different type of conflict. In this case, Obama “urged” Putin to withdraw troops and enacted sanctions (along with the EU, at times), but avoided the stark language of threats he had utilized in the Syrian crisis. In many ways, however, the outcome was similar from the U.S. perspective: a failure to accomplish its policy goals and a looming concern that it would pay a price for its inaction in the coin of credibility.

Methods and Data

We seek to bring a new type of evidence to bear in these debates. The analyses and lessons described in this paper come from a series of surveys carried out over the last several years, though the data presented here is done so for the first time. This type of method is especially valuable for four reasons. First, if we reputations are defined as beliefs about the traits of particular actors, survey methods are particularly helpful at measuring the content of beliefs. They are particularly better-suited in this regard than the large-N literature on reputation, which must rely on rough behavioral proxies in lieu of capturing the content of the beliefs themselves. Second, unlike much of the survey-based research in IR, our sample here includes foreign leaders: current and former members of the Israeli Knesset. In this sense, if the central focus of reputation-based arguments in US foreign policy is the reputational costs accrued in the eyes of allies and adversaries, the suite of


Figure 2 estimates the proportion of reputational language in daily newspaper coverage of the crisis in Crimea (from January 15-July 15, 2014) and Syria (from July 1-December 31, 2012), superimposed by a loess smoother (in blue). On average, reputation, resolve, honor, and credibility are discussed in 9.7% of all articles about Crimea, and 10.5% of all articles about Syria during these six-month windows. For Crimea, the denominator here consists of 49,248 news articles available on LexisNexis (from January 15-July 15, 2014) mentioning Crimea||Ukraine)&&(crisis|war|uprising|revolt|conflict|violence|troops|rebellion|rebels|annexation); for Syria, the denominator consists of 37,937 news articles available on LexisNexis (from July 1-December 31, 2012) mentioning Syria&&(crisis|war|uprising|revolt|conflict|violence|violent|troops|rebellion|rebels|revolution|terror). In both cases, the numerator retains articles containing the words reputation|resolve|honor|honour|credible|credibility|credibly, omitting references to phrases found to be unrelated to the quantities of interest (resolved, resolves, to resolve, honorable, honourable, an honor, an honour, honoring, honouring, honorary, honourary, honorific, honourific, incredible) after manually inspecting a random subset of the search results.
tools we employ here enables us to study the question directly. Third, and relatedly, although there has been a barrage of survey or experimental work studying audience costs in the past decade, it tends to study audience costs from the perspective of the domestic audience, rather than the international one, even though the formal model whose microfoundations the experiments are testing is dyadic.\textsuperscript{60} Indeed, writing nearly half a century ago, Robert Jervis warned that we shouldn’t assume that policymakers are concerned about reputation just for domestic political reasons.\textsuperscript{61} Understanding the international dimensions of reputation costs using this method is thus particularly valuable.

Figure 3: Diagnostic polling questions, 2000-2015

We were unable to find any existing survey data capable of testing the evaluation phase of reputational models; there are some survey questions that speak to the diagnostic phase — but they aren’t regularly fielded until 2015, following the crises in Crimea and Syria. Nonetheless, the fact that polling firms do not think to ask allied publics whether the US would come to their defense until after these two crises is itself suggestive.

Finally, we note that although surveys are well-suited to studying the perceptual component of reputation, there are remarkably few existing surveys capable of addressing the reputational questions we are interested in. There are no international polls we are aware of that capture the

\textsuperscript{60}Tomz, ”Domestic audience costs in international relations: An experimental approach”; Trager and Vavreck, ”The political costs of crisis bargaining: Presidential rhetoric and the role of party”; Levendusky and Horowitz, ”When Backing Down is the Right Decision: Partisanship, New Information, and Audience Costs”; Davies and Johns, ”Audience costs among the British public”; Chaudoin, ”Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions”; Kertzer and Brutger, ”Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory”; Levy et al., ”Backing Out or Backing In? Commitment and Consistency in Audience Costs Theory”; Kai Quek, ”Type II Audience Costs,” Journal of Politics 79, no. 4 (2017): 1438–1443.

evaluation phase, which would involve asking respondents to assess how the reputation of various actors have been affected by particular events. As Figure 3 shows, Pew, Gallup and several local polling firms have fielded polls in foreign countries in recent years that potentially speak to the diagnostic phase. However, two points are worth noting here. First, the questions do not explicitly ask respondents about the link between an actor’s past actions and future behavior, but rather simply ask about future behavior, such as whether respondents believe the United States would defend their country if attacked.\footnote{62} Given this limitation, one could still use these questions to get at the diagnostic phase if they were fielded repeatedly through time, such as before and after an international crisis. However, one of the striking patterns in Figure 3 is that these questions only begin to be asked with any regularity in 2015 — after the crises in Crimea and Syria we describe here. As Krebs notes, the questions polling firms choose to ask — and when — can itself be informative.\footnote{63}

The fact that no polling firm in the past two decades thought to ask allied publics whether they thought the United States would be willing to defend them until after the crises in Crimea and Syria suggests the matter was not seen as being in question until then.\footnote{64} This underlies the value of these two contemporary cases for the study of reputation.

