A Dispositional Theory of Reputation Costs

Ryan Brutger and Joshua D. Kertzer

Abstract Politicians frequently turn to reputational arguments to bolster support for their proposed foreign policies. Yet despite the prevailing belief that domestic audiences care about reputation, there is very little direct evidence that publics care about reputation costs, and very little understanding of how. We propose a dispositional theory of reputation costs in which citizens facing ill-defined strategic situations turn to their core predispositions about foreign affairs in order to weigh competing reputational dimensions. Employing a diverse array of methodological tools—from vignette-based survey experiments to automated text analysis—we show that the mass public has a "taste" for reputation, but understands it in fundamentally different ways, with hawks concerned about the negative reputational consequences of inconsistency, and doves equally concerned with the negative reputational consequences of belligerence and interventionism. In illustrating how reputation costs are in our heads, our findings offer both good and bad news for theories of reputation in IR.

Politicians frequently turn to reputational arguments to bolster support for their proposed policies: Lyndon Johnson famously lamented that if the United States withdrew from Vietnam, the US "would be seen as an appeaser and we would find it impossible to accomplish anything for anybody anywhere on the entire globe," while Ronald Reagan claimed that if the United States lost in Central America, "our credibility would collapse and our alliances would crumble."[1] International Relations (IR) theory tells us that even though we can never fully overcome "the problem of other minds" and directly access the motivations of others, the public reasons actors resort to when justifying their policies are nonetheless worth paying attention to because they reflect their underlying beliefs about the types of arguments that will resonate with their audience.[2] Thus, although it is possible that the elites making reputational arguments are simply speaking to other elites rather than the public at large, the IR literature generally assumes that the same "cult of reputation"

This is one of several joint articles by the authors; the ordering of names reflects a principle of rotation. Thanks to Cameron Ballard-Rosa, Mark Crescenzi, Jeff Frieden, Nikhar Gaikwad, Charlie Glaser, Michael Horowitz, Peter Katzenstein, Amanda Kennard, Katja Kleinberg, Akos Lada, Melissa Lee, Tyler Pratt, Brian Rathbun, Jonathan Renshon, Rob Schub, Alastair Smith, Brandon Stewart, Rob Trager, Keren Yarhi-Milo, and George Yin for helpful comments and conversations, as well as audiences at ISA and Peace Science, and Jon Pevehouse and two anonymous reviewers. We gratefully acknowledge the support of the Niehaus Center for Globalization and Governance, the Fund for Experimental Social Sciences, and the Center for International Security Studies at Princeton University.

1. Mercer 1996, 2; Milliken 1996, 223.
2. Krebs and Jackson 2007; Morgenthau 1985, 5; Risse 2000.

International Organization 72, Summer 2018, pp. 693–724
© The IO Foundation, 2018
doi:10.1017/S0020818318000188

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

that consumes policymakers captures members of the public as well.[3] Indeed, reputational logics are posited to explain why publics punish leaders deemed culpable for unsuccessful wars, or penalize decision makers who sully the national honor by backing down on threats.[4] If publics respond to reputational rhetoric, elites have all the more incentive to invoke it.

Yet we have relatively little direct individual-level evidence testing whether (and exploring how) ordinary citizens care about reputation in foreign policy. The closest political scientists have come to studying the microfoundations of reputational judgments in crisis bargaining comes from a burgeoning literature on audience costs, but the reputational mechanism in this research program tends to be imputed rather than formally tested. In this paper, we turn to survey experiments to explore the kinds of reputational judgments ordinary citizens make in response to their country's behavior on the world stage. Our experimental evidence is intended to build a better understanding of the microfoundations of reputation costs in IR.

We propose a dispositional theory of reputation costs in which citizens facing ill-defined strategic situations turn to their core predispositions about foreign affairs in order to weigh competing reputational considerations. As a result, although the mass public indeed has a "taste" for reputation, it understands reputation in fundamentally different ways, with hawks concerned about the negative reputational consequences of inconsistency, and doves equally concerned with the negative reputational consequences of belligerence and interventionism. In this sense, our theoretical approach departs dramatically from canonical models of reputation in IR. Whereas classic models of reputation costs tend to model reputations as *unidimensional* by focusing on only one trait at a time, such that the crisis bargaining literature treats reputation as synonymous with reputation for resolve, we model reputations as *multidimensional*–actors can have reputations for multiple traits simultaneously, which raises important theoretical questions about how actors aggregate across these competing considerations. Whereas classic models of reputation costs view reputation as *situationally* derived, a property of the strategic environment that different actors assess in similar ways, we show that reputational costs are heavily *dispositionally* derived, in that different observers can look at the same action and draw different reputational conclusions. Whereas classic models of reputation costs treat reputational considerations as conceptually distinct from assessments of the intrinsic interests at stake in a dispute, we suggest that the two are endogenous: the same dispositions that determine how individuals assess the intrinsic merits of policies affect their estimates of the policies' reputational consequences.

Our findings thus offer both good and bad news for domestic theories of reputation in IR. The bad news is that because assessing reputation is an ill-structured problem, reputation costs are largely in our heads, driven more by our prior dispositions about foreign policy than the actual behaviors leaders engage in on the world stage. The

3. Saunders 2015; Tang 2005.
4. Croco 2011; Fearon 1994.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

good news is that reputation costs *are* in our heads: the mass public does indeed care about reputation, such that reputation costs are politically consequential—even if abiding by a different logic than that advanced by canonical models of reputation in IR.

The discussion that follows proceeds in three parts. First, we ground our investigation of reputation costs in the existing literature, which has tended to assume the resonance of reputation costs in the general public, but has rarely systematically explored their microfoundations. Second, drawing on research in foreign policy decision making and political psychology, we offer our dispositional theory of reputation costs in which actors in ill-defined situations turn to core dispositions to assess reputation costs—contrasting this with the typical portrayal of reputation in IR in which reputations tend to be treated as situational, and reputation costs as distinct from approval of the policies themselves. Third, we discuss our experimental design, and delve into our results, showing that although domestic audiences indeed care about reputation costs, they disagree on what generates them in the first place, with hawkish respondents emphasizing the negative reputational consequences of inconsistency, and dovish respondents equally concerned about the reputational consequences of belligerence. We find further evidence of this heterogeneity using structural topic models which show that the language hawks and doves use when they discuss reputation costs varies meaningfully and systematically, and moderated mediation models which show that although reputation costs are indeed present for both hawks and doves, domestic audiences are more polarized about reputation than our models in IR often assume. The findings have important implications for a variety of questions, ranging from the domestic politics of crisis bargaining, to the psychology of hawks and doves, to the consequences of complexity for our theoretical models in IR.

Reputation Costs in International Relations

There is a variety of work in IR that presumes that public opinion produces incentives for policymakers, which then shapes state behavior. Many of these models specifically focus on reputation costs.[5] For these models to work, two assumptions must be met. First, the mass public must care or otherwise have preferences about reputation, and second, policymakers must respond to these concerns.[6] In this piece, we focus on establishing this first link, exploring the microfoundations of reputation costs.

Although the literature on reputation in IR is vast and polyphonic, it suggests three broad points of particular importance for our purposes.[7] First, reputations are beliefs

5. Croco 2011; Fearon 1994; McGillivray and Smith 2008; Potter and Baum 2014.
6. On rational actors having a "taste" for reputation, see Becker 1974.
7. For recent reviews, see Dafoe, Renshon, and Huth 2014; Sharman 2007.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

about a trait or tendency of an actor. They are Janus-faced, informed by behavior in the past, and used to predict behavior in the future.[8] When a policy negatively affects others' beliefs about a characteristic that an actor values, she incurs a *reputation cost*.

Second, although scholars disagree about how much control actors have over their reputations,[9] whether reputations adhere to leaders or states,[10] and why actors care about their reputations in the first place—whether because of instrumental reasons stemming from a logic of consequences (e.g., states cultivating a reputation for resolve to prevent being seen as an easy target), or normative reasons deriving from a logic of appropriateness (e.g., leaders carrying out actions to preserve the national honor), there is little disagreement that decision makers care about them, particularly with regard to reputations for resolve.[11] Walt claims the US suffers from a "credibility addiction" in which a desire to avoid reputation costs leads policymakers to pursue policies that harm the national interest, while Tang notes that policymakers subscribe to a "cult of reputation," a "persistent obsession" with reputation costs manifested in Henry Kissinger's proclamation that "no serious policymaker could allow himself to succumb to the fashionable debunking of 'prestige,' or 'honor' or 'credibility.'"[12]

Third, because international relations is characterized by two-level games, reputation costs have both a foreign and domestic component.[13] Abroad, reputation matters as a *first-order belief* about a country or leader's characteristics. Many times when we study whether reputation matters in IR, we are interested in exploring the content of these first-order beliefs: how did the Germans make judgments about British and French credibility during the appeasement crises of the late 1930s?[14] What inferences did the Soviets draw from American defeats in Vietnam?[15] At home, however, reputation matters as a *second-order belief*: what domestic constituencies think others think about the country or leader's characteristics.[16] The domestic side of reputation costs may partially explain why leaders seem to care so much about reputation even though the empirical record suggests foreign rivals may not draw inferences from past actions in the manner that policymakers assume.[17] As Fearon points out, governments are on average more likely to lose power internally than by foreign invasion, which means that if domestic audiences care about their country's reputation, losing face can have more immediate consequences at home than abroad, whether by