For this study, we fielded four surveys on three different populations: two samples of elites, and one mass public (Table 1 provides an overview of our data sources, which we describe below; more detail can be found in Appendix §2). “Elite” as it is used to describe survey or experimental samples in international relations typically refers to one of two different types of samples. The first are samples of subject-matter experts, as in Philip Tetlock’s studies of expert political judgment, or a series of surveys fielded by Cheryl Koopman and collaborators on the readership of International

\footnote{62} For example, in 2017, Pew asked respondents in a number of NATO member countries whether they thought the US would use “military force to defend their country if they got in a serious conflict with Russia.”


\footnote{64} Importantly, this gap is not due to an absence of questions about US foreign policy: we are able to find polls in foreign publics about US foreign policy fielded as early as 1991; as Figure 3 shows, however, questions about US credibility only begin in 2015, despite events in US foreign policy in the intervening time period where US credibility might be perceived as being in question, such as the US withdrawal from Somalia, the war in Kosovo, or events in Iraq before the surge.
Security. The second are samples of government officials or decision-makers. We look at both here, but with a crossnational twist.

Our first elite sample consists of foreign decision-makers: current and former members of the Israeli Knesset. Israel is an important case in which to study decision-making in international security. Despite the absence of formal alliance commitments with one another, the US and Israel have a remarkably close informal alliance, reflected not just in the outsized volume of military aid the former gives to the latter, but in the extent to which both proponents and critics of US foreign policy in the Middle East see the two countries as bound together by a “special relationship” defined by shared perceptions of geopolitical threats in the region. Although each of these relationships differ in a variety of important ways, Israeli decision-makers thus represent a more militarily powerful variant of a particular class of foreign leaders, occupying a similar structural position as Japan and South Korea in East Asia, Poland and Estonia in Eastern Europe, and so on: smaller powers dependent upon American backing in order to provide for their security, and thus the very type of ally that deterrence theory tells us should be invested in the maintenance of American credibility.

Table 1: Data Sources

<table>
<thead>
<tr>
<th></th>
<th>Fielded:</th>
<th>Subject pool</th>
<th>location</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign leaders</td>
<td>July–October, 2015</td>
<td>Members of the Israeli Knesset</td>
<td>Israel</td>
<td>89</td>
</tr>
<tr>
<td>Foreign public I</td>
<td>October 6-9, 2015</td>
<td>Representative sample of Israeli public</td>
<td>Israel</td>
<td>1599</td>
</tr>
<tr>
<td>Foreign public II</td>
<td>January 18-25, 2016</td>
<td>Representative sample of Israeli public</td>
<td>Israel</td>
<td>1111</td>
</tr>
<tr>
<td>American IR scholars</td>
<td>September 28-30, 2015</td>
<td>TRIP survey</td>
<td>United States</td>
<td>692</td>
</tr>
</tbody>
</table>

The Knesset study was fielded the study from July - October 2015. We began with a “universe” of Knesset members that included anybody that had served since 1996, including members of the


14th - 20th Knessets, a total of 408 individuals. After gaining the contact information for the current and former members, we were left with 288 potential candidates (all 120 current members plus 168 former members for whom we could verify contact information). Of those 288 potential subjects, 89 participated, leaving us with a 31% response rate, relatively high for surveys of this type. The Knesset sample is described in Table 2 and our recruitment procedures are described in Appendix §2. 25% of our participants were currently serving in the Knesset; the rest of the sample (75%) was composed of former Knesset members. They were also highly experienced in IR-relevant contexts: 64% had active combat experience, and 67% had experience serving as members of the Knesset’s Foreign Affairs and Defense committee. In addition to experience along the dimension of military strategy, they also had considerable political experience: our participants served an average of 3 terms in Parliament (with some as high as 9 terms). While 61% of the Knesset subjects had never served as a Minister, 29% had been at least a Deputy Minister, and fully 12% of our sample was in our highest category of elite experience, such that our participants include individuals who had served as Cabinet members, and even Prime Minister.68

Our second elite sample is composed of subject matter experts: American IR scholars, recruited by the Teaching, Research and International Policy (TRIP) survey. This sample is “elite” in the sense of the term used by previous elite surveys by Koopman et al and Tetlock: respondents who, through training, experience or both, can be accurately described as subject-matter experts.69 Our data from the TRIP survey was fielded by the College of William & Mary as part of one of their “snap polls” of IR scholars in September, 2015. Of the 4,086 scholars polled, 694 responded, for a response rate of approximately 17%. To be included in the sample, individuals must be employed at a U.S. university in a political science department or policy school and teach or conduct research on issues that cross international borders.70 The TRIP sample thus provides us with a valuable reference point with which to compare our responses from the Knesset members, to assess how similarly or differently reputation costs are assessed at home versus abroad, by two samples that

68. Given the complex nature of recruiting political leaders for social science research, it is worth considering how representative our leader sample is of the universe of Israeli political leaders from the time frame we examined. This requires thinking through how our subjects compare both to the overall universe of Knesset members from 1996 onwards (all 408 potential respondents), as well as the sampling frame (288 subjects, since this figure does not include members who passed away, were too sick to participate or for whom we could not get contact information). In Appendix §2, we conduct supplementary analyses showing that our sample is fairly representative, although not surprisingly, current Knesset members were less likely to participate than former members.


Table 2: Knesset Sample (N = 89)

<table>
<thead>
<tr>
<th>Proportion of respondents</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Knesset Member:</strong></td>
<td></td>
</tr>
<tr>
<td>Current</td>
<td>25%</td>
</tr>
<tr>
<td>Former</td>
<td>75%</td>
</tr>
<tr>
<td><strong>Exp. on Foreign Affairs/Defense Committee</strong></td>
<td></td>
</tr>
<tr>
<td>...as backup or full member</td>
<td>67%</td>
</tr>
<tr>
<td>...as full member</td>
<td>54%</td>
</tr>
<tr>
<td><strong>Highest level of experience</strong></td>
<td></td>
</tr>
<tr>
<td>...not a Minister</td>
<td>58%</td>
</tr>
<tr>
<td>...Deputy Minister</td>
<td>29%</td>
</tr>
<tr>
<td>...Cabinet Member or higher</td>
<td>12%</td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td>84%</td>
</tr>
<tr>
<td>Served in military</td>
<td>95%</td>
</tr>
<tr>
<td>Active combat experience</td>
<td>64%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>61</td>
</tr>
<tr>
<td>Terms in Knesset</td>
<td>3.0</td>
</tr>
<tr>
<td>Military Assertiveness</td>
<td>0.61</td>
</tr>
<tr>
<td>Right Wing Ideology</td>
<td>0.45</td>
</tr>
<tr>
<td>Hawkishness (Arab-Israeli conflict)</td>
<td>0.39</td>
</tr>
<tr>
<td>International Trust</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Note: individual differences in bottom four rows scaled from 0-1.

are each “elite” in different ways: one through virtue of expertise and training, the other through their experience, position and official responsibilities.71

Our final source comes from two samples of the Israeli public (ISRAEL PUBLIC I & II), which were obtained using iPanel, an Israeli polling firm that has been used effectively by other recent surveys and experiments.72 They were fielded in October, 2015 and in January, 2016. Both samples were representative of the Israeli Jewish population, and stratified based upon gender, age, living area and education.73 Fielding these two surveys on national samples (the first of which was fielded at roughly the same time as the TRIP survey and Knesset survey) is helpful for two purposes. First, because elite samples are inevitably small in size (by definition, the bigger the sample, the less elite it is), it allows us to replicate our findings on a much larger sample. Second, because one of the

71. For an earlier example of this type of comparison between different types of elites, see Holsti and Rosenau, “The domestic and foreign policy beliefs of American leaders”; Ole R Holsti and James N Rosenau, “Liberals, populists, libertarians, and conservatives: The link between domestic and international affairs,” International Political Science Review 17, no. 1 (1996): 29–54.
73. Our focus on the Israeli Jewish population alone in this survey is due entirely to logistical constraints, specifically the inability of online polling companies in Israel to provide anything close to a representative sample of the minority Israeli Arab population.
critiques of survey or experimental work is that the results are time-bound, replicating the study four months later allows us to test how reputational effects decay over time.

METHODS

In each study we presented respondents with eight items (ten, in the TRIP survey), spread across two grids, one in regard to Syria and the other to Crimea. In each grid, we focus specifically on the evaluation phase of reputation models. The first grid asks respondents to assess the reputation costs in Syria incurred by a variety of actors using a measure of reputation costs developed by Brutger and Kertzer:

Thinking about recent events in Syria, do you think the reputation of each of the following has been strengthened, weakened, or not affected by the way they have handled the situation in Syria?

The response options ranged from “Significantly weakened” (1) to “Significantly strengthened” (5), and respondents made reputational assessments for each of the following actors:

- The United States
- President Obama
- Russia
- The United Nations
- President Putin (TRIP survey only)

They then were presented with the same set of questions, about events in Crimea.  

Four points are of note here. First, to maximize breadth, we measure reputation costs in regards to a range of actors. Second, we deliberately ask about reputation in general terms, rather than specifying specific types. Reputations are multi-dimensional, in the sense that actors may have reputations for multiple attributes simultaneously, and the behavior that may help one type of reputation may harm another. While scholars often focus on one type of reputation at a time, doing so in this context...

74. In the TRIP survey, participants were randomly assigned to receive only one of the two grids because of space constraints.
would create more problems than it would fix. Asking respondents, for example, specifically for their inferences about U.S. reputation for resolve would have ignored potential changes in other potentially relevant types of reputation, like reputation for standing up for human rights, abiding by international commitments, and so on. Rather than priming participants to weight particular types of reputational considerations more than others — which would likely bias the results in favor of canonical models of reputation — we ask respondents to make reputational assessments more generally, freeing us from excluding specific types of reputation our respondents might care about. This method dovetails with questions asked by major survey organization such as Pew, who often elicit beliefs about overall reputation rather than reputations for specific traits or qualities. Relatedly, we do not prime participants by mentioning the infamous “red line” threat of 2012; we merely ask about the “recent events” in Syria. Doing so avoids putting our thumbs on the scale and biases us against finding any difference between the two crises.

Finally, as we noted earlier, our survey focuses on the evaluation phase of reputational models, rather than the diagnostic phase. Our results thus speak to respondents’ reputational assessments, not the weight this consideration has in planning future behavior. For ease of interpretation, in our discussion of the results, we have re-centered the data such that the scale midpoint (“Not affected”) is 0, whereupon positive values indicate the actor’s reputation was positively affected (reputational gains), and negative values indicate the actor’s reputation was negatively affected (reputational costs). Our main interest is in the reputational inferences drawn about the United States, President Obama, and Russia, so we focus primarily on them in the discussion below, although the complete results, including those for the United Nations and President Putin, are presented in Appendix §1.