8. Dafoe, Renshon, and Huth 2014, 372; Mercer 1996, 6.
9. Huth 1997; Mercer 1996.
10. Brewster 2009; Jervis 1982; Renshon, Dafoe, and Huth 2018.
11. O'Neill 1999; Schelling 1966; Yarhi-Milo 2018.
12. Tang 2005, 35; Walt 2015.
13. Putnam 1988.
14. Press 2005, chap.2.
15. Hopf 1991.
16. On higher-order beliefs and reputation, see Dafoe, Renshon, and Huth 2014; O'Neill 1999. For another application of second-order beliefs, see Mildenberger and Tingley 2017.
17. Hopf 1991; Mercer 1996; Press 2005; Shannon and Dennis 2007; Snyder and Diesing 1977, 187, but see Weisiger and Yarhi-Milo 2015.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

impeding leaders' legislative agendas, or causing them to lose office altogether.[18] Thus, IR scholars point to reputational mechanisms to explain why publics punish leaders deemed culpable for unsuccessful wars, why domestic audiences punish leaders who back down on threats in foreign policy crises, and why states use secrecy to dampen escalation pressures from both domestic and foreign audiences.[19]

Given the domestic side to reputation costs, it should not be a surprise that policymakers routinely invoke reputational logics to justify their policy choices, particularly regarding military interventions, where decision makers often face a tradeoff between saving face and losing lives.[20] In a public radio address at the outset of the Korean War, Harry Truman justified American intervention by pronouncing that failing to stand up to aggression in Korea "would be an open invitation to new acts of aggression elsewhere."[21] In a public television address justifying the invasion of Cambodia, Richard Nixon suggested that the reputation costs that would ensue would damage not only the interests of the United States, but the international community, expressing concern that "if, when the chips are down, the world's most powerful nation … acts like a pitiful, helpless giant, the forces of totalitarianism and anarchy will threaten free nations and free institutions throughout the world."[22] In publicly televised addresses explaining his decisions to send troops to Somalia and Haiti in 1993 and 1994, Bill Clinton repeatedly emphasized reputation costs, warning that in the absence of intervention "our credibility with friends and allies would be severely damaged," stressing the importance of upholding "the reliability of the commitments we make and the commitments others make to us," and warning that failing to respond would lead to "open season on Americans."[23] Both advocates and critics of the 2003 Iraq War couched their arguments in terms of reputation, with the former alleging that withdrawal from Baghdad would embolden American enemies to unleash further attacks on the United States, and the latter arguing that continued American presence in Iraq was soiling America's reputation with its allies.[24]

The frequency with which leaders resort to reputational rhetoric in these addresses is important because the public reasons leaders invoke when seeking to justify or legitimate their policies reflect their underlying beliefs about the types of arguments that will resonate with their audience, and thus the normative environment in which they are embedded.[25] It is possible of course, that politicians invoking reputational

18. Gelpi and Grieco 2015; Fearon 1994, 581.
19. Carson 2016; Croco 2011; Poznansky 2015; Tomz 2007a.
20. Kertzer 2016
21. Press 2005, 2.
22. Morgan 1985, 139.
23. Clinton 1993, 1994.
24. McGovern and Polk 2006. That many of these examples include cases of reputation costs being invoked despite the absence of a public threat shows the merit of studying reputation costs outside of the narrower confines of "audience cost" frameworks.
25. Goddard and Krebs 2015; Krebs and Jackson 2007; Mitzen 2005; Müller 2004; Risse 2000; Schimmelfennig 2001. Typically, the rhetorical action literature has to confront the question of whether actors themselves believe the reasons they are invoking, with some scholars conceding that we might not be able to tell whether an actor has internalized a norm or is merely employing it strategically.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

justifications are speaking primarily to other politicians, but the fact that decision makers employ these types of arguments publicly as much as they do privately points to a belief that reputational consequences count as "publicly acceptable reasons." If elites did not think the public cared about reputation, they would use these arguments with each other, but not with the rest of us.

Despite the prevailing belief that domestic publics indeed care about reputation costs, there is very little direct empirical evidence that publics have a "taste" for reputation, and very little theoretical understanding of how. Experiments testing market-entry deterrence games in economics, for example, have found that participants generally care less about reputation than formal models predict.[26] Developmental psychologists have found evidence of concerns about reputation costs in children as young as five, but it is unclear how these concerns translate into a political domain where the reputations at stake belong to leaders and countries.[27]

In an explicitly political domain, a proliferating body of research on domestic audience costs has pointed to reputation as the mechanism underpinning the drop in public support that occurs for leaders who fail to follow through on their threats in foreign policy crises, but in the experimental studies best suited to testing reputation costs, the reputational mechanisms tend to be inferred indirectly or addressed in brief, rather than as the subject of systematic investigations in their own right.[28]

We believe pursuing microfoundations for reputation costs is thus a valuable enterprise: as Kertzer argues, building microfoundations for difficult-to-measure constructs like reputation and resolve is especially beneficial because without adequate microfoundations, political scientists are more likely to veer toward tautology, inferring the effects of the construct from the same outcomes we use it to explain.[29] Brewster also levies this charge against the study of reputation, where she notes that unless we can directly measure "reputation costs … any claim about the power of reputation remains non-falsifiable and therefore has less theoretical force."[30] As Gelpi and Grieco point out, one of the strengths of the surge of recent experimental work on audience costs is that it provides direct evidence for the audience cost mechanism in a manner that previous observational work could not.[31] Exploring the microfoundations of reputation costs and providing direct evidence for their existence would perform a similarly valuable service.

Given the perceived strength of the "cult of reputation," however, there is less doubt here whether leaders believe in the reputational arguments they make. Nonetheless, this is a secondary point for our purposes, since regardless of whether leaders themselves have internalized the importance of reputation costs, the fact that they are are willing to employ reputational justifications strategically signals their second-order belief about whether their audiences have internalized it.

26. Tingley and Walter 2011.

27. Hill and Pillow 2006.

28. Brutger 2017; Davies and Johns 2013, 2016; Fearon 1994; Kertzer and Brutger 2016; Levendusky and Horowitz 2012; Potter and Baum 2014; Schultz 2001; Tomz 2007a; Trager and Vavreck 2011; Weeks 2008.

29. Kertzer 2017a.

30. Brewster 2011, 324. For important exceptions, see Levy et al. 2015; Wallace 2015.

31. Gelpi and Grieco 2015, 441.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

Microfoundations for Reputation Costs

Our theory of the domestic politics of reputation costs is built on the premise that international relations is replete with what decision theorists and foreign policy scholars call *ill-structured* problems.[32] Unlike the emphasis of much rationalist and constructivist work on common knowledge or intersubjectivity, many problems in IR are characterized by complex or ambiguous environments where meaningful probability distributions are unknown, and actors often disagree on the rules of the game, or what situation they are facing.[33] Given that actors are often uncertain about the consequences of policy choices, and do not possess their own reputations and thus can only access them by drawing inferences about the beliefs of others, assessing reputation costs is a daunting task.[34]

What makes calculating reputation costs particularly difficult is that reputation is *multidimensional*: actors can have reputations for multiple attributes simultaneously.[35] Indeed, actors in IR can have reputations for a panoply of characteristics, from resolve, to empathy, reliability, compliance, hostility, repaying sovereign debt, respect for human rights, honesty, and so on.[36] Behavior that may help one type of reputation may hinder another. Following through on a threat may help foster a positive reputation for consistency, for example, but also a negative reputation for belligerence. Thus, to make our theoretical models more tractable, classic models of reputation tend to bracket its multidimensional nature and focus on only one type of reputation at a time—in the crisis-bargaining literature, usually reputation for resolve.[37] However, actors in international politics, whether elites or masses, do not have the same luxury. Thus, the public's overall assessment of how foreign policy actions shape their country's reputation involves weighting multiple types of reputation simultaneously.[38]

In treating reputation as unidimensional, existing scholarship on reputation has neglected to explore how actors weigh competing reputational considerations. To be clear, we are not faulting IR scholars for focusing on the particular types of reputation that interest them. As Lake and Powell note, "all theories attempt to simplify a complex reality, and … reflect judgments made by theorists as to what to put into their analyses and what to leave out."[39] In so doing, however, IR scholars have ended

32. Hermann 1990; Levy 1994; Simon 1973; Stein 1994; Voss and Post 1988.

33. Fearon and Wendt 2002; Grynaviski 2014; Kertzer 2017b; Nelson and Katzenstein 2014; Stein 2002. Blyth notes that "structures do not come with an instruction sheet." 2003, 695.

34. Herrmann 1988; Mercer 1996.

35. Finnemore 2009, 74–75; Mercer 1996, 7.

36. Crescenzi 2007; Crescenzi et al. 2012; Guisinger and Smith 2002; Guzman 2008; Huth 1997; Kertzer 2016; LeVeck and Narang 2017; Murdie and Davis 2012; Sartori 2005; Simmons 2010; Tomz 2007b; Yarhi-Milo and Holmes 2017.

37. Dafoe, Renshon, and Huth 2014; Weisiger and Yarhi-Milo 2015.

38. On the weighting of multiple attributes in choice tasks, see Keeney and Raiffa 1993; Mellers, Schwartz, and Cooke 1998.

39. Lake and Powell 1999, 14.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

up with what Fearon and Wendt call a "tacit ontology," where simplifying assumptions made for purely analytic reasons end up shaping the questions we ask—and thus, the answers we get—about the constructs we study.[40] The fact that reputation costs become exponentially more difficult to calculate if we allow actors to have reputations for multiple traits simultaneously—such that we need to deliberately simplify our models to avoid theoretical chaos—is theoretically important because the same computational complexity that bedevils analysts is also confronted by the actors whose behavior we seek to model.[41] How do actors in IR go where IR scholars fear to tread?