Results

We begin with Figure 4, which simply plots the distribution of estimated reputational consequences incurred as a result of events in Crimea (the top row) and Syria (the bottom row) by three of our actors of most central theoretical interest: President Obama (in the left-hand column), the United

---

76. Dafoe, Renshon, and Huth, “Reputation and Status as Motives for War.”
77. We base this measurement strategy on that employed by Brutger and Kertzer, “A Dispositional Theory of Reputation Costs.”
78. See, for example http://www.people-press.org/question-search/?qid=1827792&pid=51&ccid=51.
79. Because each survey was fielded at one time, we are unable to directly measure how much respondents’ beliefs have changed – only how much they report them as having changed.
The figure depicts the distribution of reputational consequences incurred by a variety of actors (President Obama, Russia, and the United States) as a result of events in Crimea (the top row) and Syria (the bottom row). To simplify the interpretation of the results, the x-axis in each panel is rescaled into three categories (corresponding to reputational losses, reputational gains, and no perceived reputational change). The estimates from foreign decision-makers — members of the Knesset — are depicted in peach, and from American IR scholars in turquoise. Thus, the figure depicts at least two important dynamics: first, as comparing the top and bottom rows shows, both samples perceive the US and Obama as incurring steeper reputational costs in Syria (where the President issued an explicit public threat) than Crimea. Second, as comparing the peach and turquoise distributions in each panel shows, American IR scholars consistently perceive the United States and Obama has incurring a lower magnitude of reputational costs than our sample of foreign decision-makers do.
States (in the right-hand column) and Russia (the middle column). In each panel, the distribution of responses from our foreign decision-makers are displayed in peach, and the responses from American IR scholars are displayed in turquoise; the x-axis has been rescaled to depict three types of respondents: those who perceived reputation costs, those who perceived reputation gains, and those who perceived no reputational change.

**UNEquivocal losses for the U.S.; mixed results for Russia**

Thus, Figure 4 suggests three patterns. First, across both of these samples, in both of these conflicts, our respondents perceive significant reputation costs incurred by the United States and its president; as we show in Figure 6, our two Israeli public samples reach a similar conclusion. In contrast, IR scholars and foreign decision-makers are less on the same page about the reputational consequences incurred by Russia: both samples see Russia as bolstering its reputation as a result of events in Syria, but whereas foreign decision-makers perceive Russia to be doing the same in Crimea, American IR scholars are more likely to see Russia as harming its reputation. Importantly, this pattern is different than what is often reported in the literature. Reputation cynics – and scholars writing more generally about reputation, like Yarhi-Milo — have consistently found evidence that leaders are concerned about their own reputations; we show here that foreign observers perceive reputation costs too. At the same time, the bimodal distributions in the top-middle panel reinforce the extent to which reputation is multidimensional, such that actors assessing reputation costs must weigh multiple reputational considerations simultaneously: even as Russia is bolstering its reputation for resolve, it is harming its reputation for compliance with international law, for example.

---


81. On the multidimensionality of reputation, see Brutger and Kertzer, “A Dispositional Theory of Reputation Costs.” Israel has a large number of immigrants from the former Soviet Union, which might be germane when asking survey respondents about reputation costs paid by Russia in particular — or reputation costs paid by any actor in response to events in Crimea, more generally. Yet a series of Wilcoxon rank-sum tests in both of the Israeli mass public samples finds little evidence that respondents who immigrated from the former Soviet Union attribute reputation costs significantly differently than those who were born in Israel. Of the sixteen tests, the only one significant at the 95% level suggests that in the second public sample, respondents born in Israel attributed steeper reputation costs to the United Nations in Syria than respondents born in the former USSR ($p < 0.02$), but the difference effect is not significant in the other public sample ($p < 0.14$), and the results for Crimea are never statistically significant.
Second, and more interesting, however, is what emerges when we compare the reputational consequences of Crimea (in the top row) and Syria (in the bottom row). Substantively, the Crimea-Syria comparison is interesting because these crises occurred in the same geopolitical climate, at roughly the same time period, and involved many of the same actors. As many others in IR have pointed out, leaders both rarely make direct coercive threats, and even more rarely make those threats and fail to follow through, an outcome that is “off the equilibrium path.”\(^{82}\) Importantly, President Obama famously made a public threat in the Syria crisis — drawing a “red line” in August 2012 in response to the Syrian leader’s use of chemical weapons — and then backing down, accepting a diplomatic compromise brokered by Russia resulting in the removal of large stockpiles of Syrian chemical weapons. The event is often understood as an indictment of audience cost theory, since the American public, reticent to enter into yet another war in the Middle East, largely approved of the president’s decision to stay out; Gallup noted that American support for US intervention in Syria was “on track to be among the lowest for any intervention Gallup has asked about in the last 20 years.”\(^{83}\) In contrast, the White House made clear its disapproval of Russian actions in Crimea, imposing several rounds of sanctions beginning in March of 2014, but avoided making explicit threats to intervene.

Comparing the distributions between the top and bottom rows of the panels shows that both samples see the US as paying steeper reputation costs in Syria than in Crimea; as we show in Figure 6, our two Israeli public samples reach a similar conclusion. This finding is consistent with one of the predictions of the canonical audience cost model, which argues that leaders should pay a cost for making explicit coercive threats and not following through on them.\(^{84}\) At the same time, however, it encourages us to look beyond the narrow purview of a domestic punishment mechanism. As Christopher Gelpi and Joseph Grieco note, audience cost theory has been overly narrow in its focus on domestic electoral politics — such that failing to find evidence of leaders being ousted from office after making public threats is erroneously held as an indication that foreign policy misadventures are costless.\(^{85}\) Our findings raise the specter of an alternative mechanism,


\(^{83}\) http://news.gallup.com/poll/164282/support-syria-action-lower-past-conflicts.aspx


in which leaders who back down on public threats suffer reputation costs in the eyes of foreign audiences.