We argue that, as is often the case when facing ill-structured problems, individuals turn to heuristics: cognitive shortcuts that seek to make complex problems simple.[42] Two such simplifying shortcuts are relevant to the study of reputation costs. First, members of the public can defer to elite decision makers to define the situation for them, consistent with an influential literature on elite cue-taking in foreign policy.[43] Indeed, much of the presidential rhetoric we described can be thought of as instances of leaders attempting to construct reputation costs in the minds of their constituents, although how much the public swallows what elites are feeding them is an open question—as is the question of how elites themselves come to make these assessments.[44] Second, rather than being dependent upon elites to define reputation costs for them, members of the public can rely on their pre-existing dispositions or orientations toward foreign affairs, an instance of a broader phenomenon of dependence on prior belief systems in the formation of political judgments.[45] Even foreign policy experts have been found to be "prisoners of our preconceptions," "more theory-driven than data-driven in analyzing and making recommendations about international relations."[46] Although there are important differences between these two images of public opinion in foreign affairs—the former is top-down, the latter is bottom-up—these two mechanisms complement one another: research on persuasion shows us that messages are most likely to resonate when they tap into pre-existing beliefs or orientations, such that understanding when messages work

40. Fearon and Wendt 2002, 35.

41. On the importance of procedural rationality for our theories of IR, see Rathbun, Kertzer, and Paradis 2017.

42. Gigerenzer 2008.

43. Berinsky 2009; Zaller 1992.

44. Kertzer and Zeitzoff 2017; Sylvan and Voss 1998.

45. Fiske and Taylor 2013; Herrmann, Tetlock, and Visser 1999; Holsti 2004; Rathbun et al. 2016. We borrow from attribution theory in social psychology to use "dispositional" here to refer to characteristics or attributes of actors themselves: in our case, general orientations or belief systems toward foreign policy. See Ross 1977. For a similar application of this framework, see Kertzer 2017b.

46. Koopman, Snyder, and Jervis 1990, 716; Tetlock 1999. See also Murray 1996. In this sense, it is striking how our understanding of public opinion in foreign policy has evolved: the concern of the Almond-Lippmann consensus was that the mass public is too data driven, such that public opinion is inherently unstable; political scientists are now concerned the public is too theory driven, assimilating information about the world through ideological blinders. See See Koopman et al. 1995, 378; Nyhan and Reifler 2010.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

requires understanding individuals' predispositions or belief systems.[47] Our focus here is thus on this latter mechanism, and the role that general predispositions about foreign affairs play in shaping calculations of reputation costs in foreign policy.[48]

Although there are any number of dispositions relevant to the study of reputation costs in foreign policy crises, we focus predominantly here on *militant assertiveness*, the ubiquitous division between hawks and doves that plays a prominent role in almost every model of the structure of American foreign policy attitudes.[49] Hawks and doves differ from one another in their beliefs about the motivations of adversaries in the international system and thus their assessments about the wisdom and efficacy of the use of force. We argue that these differences cause them to weight different types of reputations differently, and ultimately assess reputation costs in fundamentally different ways. Hawks tend to embrace "deterrence mindsets" that understand adversaries as being motivated by expansionist goals, thereby placing great weight on demonstrating resolve and carrying out assertive foreign policy postures in order to dissuade others from taking advantage of them.[50] Hawks thus think about reputation predominantly through the prism of resolve, whereby acting inconsistently and backing down from a threat is viewed as undermining resolve and generating reputation costs. Hawks only care about "reputation for honesty," for example inasmuch as it implicates the credibility of one's threats; when Schelling claimed that reputation is "one of the few things worth fighting over," he was not talking about, say, a reputation for compliance with international law.[51] Doves, on the other hand, subscribe to the "spiral model," where conflict is often the result of misperceptions.[52] They thus perceive making threats and intervening in the affairs of other countries as costly and counterproductive, undermining opportunities for cooperation that would otherwise take place. Belligerence thus is understood as having negative reputational consequences by causing others to see a hypocritical disconnect between the liberal values we preach and the foreign policies we practice.[53] Hawkish and dovish segments of the public should therefore perceive substantially different reputational costs from one another, even in response to the exact same strategic situation.

It is worth noting how our theory differs from how reputation is often treated in IR. First, for entirely reasonable analytic purposes, classic models of reputation costs treat reputations as unidimensional by focusing on one trait at a time. Reputation is frequently modeled as "a probability distribution … regarding the likelihood that

47. Brewer 2001; Druckman 2001.

48. We bracket the question of the origins of these foreign policy orientations, which may be a product of personal values, a response to discrete psychological needs, or a function of our genetic makeup. See Jost et al. 2007; McDermott et al. 2009; Rathbun et al. 2016.

49. For example, Chittick, Billingsley, and Travis 1995; Herrmann, Tetlock, and Visser 1999; Holsti 2004; Hurwitz and Peffley 1987; Modigliani 1972; Wittkopf 1990.

50. Glaser 1992; Jervis 1976; Tetlock 1983.

51. Sartori 2005; Schelling 1966.

52. Jervis 1976.

53. Finnemore 2009; Kertzer and Brutger 2016.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

[an] agent is of one type rather than another," with the crisis bargaining literature frequently treating "reputation" as shorthand for "reputation for resolve."[54] In contrast, our theory understands reputation as multidimensional. Assessing reputation costs involves aggregating across multiple types of reputation simultaneously, forcing our attention on important questions about how this aggregation process takes place.

Second, the IR literature largely tends to study reputation situationally, pointing to how features of the strategic environment—such as the types of actions leaders pursue, or the number of opportunities for future interaction—incur reputation costs, so that it is assumed that different types of observers will reach identical conclusions from looking at the same behavior.[55] Huth notes that our theories of reputation in IR tend to assume: "1) The cognitive processes by which reputational inferences are drawn are similar across individuals; and 2) that the content of whatever inferences are drawn are also similar across individuals."[56] Similarly, Mercer notes that deterrence theory tends to assume that "everyone will view our behavior similarly."[57] This is largely the case because IR scholars frequently model reputation in well-defined strategic environments, whether in the form of formal models that assume the aggregation question away, or in stylized laboratory experiments where the situation is already defined for participants by the experimenter. In contrast, we suggest that the ill-defined nature of the strategic environment in many foreign policy crisis situations means that assessing reputation costs has an important dispositional component, such that hawks and doves will assess reputation costs differently. This also differentiates our approach from other work in IR that incorporates hawkishness into discussions of reputation, like Diesing, who assume that hawks or "hardliners" tend to care about reputation costs, while doves do not.[58] In contrast, we argue that both hawks and doves care about reputation—they just aggregate reputational considerations in different ways.

Third, classic approaches to reputation in IR often treat reputational considerations at stake as distinct from an actor's intrinsic interests. Jervis notes that when actors back down in a foreign policy crisis, they sacrifice both *intrinsic interests*—the tangible objects at stake in the conflict, and *strategic interests* or reputation costs.[59] Many of the fiercest debates in deterrence theory revolve around the relative important of each type of factor; for Hopf and Press, for example, intrinsic interests trump past actions so that reputation costs are rarely as costly as we think.[60] Thinking about reputation costs as having an important dispositional component calls this dichotomy into question. If the same dispositions that determine how individuals assess the

54. Bar-Isaac and Deb 2014, 44.
55. Mattes 2012; Sechser 2010; Walter 2006; Weisiger and Yarhi-Milo 2015.
56. Huth 1997, 79.
57. Mercer 1996, 7.
58. Snyder and Diesing 1977.
59. Jervis 1976.
60. Clare and Danilovic 2012; Hopf 1991; Huth 1999; Schelling 1960; Press 2005. Critiques of audience cost theory similarly play up the distinction: for Snyder and Borghard 2011 and Chaudoin 2014, for example, the public's substantive policy preferences trump reputational concerns about consistency.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

intrinsic merit of policies also affect their estimation of the policies' reputational consequences, intrinsic and strategic interests are intertwined.

Although our theory of the microfoundations of reputation costs perhaps draws on more explicitly psychological roots than some treatments of reputation in IR, it is not an "irrational" theory of reputation; psychology and rationality are not opposites.[61] If rational choice is about the study of choice under constraint, our interest here is in how individuals constrained by ill-defined strategic situations rely on prior beliefs to assess reputation costs.[62]

Methodology

To examine how domestic audiences think about reputation costs in foreign policy, we fielded two online survey experiments in the summers of 2013 and 2014. The main experiment we focus on in our analyses was fielded in the summer of 2014 on a sample of 589 registered American voters recruited by Survey Sampling International (SSI).[63] We turn to survey experiments for two reasons. First is identification: reputation costs are difficult to study in an uncontrolled environment, since strategic leaders should shy away from incurring reputation costs in the first place. The experiment allows us to manipulate the strategic behavior of the president while holding other things constant, and test how participants assess the reputational consequences. Second is measurement: if we are interested in testing a dispositional theory of reputation costs, the suite of survey methods we employ here—from clean measures of hawkishness borrowed from political psychologists to identify our hawks and doves, to automated text analysis to show how they talk about reputation differently—is uniquely well-suited to answering the question we are interested in.

The experiment began by presenting respondents with an introduction:

> The following questions are about US relations with other countries around the world. You will read about a situation our country has faced many times in the past and will probably face again. Different leaders have handled the situation in different ways. We will describe one approach US leaders have taken, and ask whether you approve or disapprove.