Of course, eliciting reactions to real-life paired crises such as Syria and Crimea sacrifices experimental control and limits our ability to make causal claims. Syria and Crimea are dissimilar in many ways besides the fact or whether a U.S. President made a threat. The cost of US intervention was lower in Syria than in Crimea, for example. Similarly, it is possible that geographic proximity that might shape peoples’ reactions to the crises. For example, Israeli respondents may assess steeper reputation costs to the US as a result of Syria than Crimea because it neighbors the former and not the latter. However, the fact that we see this same pattern among American IR scholars — who perceive stronger reputation costs for the U.S in Syria compared to Crimea — suggests that geographic proximity is not “doing the causal work” that we are assigning to failing to follow through on a threat.

Reputation costs paid internationally, not domestically

Third, although American IR scholars clearly perceive the US as incurring reputation costs in both Syria and Crimea, it is striking to note that they estimate these costs as being significantly less severe than our sample of foreign decision-makers do, the differences calculated numerically in Table 3. The same is also true when we compare American IR scholars with the two Israeli public samples.

<table>
<thead>
<tr>
<th>Context</th>
<th>Actor</th>
<th>TRIP</th>
<th>Knesset</th>
<th>Difference</th>
<th>Wilcoxon test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syria</td>
<td>US</td>
<td>-0.90</td>
<td>-1.09</td>
<td>0.19</td>
<td>p &lt; 0.03</td>
</tr>
<tr>
<td>Syria</td>
<td>Obama</td>
<td>-0.82</td>
<td>-1.16</td>
<td>0.34</td>
<td>p &lt; 0.00</td>
</tr>
<tr>
<td>Syria</td>
<td>Russia</td>
<td>0.53</td>
<td>0.54</td>
<td>-0.01</td>
<td>p &lt; 0.97</td>
</tr>
<tr>
<td>Crimea</td>
<td>US</td>
<td>-0.58</td>
<td>-0.86</td>
<td>0.28</td>
<td>p &lt; 0.00</td>
</tr>
<tr>
<td>Crimea</td>
<td>Obama</td>
<td>-0.53</td>
<td>-0.91</td>
<td>0.38</td>
<td>p &lt; 0.00</td>
</tr>
<tr>
<td>Crimea</td>
<td>Russia</td>
<td>-0.22</td>
<td>0.40</td>
<td>-0.62</td>
<td>p &lt; 0.00</td>
</tr>
</tbody>
</table>

Note: positive differences indicate foreign decision-makers perceive larger reputation costs than American IR scholars.

Why do we see American IR scholars go relatively easier on the United States when assessing reputation costs than our sample of foreign decision-makers do? There are several possibilities. One possibility is that IR scholars are more likely to self-define as realists, who may be more sensi-
tive to considerations of capabilities and interests, and have been especially critical of the American foreign policy establishment’s preoccupation with reputation. Breaking down the reputational assessments of IR scholars by their self-identified paradigmatic attachments, however, finds that realist respondents never estimate significantly smaller reputation costs than non-realists — and, in the case of Obama’s reputation costs for Crimea, actually perceive larger reputation costs than non-realists. 86

Another is simply a motivated reasoning or ethnocentrism story, in which Americans feel a sense of attachment to their country, and thus underestimate the negative conclusions others will draw from their behavior, and are less critical of their foreign policy than foreigners are. 87 Yet if elites tend to be more cosmopolitan and less nationalist than the public as a whole, for good or for ill, American IR scholars should presumably be less susceptible to this tendency than the American public writ large. 88

An alternative explanation is that American IR scholars tend to be on the political left, making them more likely to support the policies of a Democratic administration, and less likely to perceive these policies as having negative reputational consequences. 89 This is particularly the case because our foreign decision-maker sample is from Israel, a country that has steadily moved to the political right in recent years, partially as a result of its security environment. 90 Indeed, we find considerable ideological disagreement about reputation costs amongst our IR scholars, with right-leaning American IR scholars perceiving significantly larger reputation costs for the United States than left-leaning ones. Interestingly, however, we also find similar ideological disagreement amongst our Knesset respondents, and left-leaning Knesset members perceive steeper reputation costs than left-leaning American IR scholars do. However, as we show in Appendix §1.1, if we re-weight our TRIP sample to match the ideological composition of our Knesset sample, we still find evidence of the “home-away” gap, suggesting that our finding cannot be attributed entirely to the ideolog-

86. The results from a set of Wilcoxon rank-sum tests find that realists and non-realists don’t significantly differ in their assessments of the US’s reputation costs for Syria (p < 0.57), Obama’s reputation costs for Syria (p < 0.74), or the US’s reputation costs for Crimea (p < 0.31), but do differ in their assessments of Obama’s reputation costs for Crimea (p < 0.07), with realists perceiving Obama as paying steeper reputation costs than non-realists do.
ical leanings of the Academy. Moreover, our sample also includes many former decision-makers, from sessions of the Knesset that was more left-leaning than it is today, such that if anything this analysis underestimates the magnitude of the reputation costs perceived by those decision-makers currently in power.