Respondents then read about a hypothetical international crisis where a foreign country sent its military to invade a neighboring country.[64] After reading that "A

61. McDermott 2004; Mercer 2005; Rathbun, Kertzer, and Paradis 2017.

62. Snidal 2002, 74.

63. See Appendix §2 for sample characteristics and more details about the survey and sampling methodology.

64. Consistent with Herrmann, Tetlock, and Visser 1999, Tomz 2007a, and others, we chose not to identify the foreign countries by name to avoid confounding the scenario with country-specific biases, and speak directly to the existing literature. The results are also robust when additional contextual information is provided to respondents, a point we discuss further in Appendix §4.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

country sent its military to take over a territorial region in a neighboring country," respondents were randomly assigned to one of three experimental conditions varying the president's behavior. In the "Stay Out" condition, respondents read that the president announced "the United States would stay out of the conflict," whereupon the attacking country continued to invade. In the "Not Engage" condition, respondents read that the president announced that "if the attacking country continued to invade, the United States military would immediately engage and attempt to push out the attacking country"; the attacking country continued to invade, but the president did not engage. In this sense, the experimental setup is similar to those used in classic audience cost experiments, which are also interested in how domestic publics respond to presidents who back down after making a threat.[65] However, as Kertzer and Brutger show, these designs confound punishing the president for backing down with punishing the president for threatening to get involved in the first place.[66] We therefore use a third experimental condition to disaggregate these two treatments: an "Engage" condition, in which the president threatens force (announcing that "if the attacking country continued to invade, the United States military would immediately engage and attempt to push out the attacking country"), and then follows through on its threat and orders the US military to engage. Comparing the "Engage" treatment to the "Stay Out" treatment allows us to hold constant the effects of inconsistency while varying whether or not the president threatened force. In contrast, comparing the "Engage" and "Not Engage" conditions holds constant the president's threat of force but varies whether or not the president acted inconsistently. The experiment thus allows us to test how participants assess the reputational consequences of inconsistency and belligerence in foreign policy crises.[67]

Following the scenario, we administered our dependent variable, asking participants to assess the reputation costs borne by both the president and the country as a whole ("On a scale of 1–5, how much damage do you think there would be to the President's [America's] reputation as a result of the President's handling of the situation?"), on a response scale ranging from "No damage" to "A lot of damage" such that higher values indicate steeper reputation costs. We chose to ask respondents about both the leader's and the country's reputation since the IR literature is divided about whether reputations are leader specific, country specific, or linked.[68] While our theory is agnostic about the relative importance of leader and country-specific reputations, we find nearly identical results for both leader- and country-specific

65. For example, Tomz 2007a; Trager and Vavreck 2011; Levendusky and Horowitz 2012.
66. Kertzer and Brutger 2016.
67. For a complete discussion of disaggregating inconsistency and belligerence costs, see Kertzer and Brutger 2016. To isolate the treatment effects of inconsistency and belligerence, the experiment holds outcomes constant with the hostile foreign country taking control of 20 percent of the contested territory and no casualties being incurred. This ensures that the treatment effects are not motivated by varying policy outcomes and instead are limited to the president's actions in the crisis.
68. Brewster 2009; Guisinger and Smith 2002; Jervis 1982; Renshon, Dafoe, and Huth 2018.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

reputations, a point we discuss further in Appendix §5. In addition to measuring reputation costs, the study also included a standard battery of demographic characteristics, and a measure of militant assertiveness (or hawkishness) modified from Herrmann, Tetlock, and Visser.[69]

Results

Our analysis follows three steps. First, we present our average treatment effects, showing how our participants assessed reputation costs in response to belligerence and inconsistency; we then add our measure of militant assertiveness and a number of other individual-level variables to study the dispositional predictors of reputation costs. Second, we show how hawks and doves assess reputation costs in fundamentally different ways, both by presenting evidence of heterogeneous treatment effects, and using structural topic models to show that these differences are mirrored in the language hawks and doves use when they talk about reputation. Third, having explored the causes and character of reputation costs, we turn to their consequences, showing our participants punish leaders who suffer reputation costs—they just disagree on the policies that produced them.

Assessing Reputation Costs

We begin by estimating the average treatment effects, showing how the strategy the president employed—either belligerence, or inconsistency—affected the average leader-level and country-level reputation costs provided by our participants.[70] We find that inconsistency on the part of the president generates a significant reputation cost. The cost to the president's reputation is 0.755 ($p < 0.00$) and the country's reputation is 0.711 ($p < 0.00$), representing a 19 percentage point reputation cost for the president and a 18 percentage point reputation cost for the country.[71] In contrast to the strong inconsistency effect, we do not find that belligerence on average generates significant reputation costs for the president or the country, with the average treatment effects being −0.099 ($p < 0.188$) and −0.029 ($p < 0.394$) respectively, although as we suggest, these average treatment effects belie considerable heterogeneity across respondents.

To assess the role of individual-level characteristics on perceptions of reputation costs, we also model the generation of reputation costs another way. Since the two reputation cost variables—the president's and country's reputation—are clearly correlated with one another ($r = 0.86$), rather than estimate separate regression models

69. Herrmann, Tetlock, and Visser 1999. See Appendix §1 for our dispositional instrumentation.

70. The sample was well balanced across treatment groups, as shown in Appendix §3.

71. We use bootstrapped treatment effects as our quantities of interest, with 1,500 draws with replacement.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

assuming the errors are uncorrelated, we more efficiently estimate a seemingly unrelated regression (SUR) model in Table 1. The SUR not only allows us to estimate our treatment effects on our two dependent variables simultaneously, but allows us to do so while controlling for a host of dispositional characteristics and demographic variables, giving us insight into the role that observer-level characteristics play in shaping perceptions of reputation costs.

As before, we see that on average, inconsistency significantly increases perceived reputational damages to both the country and the president (with effect sizes of 0.69 and 0.74 on a five-point scale, respectively). Yet we also see that even controlling for the president's strategic choices, certain types of people are systematically more likely to perceive the existence of reputation costs than others. Respondents higher in militant assertiveness, with more hawkish foreign policy preferences, generally perceive steeper reputation costs than individuals with more dovish foreign policy preferences, while individuals higher in international trust are significantly less likely to perceive reputation costs than their less-trusting counterparts. Higher levels of income are positively associated with the magnitude of reputation costs for the president, but not for the country as a whole. In general, then, these results

TABLE 1. *Modeling perceptions of reputation costs*

	Perceived reputation cost to:	
	Country	*President*
INTERCEPT	2.69***	2.561***
	(0.232)	(0.231)
BELLIGERENCE	−0.024	−0.09
	(0.118)	(0.118)
INCONSISTENCY	0.69***	0.739***
	(0.117)	(0.117)
MILITANT ASSERTIVENESS	0.597**	0.652***
	(0.246)	(0.246)
INTERNATIONAL TRUST	−0.262**	−0.254**
	(0.11)	(0.11)
NATIONAL CHAUVINISM	−0.236	−0.382
	(0.257)	(0.257)
DEMOCRAT	−0.142	−0.04
	(0.107)	(0.107)
REPUBLICAN	0.136	0.201*
	(0.119)	(0.119)
EDUCATION	−0.028	−0.02
	(0.051)	(0.051)
INCOME	0.04	0.064**
	(0.03)	(0.03)
MALE	0.086	0.114
	(0.091)	(0.091)
N	586	586
Adj. R^2	0.119	0.127

Notes: Coefficient estimates from a seemingly unrelated regression model. Reference category for partisanship: independents. *$p < .1$; **$p < .05$; ***$p < .01$.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

tell us that there remains important variation in assessments of reputation costs even when controlling for the president's behavior on the world stage, and that we significantly improve our explanatory power when we take these dispositional characteristics into account.[72]

How Do Hawks and Doves Understand Reputation Costs?

Heterogeneous Treatment Effects

If there is an important dispositional element to reputation costs, however, we would also expect it to manifest itself in terms of heterogeneous treatment effects: because they rely on their predispositions toward foreign affairs when aggregating reputational considerations, hawks and doves should offer significantly different reputational assessments in response to the exact same foreign policy behavior.

We thus split our sample into hawks and doves, plotting the bootstrapped distributions of reputation cost treatment effects in each subsample in Figure 1, along with the full sample results for purposes of comparison.[73] Hawks perceive significant reputation costs to leaders who act inconsistently, but believe the president and country's reputations are improved when the leader initiates threats, consistent with models that argue leaders can build stronger reputations by initiating conflicts and proactively building their reputation.[74] Doves, however, disagree: in contrast to those who believe that "threats have the virtue of being free when they succeed," doves believe leaders' reputations are damaged when they initiate threats, regardless of the subsequent outcome.[75] Doves thus weight these reputational considerations in a systematically different fashion than hawks, perceiving both belligerence and inconsistency as reputationally costly. The results are consistent with a dispositional theory of reputation in which individuals draw on their pre-existing foreign policy orientations to make reputational judgments. The divergence between doves and hawks suggests limitations on the extent to which domestic audiences agree about reputation costs, reminding us that audiences' perceptions of reputation costs are closely tied to their foreign policy orientations and policy preferences, and that leaders have to manage their reputations among distinct domestic audiences with contrasting views of reputation.

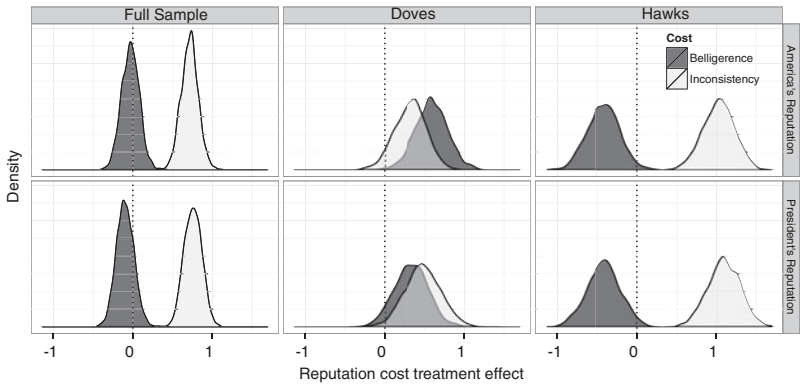
72. A pair of Wald tests shows that incorporating the individual-level characteristics significantly improves model fit (for America's reputation: $F = 3.792$, $p < 0.001$; for the president's reputation: $F = 4.213$, $p < 0.000$).

73. We define hawks as those participants who score in the top quartile for militant assertiveness and doves as those individuals in the bottom quartile for militant assertiveness.

74. Clare and Danilovic 2010.

75. Dafoe, Renshon, and Huth 2014, 378.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press



Note: Figure 1 shows the reputation cost (where higher values indicates steeper costs) in response to both inconsistency and belligerence. These density plots show that hawks and doves believe reputations are damaged for different reasons, showing a strong connection between foreign policy orientation and beliefs about reputation.

FIGURE 1. *Hawks and doves assess reputation costs in significantly different ways*

How Do Hawks and Doves Talk About Reputation?

To further probe why doves and hawks perceive reputation differently, we look to how doves and hawks differ in their discussion of reputation and perceptions of reputational damage. Since our primary experiment does not include open-ended response questions, we turn to a supplementary experiment that follows the same overarching structure. This supplementary experiment was conducted using a sample of 1,173 respondents recruited using Amazon Mechanical Turk in the spring of 2013. Following their evaluation of how much damage respondents believed there to be to the country and president's reputation, respondents were then asked to "Please briefly explain your response about the president's [or country's] reputation." We analyze these open-ended responses using a structural topic model (STM), a semi-automated text-analysis model drawing on advances in machine learning, to test whether the language our respondents use when they talk about reputation costs in our open-ended responses systematically varies with their predispositions toward militant assertiveness.[76] Since reputations are multidimensional, we expect that respondents will invoke a variety of types of reputational concerns when justifying their reputational assessments, but in line with our dispositional theory, we expect these considerations to arise in varying frequency among hawks and doves, such that the two groups will talk about reputation in fundamentally different ways.

Although the technical details of STMs are beyond the scope of this paper, one can think of them as an automated text-analysis technique that models text as a mixture of

76. See Roberts et al. 2014 for an accessible introduction. For an application, see Tingley 2017.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

semantically interpretable "topics." Because STMs are unsupervised, they allow researchers to "discover" topics in the text rather than assume their existence in advance, as would be the case in a supervised text analysis method in which our prior expectations about why doves and hawks have different reputational concerns might unconsciously drive our findings. Unlike with more traditional mixed-membership models used in text analysis like latent Dirichelet allocation (LDA), STMs let us leverage information about respondents (in our case, the participant's level of militant assertiveness, and the treatment to which the participant was randomly assigned in the experiment) when structuring the topics, rather than assuming that the content and prevalence of topics are constant across all of our participants in the study. In this way, STMs offer a statistical framework in which researchers can study discourse as a dependent variable.

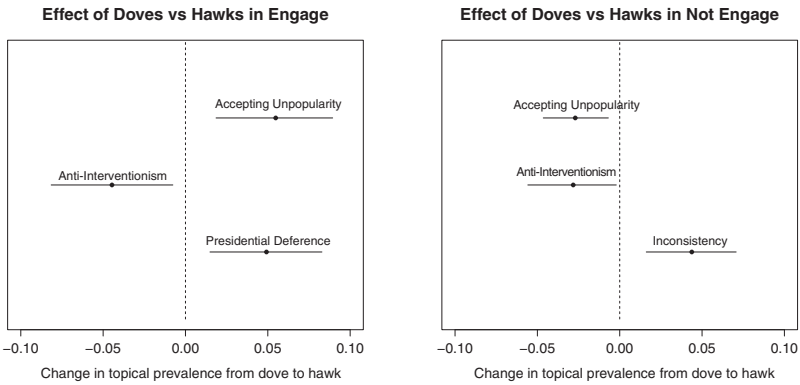
We therefore estimate a set of STMs on our participants' responses to the open-ended questions about reputation costs, using the models to identify whether certain topics are more likely to appear in the responses of doves and hawks, conditional on the assigned treatment in the experiment.[77] We look at the results in two steps. First, Figure 2 tells us how doves and hawks assigned to the same treatment condition differ in the language they use when discussing reputation costs, as measured by changes in topical prevalence. Second, to derive the labels for each topic and better illustrate what they mean substantively, Figure 3 presents a set of representative responses from participants.

For the Engage condition, the results in the left-hand panel of Figure 2 tell us that doves and hawks significantly differ from one another on three substantively clear topics, representative responses for which are shown on the left panel of Figure 3.[78] First, when the president threatens force and follows through on it, hawks are significantly more likely to make comments regarding the inevitability of unpopularity. "When you're the parent," one respondent writes, "there is always someone who won't like you." "The president needs to make tough decisions no matter what effect it will have [on] his reputation," another notes. These comments, which implicitly downplay the importance of reputation costs when threats are made, are significantly more likely to be made by hawks than by doves. Doves, on the other hand, are significantly more likely to respond to the president's use of force by discussing reputation costs in terms of anti-interventionism: "The American reputation is that of a bully and the result is evidence of that," one respondent writes. Against outward-looking understandings of the American national interest exemplified by

77. All of the statistical analysis for the topic models was conducted using the STM package in R. Given the extremely high correlation between reputation costs attached to America and reputation costs attached to the president, we pool these responses here for each respondent, to increase the amount of text the STM has to work with. We also present separate results for both America's and the president's reputation in Appendix §6 and find that similar topics are significant regardless of how the results are analyzed. For model selection, we employ the manyTopics function, which indicates a model with thirteen topics is appropriate, based on exclusivity.

78. For a list of the highest probability words for each topic, see Appendix §6.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press



Notes: This figure shows the change in the topic proportion when shifting from doves to hawks. Estimates are made by incorporating estimation uncertainty for the topic proportions into the uncertainty estimates. Only topics with a statistically significant difference between groups are displayed. Thus, in the Engage condition, for example, hawks are significantly less likely to make anti-interventionist arguments about reputation costs than doves are. In the Not Engage condition, a fourth topic differed in prevalence between hawks and doves, but it lacked a clear substantive interpretation and is thus omitted from the main analysis, consistent with norms in the text analysis literature. See Roberts, Stewart, and Tingley 2014.

FIGURE 2. *Difference in topic proportion between doves and hawks*

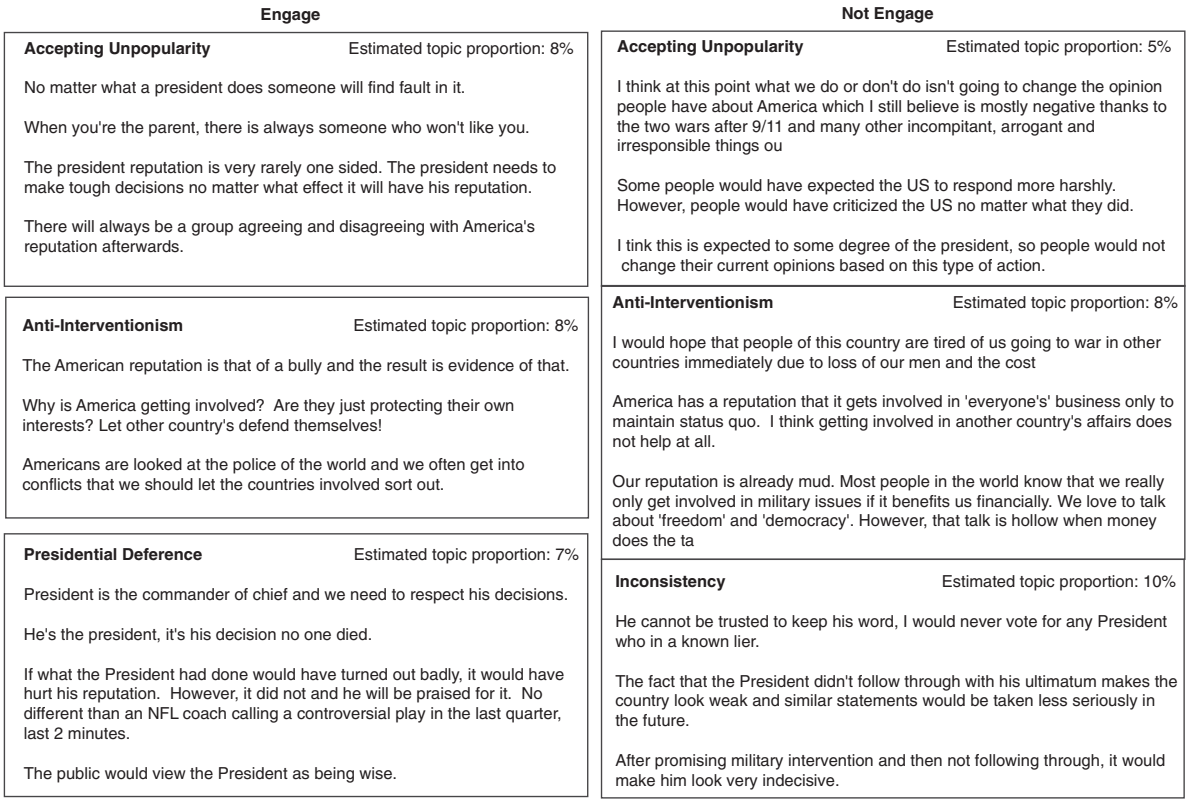
what Mead called the "Hamiltonian" and "Wilsonian" schools of American foreign policy, doves are more likely to denounce America's role as global police: "Why is America getting involved?" one respondent laments. "Let other [countries] defend themselves!"[79] Third, hawks are significantly more likely than doves to couch their discussions of reputation costs in terms of deference to the president, either pointing to the president's wisdom in producing a good foreign policy outcome ("He's the president, it's his decision, no one died"), or generally accepting the president's judgment ("President is the commander [in] chief and we need to respect his decisions"). These types of comments, with their emphasis on deference to authority, are consistent both with the "rally around the flag" phenomenon, and the general extent to which hawkish beliefs are associated with a greater emphasis on deference to authority and the entrenchment of hierarchy.[80]

For the Not Engage condition, hawks and doves significantly differ from one another on four topics, three of which have a clear substantive interpretation, as presented in the right-hand panel of Figure 3. Whereas hawks were more likely to accept unpopularity or downplay reputation costs when the president makes a threat and follows through, the right-hand side of Figure 3 shows that when the president makes a threat and backs down, doves are more likely to make these arguments. "Some people would have expected the US to respond more harshly. However, people would have criticized the US no matter what they did," one respondent

79. Mead 2002.
80. Mueller 1973; Kertzer et al. 2014.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press



Note: This figure shows representative open-ended responses for each topic, selected for coherence among the ten most highly associated responses for each topic estimated using findThoughts from the STM package. Some responses are truncated by the statistical software. See Roberts et al. 2014.