A final interpretation of this pattern might be attribution bias. As Mercer notes, making reputational judgments involve making an inference about an actor’s disposition or type, rather than the broader situation or environment the actor happens to face.\textsuperscript{91} The fundamental attribution error holds that we tend to neglect the role of situational causes when explaining the behavior of others; we acknowledge the role of mitigating situational factors when analyzing our own actions, but downplay them when analyzing the actions of others.\textsuperscript{92} Thus, Americans might perceive their country’s reputation costs as being smaller than foreign observers do because they place a greater weight on these environmental constraints.

Regardless of its cause, however, the presence of this perceptual gap between American IR experts and foreign observers is of considerable theoretical importance. Critics of Obama’s foreign policy legacy (particularly on the political right) have suggested that the administration’s inaction in Crimea and Syria have sullied the country’s reputation, yet Obama did not appear to pay any consequences domestically for his inaction.\textsuperscript{93} IR scholars have thus far used this to suggest a potential weakness in audience cost theory, which is sometimes criticized for exaggerating the costliness of empty threats.\textsuperscript{94} Yet, in focusing so exclusively on the domestic side of the audience cost model, IR scholars have perhaps neglected the international side; the starker reputation costs with Syria appear to have manifested themselves abroad, rather than at home.

**Strong Correlation Between Leader and Country Reputations**

At the same time, however, this begs the questions of to whom these reputation costs adhere — a question that our study is also equipped to answer.\textsuperscript{95} Figure 5 thus presents the results another way: comparing the reputation costs our respondents in each of our four samples attributed to

91. Mercer, *Reputation and international politics*.
95. Renshon, Dafoe, and Huth, “Leader Influence and Reputation Formation in International Politics.”
The figure depicts the differences between the reputational consequences attributed to President Obama versus the reputational consequences attributed to the United States as a whole. Each panel displays the distributions for each of our four samples (the Israeli Knesset, two samples of the Israeli public, and the American IR scholars surveyed by the TRIP snap poll); the left panel shows the within-subject leader-specific reputation effects in Crimea, and the right for Syria. As the preponderance of green in both panels shows, most respondents attribute the same reputational effects for the US as for President Obama. However, of the minority of respondents that do assign different reputational costs for the two actors, most attribute steeper reputation costs for the President than for the country as a whole. See Appendix §1 for an analysis of leader-specific reputation costs for President Putin.
President Obama versus the United States writ large. To facilitate a clearer interpretation, we calculate this as a within-subject estimate and trichotomize it to distinguish between three types of respondents: those who allocate the same amount of reputational costs to both Obama and the United States (depicted here in green), those who perceive Obama as bearing less reputational cost than the United States (depicted here in blue), and those who perceive Obama as bearing more reputational cost than the United States (depicted here in red).

The plot shows that reputational judgments about the US and President Obama are highly correlated with one another. Across all four samples, most respondents (65%-86%, depending on the sample and crisis) attribute identical reputational costs to both the president and the country as a whole. In this sense, the findings are consistent with Brutger and Kertzer, who find a similarly strong correlation between leader and country-specific assessments of reputation costs in the context of the domestic politics of reputation. Interestingly, however, we also see a sizable minority of respondents (14%-35%, depending on the sample and crisis) who allocate reputational costs differentially between the leader and country as a whole. As the top and bottom portions of the plot show, of the respondents who differentiated between country and leader-specific reputations, most tended to allocate greater costs to the leader than the country; this is especially the case with the Syrian conflict. The interesting exception here, again, is the TRIP survey, where American IR scholars were significantly more likely to give Obama a break than the foreign samples were.

Two points are noteworthy here. The first is that the 6-28% of respondents who allocate greater reputational costs to the leader rather than the country are doing so in a manner consistent with theories of leader-specific reputations which posit that reputations are more likely to adhere to actors who are influential in the state’s decision-making process; given the centrality of Obama’s involvement in the Syrian episode in particular, it is not surprising that the reputational consequences should fall on his shoulders. Our existing theories, however, have more difficulty explaining the existence of the blue portion of the plot: the 6-12% of respondents (depending on the sample and crisis) who give Obama a break and attribute greater reputational costs to the country as a whole, even though the episodes occurred during his presidency.

---

97. Renshon, Dafoe, and Huth, “Leader Influence and Reputation Formation in International Politics.”
98. A series of linear probability models regressing whether respondents attributed steeper reputation costs to Obama or to the United States as a whole suggests the patterns depicted in Figure 5 aren’t simply due to liberals giving Obama a break. In the TRIP sample, political ideology is not a significant predictor for either Crimea or Syria; in the two Israeli public samples, the results vary somewhat by sample and context, but politically knowledgeable respondents, older respondents, and non-Ashkenazi respondents are generally less likely to give Obama a break.
REPUTATIONAL EVALUATIONS SIMILAR IN MASS PUBLIC, AND STABLE OVER TIME

Questions about the nature of survey samples are longstanding in social science surveys and experiments. One can see a straight line running from early critiques that the “science of human behavior” was the “science of sophomores” to later concerns within political sincere that MTURK workers may be inappropriate subjects.\(^99\) Because the field of International Relations is often (though not always) concerned with political elites and leaders, scholars within that field have faced even more scrutiny: after all, it is one thing to replace telephone surveys of U.S. citizens with online marketplaces like Amazon Mechanical Turk, but quite another to ask students to “stand in” for world leaders.\(^100\) In general, these concerns about the “narrow database” that political scientists have access to ultimately coalesce around the question: would our results look different if we had gained access to different (and in IR, more elite) subjects?\(^101\)

As a result of our research design, we can address the question of how “sample-bound” our results were directly. Recall that we fielded identical survey questions to members of the Israeli Knesset and a much larger representative sample of the Israeli public simultaneously. Figure 6 provides an overview of the distributions for each response item. On first glance, the results appear strikingly similar, but we test this more formally in Table 4. There we show that, for both public samples, we fail to detect statistically significant differences between elites and the public on Syria. For Crimea, we find the public perceives slightly less steep reputation costs for the US than the Knesset members do, although not for Obama, where elites and publics do not significantly differ from one another. On the whole, then, foreign elites and their publics seem to assess reputation costs similarly.