FIGURE 3. *Representative responses for each topic*

writes. Another offers a more jaded version of this argument, noting that America's reputation has already been damaged by the belligerence of American foreign policy post-9/11: "I think at this point what we do or don't do isn't going to change the opinion people have about America." Nonetheless, the parallel with hawks in the Engage condition is striking. We see further similarities with regard to anti-interventionism. Just as in the Engage condition, doves are significantly more likely to discuss reputational costs using anti-interventionist rhetoric than hawks are. "America has a reputation that it gets involved in everyone's business only to maintain [the] status quo," one respondent writes. "I think getting involved in another country's affairs does not help at all." Finally, and consistent with traditional understandings of the key mechanism behind audience cost theory, the third topic that significantly differs between doves and hawks involves inconsistency. Hawks are significantly more likely to discuss reputation costs in terms of expressing dissatisfaction for the president saying one thing and doing another than doves are. "He cannot be trusted to keep his word, I would never vote for any President who is a known [liar]," complains one participant. The national honor is also implicated: "The fact that the President didn't follow through with his ultimatum makes the country look weak and similar statements would be taken less seriously in the future," worries another.

The value of structural topic modeling is that it is purely inductive: unlike with supervised text analysis methods, our prior expectations about how doves and hawks aggregate reputational considerations cannot steer our findings. Thus, the fact that we uncover patterns remarkably consistent with our theoretical priors about how doves and hawks assess reputation—with doves placing significant weight on the negative reputational consequences of intervention, while hawks place greater weight on the negative repuational consequences of inconsistency—is all the more striking. Interestingly, both hawks and doves demonstrate a willingness to discount potential reputation costs when the costs are produced by a policy consistent with their underlying foreign policy orientation. When the president makes a threat, consistent with hawkish predispositions, hawks indicate a willingness to accept unpopularity. Similarly, doves are also willing to discount potential reputation costs, but when the president backs away from a threat and pursues a more peaceful strategy. This pattern—consistent with motivated reasoning—thus calls the exogeneity of strategic and intrinsic interests into question: if the same dispositions that shape how we evaluate the intrinsic value of policies also shape how we estimate their reputational consequences, the conceptual distinctions between how much we value a policy intrinsically and what we think the reputational consequences of that policy will be begin to blur. This point raises some interesting implications for IR theory, which we return to in the conclusion.

Consequences of Reputation Costs

Thus far we have shown that ordinary citizens have a "taste" for a good reputation in foreign policy—a crucial exercise given the growth of theoretical models in IR in the

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

past several decades that use public opinion as the channel through which reputation costs shape state behavior. While fully validating the next link in the model—that public opinion shapes state behavior—is well beyond the scope of this paper, our study lets us continue down the causal chain by examining the relationship between reputation costs and presidential approval, which, following Gottfried and Trager, we argue constitutes a crucial component of democratic leaders' incentive structures.[81] Examining the downstream effects of reputation costs on public approval is also valuable given the particular popularity of audience cost models, which generally assume a reputational mechanism to be driving domestic punishment but rarely test it directly.[82]

In addition to being asked to assess the reputational consequences of the president's behavior in the experimental scenario, participants were also asked to assess how strongly they approved or disapproved of the president's handling of the situation, on a seven-point scale.[83] The inclusion of an approval measure allows us to examine how concerns about reputation mediate participants' approval of the president's handling of the situation. We thus estimate a series of nonparametric mediation models testing whether the effects of inconsistency or belligerence on public support is mediated through reputational assessments, and, through moderated mediation tests, whether this effect varies between doves and hawks.[84]

We begin by displaying the results of the mediation analysis for hawks and doves, separately, in Figure 4. The figure reconfirms that both hawks and doves care about reputation—they just think about it in different ways. Among doves, the average causal mediation effect (ACME) through reputation costs on our approval measure is negative both for belligerence (ACME: -0.41; $p < 0.02$) and inconsistency (ACME: -0.39; $p < 0.11$): both belligerence and inconsistency increase reputation costs, which decrease public support. In contrast, the mediation effect through reputation costs for hawks varies significantly based on the treatment. As with the case of doves, inconsistency displays a strongly negative mediation effect through reputation (ACME: -1.68; $p < 0.01$), although the magnitude of the effect is much greater for hawks than for doves. What is striking however, is that belligerence displays a significantly *positive* mediation effect through reputation in the belligerence treatment (ACME: 0.40; $p < 0.08$). Because hawks view reputation as being strengthened when the leader acts belligerent, reputation significantly improves approval for the president's handling of the crisis, which bolsters public support.

We can test these differences between hawks and doves more formally using a moderated mediation model in which the relationship between the treatments and

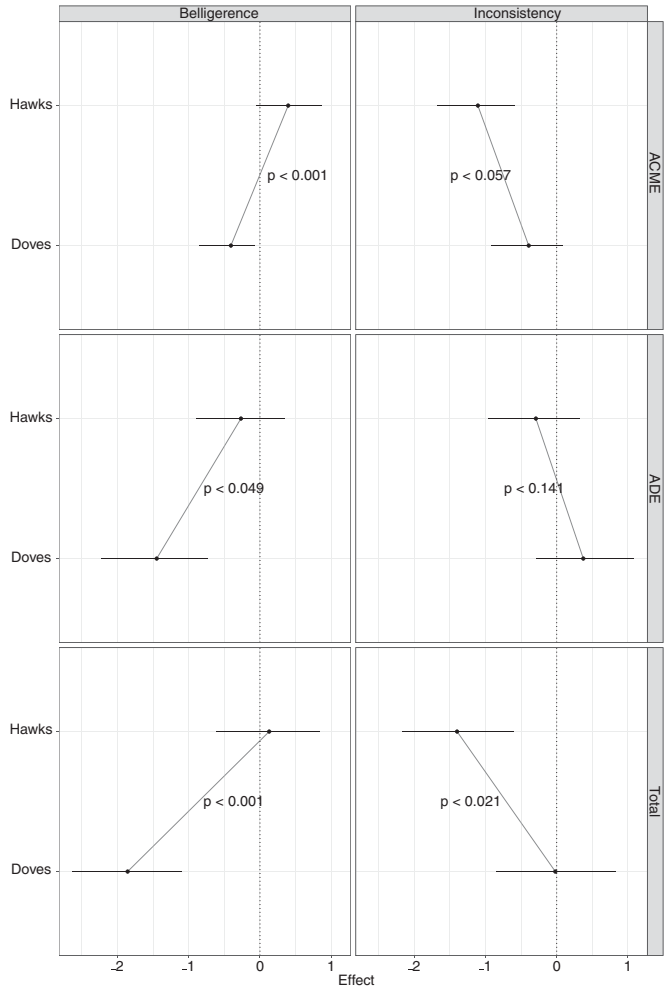
81. Baum and Potter 2015; Gottfried and Trager 2016; Kertzer 2016, 50–51.

82. Both Tomz 2007a and Levendusky and Horowitz 2012 find that many respondents use reputation as an ex post justification for their decision to disapprove of the president's actions; our interest here is in capturing this micro-level mechanism ex ante.

83. Tomz 2007a.

84. Imai et al. 2011. All models control for pre-treatment demographic characteristics; complete mediation results are presented in Appendix §7.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press



Notes: Figure plots the Average Causal Mediation Effects (ACME), the Average Direct Effects (ADE) and Total Effects from a series of nonparametric mediation models in which the effect of each treatment on approval of the president's handling of the crisis is mediated through reputation costs to the country, for hawks, and doves, respectively. Presidential approval is measured on a seven-point scale; hawks and doves are defined as those scoring in the bottom and top quartile of militant assertiveness respectively. The models control for pre-treatment measures of respondents' education, income, gender, and political ideology. The *p*-values shown on the ACME and ADE panels come from formal tests of a moderated mediation model where the effect of the treatment on the reputational mediator varies between hawks and doves; the *p*-values on the total effects panels are estimated using the interaction term between the treatment and the hawkishness dummy variable in a regression model of presidential approval on the treatment, military assertiveness, their interaction, and the pretreatment covariates described earlier. Analysis conducted using the mediation package by Tingley et al. (2014). See Appendix §7 for similar results using the president reputation as a mediator instead, or for moderated mediation tests using a continuous measure of the moderator.

FIGURE 4. *Reputational consequences of belligerence and inconsistency on presidential approval*

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

the mediator is permitted to vary with levels of hawkishness. As the p-values in Figure 4 show, the 1.02 unit difference in ACMEs for belligerence between doves and hawks is highly significant ($p < 0.001$).[85] The moderated mediation results for inconsistency also show notable differences between hawks and doves, with the negative ACME being −0.71 ($p < 0.057$) units greater for hawks than for doves.[86] These findings thus lend further support to the role of reputation as a micro-level mechanism underpinning the relationship between foreign policy behavior in crises and changes in public support, whether in an audience cost framework or in models of public opinion in foreign policy more broadly. The divergent role of reputation between doves and hawks raises new questions for how to think about the domestic politics of reputation in IR, which we turn to in the conclusion.

Conclusion

One of the striking patterns in the history of postwar American foreign relations has been the extent to which policymakers have been preoccupied with reputation, consistently invoking reputational logics when justifying their preferred foreign policy choices.[87] Morgan notes that this fixation on reputation was far more pronounced in the United States than in the Soviet Union, suggesting that it stems partially from the character of American domestic politics.[88] The fact that presidents, generals, and senators employ reputational arguments as frequently when they're speaking to the public as they do behind closed doors demonstrates the extent of the belief that reputational consequences count as "publicly acceptable reasons"; otherwise, decision makers would use these arguments with each other, but not with the rest of us. Yet despite the prevailing assumption that domestic publics care about reputation costs, we have very little direct evidence that publics care about reputation, and very little understanding of how.

To address these lacunae, we propose a dispositional theory of reputation costs in which citizens facing an ill-defined strategic situation turn to their core orientations about foreign affairs to help them weight multiple strategic considerations. We then tested this framework using a diverse array of methodological tools—from vignette-based survey experiments to automated text analysis—to gain a richer understanding of the microfoundations of reputation costs.