Thus, we see a remarkable amount of congruence in how Israeli leaders and the public whom they represent view the reputational costs of Syria and Crimea. Both groups saw Syria (and the failure to follow through on a threat) as hurting Obama’s reputation as well as that of the United States. Both respondent pools also saw the events in Crimea as harming the reputation of the

---


100. Alex Mintz, Steven B Redd, and Arnold Vedlitz, “Can we generalize from student experiments to the real world in political science, military affairs, and international relations?,” *Journal of Conflict Resolution* 50, no. 5 (2006): 757–776.

The figure depicts the distribution of reputational consequences incurred by a variety of actors (President Obama, Russia, and the United States) as a result of events in Crimea (the top row) and Syria (the bottom row). To simplify the interpretation of the results, the x-axis in each panel is rescaled into three categories (corresponding to reputational losses, reputational gains, and no perceived reputational change). The estimates from foreign decision-makers — members of the Knesset — are depicted in blue, and from each of the two national Israeli samples in red and green. The figure shows a striking congruence between the reputational judgments of the Knesset members and the public at large, as well as an absence of decay effects between the two public samples, which were fielded four months apart from one another. The only difference that is statistically significant is in the top-right panel, where Knesset members perceive the reputation costs incurred by the US in Crimea to be steeper than the public samples.

<table>
<thead>
<tr>
<th>Context</th>
<th>Actor</th>
<th>Knesset</th>
<th>Public I</th>
<th>Public II</th>
<th>Wilcoxon test I (Knesset/Public I)</th>
<th>Wilcoxon test II (Knesset/Public II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syria</td>
<td>US</td>
<td>-1.09</td>
<td>-1.03</td>
<td>-0.94</td>
<td>$p &lt; 0.65$</td>
<td>$p &lt; 0.15$</td>
</tr>
<tr>
<td>Syria</td>
<td>Obama</td>
<td>-1.16</td>
<td>-1.20</td>
<td>-1.15</td>
<td>$p &lt; 0.66$</td>
<td>$p &lt; 0.99$</td>
</tr>
<tr>
<td>Crimea</td>
<td>US</td>
<td>-0.86</td>
<td>-0.70</td>
<td>-0.71</td>
<td>$p &lt; 0.04$</td>
<td>$p &lt; 0.04$</td>
</tr>
<tr>
<td>Crimea</td>
<td>Obama</td>
<td>-0.91</td>
<td>-0.80</td>
<td>-0.81</td>
<td>$p &lt; 0.21$</td>
<td>$p &lt; 0.20$</td>
</tr>
</tbody>
</table>

Table 4: Elite-public congruence: numbers under Knesset/Public columns indicate how much reputation of actor was hurt in a given context.
U.S. and Obama, with only slight differences in the magnitude for the country’s reputation. This fits in broadly with recent research dedicated to adjudicating the suitability of non-elite samples in IR research. Some present the results of classic behavioral experiments (often related to risky decision-making and Prospect Theory) fielded on elites, then compare those results to mass public studies, or to the dozens of extant studies on student samples. Others present the results of experiments fielded simultaneously on elites and the general public and are centered on vignettes that require context-specific judgment relating to issues of credibility and reputation in IR. Thus, across different samples of elites, different contexts, and different research designs, we have a host of evidence that the judgments and decision-making of leaders may, in many cases, mirror that of the citizenry from which they are drawn.

Our research design also permits us to address the second question (about whether our conclusions are time-bound) by leveraging another unique aspect of our research design: we asked identical questions of both samples of the Israeli public, and fielded the surveys three and a half months (105 days) apart. During that time, a litany of events occurred that might have shifted the public’s beliefs. Russia began their bombing campaign at the end of September, just a few days before the first survey was fielded, but in between the two surveys, the campaign intensified dramatically, with hundreds of casualties on both the government and rebel sides. During that period, Israeli Prime Minister Benjamin Netanyahu admitted publicly for the time that Israeli forces were operating in Syria. During that time in the Ukraine/Crimea crisis, the UN General Assembly voted to appoint Ukraine to the UN Security Council, local elections were held in Ukraine, Ukraine engaged directly with rebel forces for the first time in months, explosions near four different Ukrainian power lines triggered a near-total blackout of Crimea, the EU extended their sanctions against Russia, and Putin met Obama (at the Paris conference) and later Secretary of State John Kerry in person to discuss Crimea.

Despite these large and small events, our data shows that the Israeli public’s assessment of reputation costs were highly stable between October 2015 and January 2016. We test this formally

105. For a more detailed timeline, see Figure 1.
by estimating a series of Wilcoxon rank-sum tests (see Appendix §1 for results). Our results here are quite clear: we find little evidence of reputation costs decaying or changing over time. The one exception to this is that perceptions of how much the reputation of the United States was harmed in the context of Syria did exhibit a statistically significant increase ($p < 0.009$), but the substantive effect (0.09, equivalent to 1.1 percentage points) was very small, and none of the other effects significantly change.