Some of our findings offer empirical evidence in favor of political scientists' theoretical predictions: for example, leaders who say one thing and do another incur reputation costs, which spill over into evaluations of the president's handling of foreign

85. The difference increases to 1.58 ($p < 0.001$) when using a continuous measure of militant assertiveness, as opposed to the quartile-based one used in the main analysis.

86. The difference increases to −1.58 ($p < 0.002$) when using the continuous measure of militant assertiveness.

87. McMahon 1991; Tang 2005.

88. Morgan 1985, 142, 148–49.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

policy—a causal chain long assumed by audience cost scholars, though never tested formally. Given the large shadow that reputation costs cast in our experimental results, the frequency with which reputational rhetoric is employed by any decision maker with a microphone should not be surprising.

Other parts of our findings, though, encourage us to push in new directions: although ordinary citizens in the United States indeed care or otherwise have preferences about reputation, they do so in a way that violates the assumptions of our canonical models. Whereas IR scholars have tended to model reputation as unidimensional for analytic purposes, we show that its multidimensionality raises theoretically important questions about how actors aggregate multiple reputational considerations simultaneously. Whereas IR scholars have tended to study reputation situationally, we show how this multidimensionality leads it to have an important dispositional component, as the resonance of reputational rhetoric belies the heterogeneity with which reputation is perceived.

In particular, we find that audiences' perceptions of reputation costs are closely tied with their pre-existing foreign policy orientations. Certain kinds of people—especially citizens high in militant assertiveness—are more likely to perceive reputation costs regardless of what the president does. Moreover, hawks and doves tend to understand reputation costs systematically differently: hawks predict reputation costs for inconsistency, while doves envision negative reputational consequences for both inconsistency and belligerence. Automated text analysis of participants' open-ended responses suggests that hawks and doves even use systematically different language when they talk about reputation costs: doves more often make anti-interventionist arguments about reputation costs, focusing on the negative reputational consequences that accrue from getting involved in other countries' affairs, while hawks tend to focus on inconsistency. Interestingly, both hawks and doves tend to be dismissive of reputation costs when it involves policies that they approve of (following through on a threat to use of force, for hawks; choosing ultimately not to engage, for doves). These results thus enrich our understanding of the psychology of hawks and doves, illustrating the vital roles that reputational considerations play in both types of mental models. Contrary to what Snyder and Diesing argue, it is not that hawks or "hard-liners" care more about reputation than doves do, but rather, that they understand the phenomenon in different ways.[89] Future research should thus do more to explore the origins of these orientations, as well as explore the conditions in which hawks and doves are willing to revise their second-order beliefs about reputation.

In addition to illustrating the merit of bridging the disconnect between foreign policy analysis and "mainstream" IR,[90] the findings also suggest an underappreciated paradox about deterrence theory. Deterrence theory is often built around a distinction between *intrinsic* and *strategic* interests, the former referring

89. Snyder and Diesing 1977, 188.
90. Kaarbo 2015.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

to how much you value the object or issue in dispute, and the latter reflecting the reputational consequences that retreating on this one issue will bear in the future.[91] One reason that leaders turn to reputational arguments to justify a particular policy choice involves the belief that the policy option is insufficiently attractive without an extrinsic rationale. In some variants of deterrence theory, you make a point of standing firm on issues you care intrinsically little about, precisely to eliminate all doubt about whether you would back down when vital interests are at stake.[92] From a strategic perspective, then, reputational arguments should be deployed at home primarily to persuade those who do not find a particular policy course inherently worthwhile, since those who support a policy on intrinsic grounds would not need reputational rhetoric to convince them of its merits. Yet the different ways doves and hawks think about reputation costs in our experiments raises the possibility that assessments of intrinsic and strategic interests are not as independent of one another as we often think: our participants dismissed reputation costs when it involved a policy they generally approved of, hawks reminding us that "when you're the parent, there is always someone who won't like you" when confronted with potential reputational consequences for using force, and doves noting that "people would have criticized the US no matter what they did" when asked about the potential negative reputational consequences for backing down on a threat. Disentangling the causal directionality here is beyond the domain of this paper, but it is worth noting that our pattern of results is entirely consistent with a process in which, rather than reputational considerations bolstering the degree of support for a policy, pre-existing levels of support for a policy shape perceptions of reputation costs. Scholars of reputation have long noticed the shadowboxing-like quality of policymakers' reputational rhetoric: the mismatch between the inferences policymakers assume foreign adversaries will make versus the lessons they actually take away.[93] If individuals do construct perceptions of reputational considerations backwards, inflating or deflating them based on their perceived desirability of the policy being pursued, it would offer further evidence of what Tang calls the "illogic" of reputation.[94]

Because reputation for resolve is not a valence issue, the results illustrate a complicated incentive structure for leaders engaged in foreign policy crises, given that initiating threats can simultaneously generate reputational costs and rewards among two very different domestic audiences. As economists have noted, in the face of heterogeneous audiences, "identifying a 'good' reputation itself is non-trivial."[95] Given this variation, leaders have an incentive to conduct foreign policy in a manner that balances reputational costs and benefits, which will vary based on the composition of their constituencies. Saunders notes that leaders' military intervention choices are partially a function of their causal

91. For example, Jervis 1979.
92. Ravenal 1982, 28.
93. Hopf 1991; Mercer 1996; Press 2005.
94. Tang 2005.
95. Bar-Isaac and Deb 2014, 44.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

beliefs about other states' internal institutions.[96] Our results suggest another relevant factor is leaders' beliefs about the distribution of hawks and doves in *their own* domestic constituency—a question political scientists currently know relatively little about. Scholarship that has evaluated politicians' beliefs about public opinion has largely tended to focus on domestic issues. Existing research suggests that both liberal and conservative politicians often overestimate the degree to which their constituents support conservative policies, which if true in the foreign policy realm would suggest that politicians may overestimate the reputational costs for acting inconsistently in a security crisis and similarly overstate the reputational rewards of making threats and standing firm.[97]

To fully appreciate the domestic politics of reputation, more research is needed to evaluate the extent to which politicians are in tune with their constituents' reputational concerns, and—since leaders face the same aggregation challenges that ordinary citizens do—whether they rely on the same dispositional characteristics to calculate reputation costs at home as they do abroad. Although our focus here has been on the *domestic* politics of reputation costs—and thus, reputation as a second-order belief—there is little in our dispositional theory that precludes its applicability to reputation as a first-order belief. If assessing reputation is an ill-defined problem at home, it is surely just as ill-defined abroad; the dispositions we point to here should presumably be as relevant for foreign audiences as they are for domestic ones. If so, reputation-building strategies are likely to be less straightforward than our existing theoretical models claim because states likely face multiple reputations in the eyes of different beholders simultaneously. Ultimately, our findings paint a more complete picture of how reputation is perceived by domestic audiences and the nuanced incentive structures leaders face.

Finally, our approach here has implications for how we think about the role of complexity in IR. IR scholars are well aware of the challenges that causal complexity poses for our theories of politics.[98] Many of these discussions approach complexity from a philosophical or analytic perspective, as captured in deliberations on the desirability of parsimony as a theoretical goal, or in debates about whether an instrumentalist philosophy of science—in which models are evaluated based on their utility rather than the fidelity of their motivating assumptions—is theoretically satisfying.[99] Our interest here, though, is subtly different. The same complexity that challenges us as analysts also confronts the actors whose beliefs and behavior we study. As scholars, we are familiar with the tools we have developed to study a "cloudy" world as if it were "clock-like," but outside of the cognitive tradition in foreign policy analysis, we have spent less time understanding the micro-processes through which the actors we study in IR carry out the same task—which might in turn suggest new ways that we as scholars might do the same.[100]

96. Saunders 2011.
97. Broockman and Skovron 2018; Miller and Stokes 1963.
98. Braumoeller 2003; Jervis 1998; Nelson and Katzenstein 2014.
99. Clarke and Primo 2012; Lake and Powell 1999; Waltz 1979, 7–8; Wight 2006.
100. Almond and Genco 1977; Hafner-Burton et al. 2017; Herrmann, Tetlock, and Visser 1999; Larson 1994; Sylvan and Voss 1998.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

Supplementary Material

Supplementary material for this article is available at <https://doi.org/10.1017/S0020818318000188>.

References

Almond, Gabriel A., and Stephen Genco. 1977. Clouds, Clocks and the Study of Politics. *World Politics* 29 (4):489–522.

Bar-Isaac, Heski, and Joyee Deb. 2014. What is a Good Reputation? Career Concerns with Heterogeneous Audiences. *International Journal of Industrial Organization* 34 (1):44–50.

Baum, Matthew A., and Philip B.K. Potter. 2015. *War and Democratic Constraint: How The Public Inuences Foreign Policy*. Princeton, NJ: Princeton University Press.

Becker, Gary S. 1974. A Theory of Social Interactions. *Journal of Political Economy* 82 (6):1063–93.

Berinsky, Adam J. 2009. *In Time of War: Understanding American Public Opinion from World War II to Iraq*. Chicago: University of Chicago Press.

Blyth, Mark. 2003. Structures Do Not Come with an Instruction Sheet: Interests, Ideas, and Progress in Political Science. *Perspectives on Politics* 1 (4):695–706.

Braumoeller, Bear F. 2003. Causal Complexity and the Study of Politics. *Political Analysis* 11 (3):209–33.

Brewer, Paul R. 2001. Value Words and Lizard Brains: Do Citizens Deliberate About Appeals to Their Core Values? *Political Psychology* 22 (1):45–64.

Brewster, Rachel. 2009. Unpacking the State's Reputation. *Harvard International Law Journal* 50 (2):231–69.

———. 2011. The Limits of Reputation on Compliance. *International Theory* 1 (2):323–33.

Broockman, David E., and Christopher Skovron. 2018. Bias in Perceptions of Public Opinion Among American Political Elites. *American Political Science Review*. Published online 22 March <https://doi.org/10.1017/S0003055418000011>.