Conclusion

Debates about reputation and credibility in international politics — spanning three generations of research — have loomed large in both the academic and policy-making communities. In this article, we have sought to make two contributions. First, we suggest that the “reputation debate” in IR is really two distinct debates, each focusing on a different phase of reputational inference, with some scholars investigating the evaluation phase (do observers draw reputational inferences in response to past actions in the manner our theories suggest?) and others focusing on the diagnostic phase (how much do observers weight past actions when formulating present behavior?). Second, since reputations are beliefs, we introduce a new type of evidence to the debate, leveraging a unique survey research design asking identical questions from respondents from four samples in two countries. Together, they shed insight on what we see as three critical questions relating to the diagnostic phase of reputation: (1) do reputational costs exist (and if so, when are they generated)? (2) Who pays these costs: states, leaders or both? (3) What are the international consequences of backing down?

A number of important findings emerge from our data. First, all of our respondents see the United States and President Obama as having incurred significant reputational costs as a result of both Syria and Crimea. Not all events affect reputations equally, however, and despite Syria and Crimea having occurred at roughly the same time and involving similar actors, they led to drastically different reputational costs. Of course, one of those instances involved a rare phenomenon in world politics: a leader making a clear, dramatic “red line” threat (rare enough on its own) and then deciding not to follow through once challenged. Our data indicates the effects of this action: every one of our samples saw the U.S. as paying reputation costs in Syria. It seems clear that failing to follow through on “red line” threats does indeed come at a cost as far as reputational beliefs
are concerned. Importantly, we do not argue that Obama’s decision not to intervene in Syria was misguided, and evaluating the success of his Syria policy is far outside the scope of our paper. But our findings do suggest — and contrary to what some supporters of Obama’s policy have maintained — his decision to back down from his threat carried important reputational costs in the eyes of a close US ally. We also find that our observers assessed the reputational cost in Syria — where Obama issue an explicit threat — to be steeper compared to Crimea. This finding underscores the importance observers attribute to signals of resolve that explicitly stake the leader’s reputation for resolve. Applying these findings to current crises, President Trump should exercise caution in issuing public hands-tying threats to the leaders of Iran, Syria, North Korea or Russia. Backing down on those threats could lead to a significant reduction in his credibility, which according to recent polls, is already low.

A second important finding of our study is that reputation costs are in the eye of the beholder. Specifically, our data clearly point to what we called the “home-away gap”: U.S. foreign policy scholars seemed far more sanguine about the reputational costs incurred by their own country in Syria and Crimea than did our Israeli respondents. While our research design cannot adjudicate between explanations for this pattern — whether attribution bias, a more liberal sample, or simple ethnocentrism — the implications do help us refine our understanding of reputation. If audience cost models in IR focus our attention on the domestic consequences of failing to follow through on threats, our data suggest a renewed focus on the international side of the equation that is more in line with the writing of Schelling and Jervis on this topic. Ketian Zhang, for example finds—from numerous interviews with Chinese officials—that China redoubled its land reclamation actions in the South China Sea following what they perceived as Obama’s weak response to the Ukrainian crisis in 2014.\footnote{Ketian Zhang, “Calculating Bully – Explaining Chinese Coercion” (PhD diss., Massachusetts Institute of Technology, 2018). See also Ketian Zhang, “Killing the Chicken to Scare the Monkey: Explaining Chinese Coercion in The South China Sea” (Working paper, 2018).}

The policy implications of this finding are clear: whatever reputational costs President Obama paid for not following through in Syria redounded internationally rather than domestically. Similarly, the chief cost to president Trump’s decision to withdraw from the Iran nuclear deal — which arguably damaged US reputation for keeping its commitments — may not be imposed by American voters but rather by international actors.

Third, our surveys helped refine our understanding of how reputations “attach” to different
actors. While most first-generation work on reputations assumed that countries were the sole owners of their reputations, more recent work has shown that reputations may adhere to leaders as well. Our surveys gave us a window into how a number of different types of observers actually apportion responsibility differentially to leaders and states. The vast majority of our respondents attributed responsibility about equally between Obama and the United States. However, a significant portion (between 14-35% depending on the sample and context) do not follow this pattern. Of those observers, most tended to “blame” Obama more than the United States, a finding that fits neatly with the theory proposed above: the domain of foreign policy is exactly where we would expect observers to assume that the U.S. President has wide discretion, and in both cases (but particularly Syria), Obama and his actions served as a focal point for understanding the crisis. One lesson to draw from those findings is that US reputation for resolve on issues of foreign policy seems to hinge especially strongly on the actions of its president. Thus, the credibility of President Trump — especially the consistency between his words and actions — will guide how observers will judge the credibility of the United States more broadly.

Finally, the survey data suggests that while allies continue to assess the credibility and reputation for resolve of the United States, they are also closely monitoring the reputation for resolve of a resurgent Russia. Our findings clearly show that in the eyes of local allies in the Middle East, Putin has played his cards quite well. The crises in Syria (and to a lesser extent in Crimea) seem to have enhanced Russia reputation for resolve, while it weakened that of the United States. Future studies could probe whether those perceptions have faded or solidified following the crises over Syria under President Trump.