Brutger, Ryan. 2017. Domestic Politics of International Compromise. Working paper.

Carson, Austin. 2016. Facing Off and Saving Face: Covert Intervention and Escalation Management in the Korean War. *International Organization* 70 (1):103–31.

Chaudoin, Stephen. 2014. Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions. *International Organization* 68 (1):235–56.

Chittick, William O., Keith R. Billingsley, and Rick Travis. 1995. A Three-Dimensional Model of American Foreign Policy Beliefs. *International Studies Quarterly* 39 (3):313–31.

Clare, Joe, and Vesna Danilovic. 2010. Multiple Audiences and Reputation Building in International Conflicts. *Journal of Conflict Resolution* 54 (6):860–82.

———. 2012. Reputation for Resolve, Interests, and Conflict. *Conflict Management and Peace Science* 29 (1):3–27.

Clarke, Kevin A., and David M. Primo. 2012. *A Model Discipline: Political Science and the Logic of Representations*. New York: Oxford University Press.

Clinton, William J. 1993. Address to the Nation on Somalia, 7 October. Available at <http://www.presidency.ucsb.edu/ws/?pid=47180>.

———. 1994. Address to the Nation on Haiti, 15 September. Available at <http://www.presidency.ucsb.edu/ws/?pid=49093>.

Crescenzi, Mark J.C. 2007. Reputation and Interstate Conflict. *American Journal of Political Science* 51 (2):382–96.

Crescenzi, Mark J.C., Jacob D. Kathman, Katja B. Kleinberg, and Reed M. Wood. 2012. Reliability, Reputation, and Alliance Formation. *International Studies Quarterly* 56 (2):259–74.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

Croco, Sarah E. 2011. The Decider's Dilemma: Leader Culpability, War Outcomes, and Domestic Punishment. *American Political Science Review* 105 (3):457–77.
Dafoe, Allan, Jonathan Renshon, and Paul Huth. 2014. Reputation and Status as Motives for War. *Annual Review of Political Science* 17 (1):371–93.
Davies, Graeme A.M., and Robert Johns. 2013. Audience Costs Among the British Public: The Impact of Escalation, Crisis Type, and Prime Ministerial Rhetoric. *International Studies Quarterly* 57 (4): 725–37.
———. 2016. The Domestic Consequences of International Over-cooperation: An Experimental Study of Microfoundations. *Conflict Management and Peace Science* 33 (4):343–60.
Druckman, James N. 2001. On the Limits of Framing Effects: Who Can Frame? *Journal of Politics* 63 (4): 1041–66.
Fearon, James D. 1994. Domestic Political Audiences and the Escalation of International Disputes. *American Political Science Review* 88 (3):577–92.
Fearon, James, and Alexander Wendt. 2002. Rationalism v. Constructivism: A Skeptical View. In *Handbook of International Relations*, edited by Walter Carlsnaes, Thomas Risse, and Beth A. Simmons, 52–72. London: Sage.
Finnemore, Martha. 2009. Legitimacy, Hypocrisy, and the Social Structure of Unipolarity: Why Being a Unipole Isn't All It's Cracked Up to Be. *World Politics* 61 (1):58–85.
Fiske, Susan T., and Shelley E. Taylor. 2013. *Social Cognition: From Brains to Culture*. Los Angeles: Sage.
Gelpi, Christopher, and Joseph M. Grieco. 2015. Competency Costs in Foreign Affairs: Presidential Performance in International Conflicts and Domestic Legislative Success, 1953–2001. *American Journal of Political Science* 59 (2):440–56.
Gigerenzer, Gerd. 2008. Why Heuristics Work. *Perspectives on Psychological Science* 3 (1):20–29.
Glaser, Charles L. 1992. Political Consequences of Military Strategy: Expanding and Refining the Spiral and Deterrence Models. *World Politics* 44 (4):497–538.
Goddard, Stacie E., and Ronald R. Krebs. 2015. Rhetoric, Legitimation, and Grand Strategy. *Security Studies* 24 (1):5–36.
Gottfried, Matthew S., and Robert F. Trager. 2016. A Preference for War: How Fairness and Rhetoric Inuence Leadership Incentives in Crises. *International Studies Quarterly* 60 (2):243–57.
Grynaviski, Eric. 2014. *Constructive Illusions: Misperceiving the Origins of International Cooperation*. Ithaca, NY: Cornell University Press.
Guisinger, Alexandra, and Alastair Smith. 2002. Honest Threats: The Interaction of Reputation and Political Institutions in International Crises. *Journal of Conflict Resolution* 46 (2):175–200.
Guzman, Andrew T. 2008. *How International Law Works: A Rational Choice Theory*. New York: Oxford University Press.
Hafner-Burton, Emilie M., Stephan Haggard, David A. Lake, and David G. Victor. 2017. The Behavioral Revolution and the Study of International Relations. *International Organization* 71 (S1):S1–S31.
Hermann, Charles F. 1990. Changing Course: When Governments Choose to Redirect Foreign Policy. *International Studies Quarterly* 34 (1):3–21.
Herrmann, Richard. 1988. The Empirical Challenge of the Cognitive Revolution: A Strategy for the Drawing Inferences About Perceptions. *International Studies Quarterly* 32 (2):175–203.
Herrmann, Richard K., Philip E. Tetlock, and Penny S. Visser. 1999. Mass Public Decisions to Go to War: A Cognitive-Interactionist Framework. *American Political Science Review* 93 (3):553–73.
Hill, Valerie, and Bradford H. Pillow. 2006. Children's Understanding of Reputations. *The Journal of Genetic Psychology: Research and Theory on Human Development* 167 (2):137–57.
Holsti, Ole R. 2004. *Public Opinion and American Foreign Policy*. Ann Arbor, MI: University of Michigan Press.
Hopf, Ted. 1991. Soviet Inferences from Their Victories in the Periphery: Visions of Resistance or Cumulating Gains? In *Dominoes and Bandwagons: Strategic Beliefs and Great Power Competition in the Eurasian Rimland*, edited by Robert Jervis and Jack Snyder, 145–89. New York: Oxford University Press.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

Hurwitz, Jon, and Mark Peffley. 1987. How Are Foreign Policy Attitudes Structured? *American Political Science Review* 81 (4):1099–120.
Huth, Paul K. 1997. Reputations and Deterrence: A Theoretical and Empirical Assessment. *Security Studies* 7 (1):72–99.
———. 1999. Deterrence and International Conflict: Empirical Findings and Theoretical Debates. *Annual Review of Political Science* 2:25–48.
Imai, Kosuke, Luke Keele, Dustin H. Tingley, and Teppei Yamamoto. 2011. Unpacking the Black Box of Causality: Learning About Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review* 105 (4):765–89.
Jervis, Robert. 1976. *Perception and Misperception in International Politics*. Princeton, NJ: Princeton University Press.
———. 1979. Deterrence Theory Revisited. *World Politics* 31(2):289–324.
———. 1982. Deterrence and Perception. *International Security* 7 (3):3–30.
———. 1998. *System Effects: Complexity in Political and Social Life*. Princeton, NJ: Princeton University Press.
Jost, John T., Jaime L. Napier, Hulda Thorisdottir, Samuel D. Gosling, Tibor P. Palfai, and Brian Ostafin. 2007. Are Needs to Manage Uncertainty and Threat Associated with Political Conservatism or Ideological Extremity? *Personality and Social Psychology Bulletin* 33 (7):989–1007.
Kaarbo, Juliet. 2015. A Foreign Policy Analysis Perspective on the Domestic Politics Turn in IR. *International Studies Review* 17 (2):189–216.
Keeney, Ralph L., and Howard Raiffa. 1993. *Decisions with Multilple Objectives: Preferences and Value Tradeoffs*. New York: Cambridge University Press.
Kertzer, Joshua D. 2016. *Resolve in International Politics*. Princeton, NJ: Princeton University Press.
———. 2017a. Microfoundations in International Relations. *Conflict Management and Peace Science* 34 (1):81–97.
———. 2017b. Resolve, Time, and Risk. *International Organization* 71 (S1):S109–36.
Kertzer, Joshua D., and Ryan Brutger. 2016. Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory. *American Journal of Political Science* 60 (1):234–49.
Kertzer, Joshua D., Kathleen Powers, Brian C. Rathbun, and Ravi Iyer. 2014. Do Moral Values Shape Foreign Policy Preferences? *Journal of Politics* 76 (3):825–40.
Kertzer, Joshua D., and Thomas Zeitzoff. 2017. A Bottom-Up Theory of Public Opinion About Foreign Policy. *American Journal of Political Science* 61 (3):543–58.
Koopman, Cheryl, Rose McDermott, Robert Jervis, Jack Snyder, and Joe Dioso. 1995. Stability and Change in American Elite Beliefs About International Relations. *Peace and Conflict: Journal of Peace Psychology* 1 (4):365–82.
Koopman, Cheryl, Jack Snyder, and Robert Jervis. 1990. Theory-Driven versus Data-Driven Assessment in a Crisis: A Survey of International Security Readers. *Journal of Conflict Resolution* 34 (4):694–722.
Krebs, Ronald R., and Patrick Thaddeus Jackson. 2007. Twisting Tongues and Twisting Arms: The Power of Political Rhetoric. *European Journal of International Relations* 13 (1):35–66.
Lake, David A., and Robert Powell. 1999. International Relations: A Strategic-Choice Approach. In *Strategic Choice and International Relations*, edited by David A. Lake and Robert Powell, 3–38. Princeton, NJ: Princeton University Press.
Larson, Deborah Welch. 1994. The Role of Belief Systems and Schemas in Foreign Policy Decision-Making. *Political Psychology* 15 (1):17–33.
LeVeck, Brad L., and Neil Narang. 2017. How International Reputation Matters: Revisiting Alliance Violations in Context. *International Interactions* 43 (5):797–821.
Levendusky, Matthew S., and Michael C. Horowitz. 2012. When Backing Down Is the Right Decision: Partisanship, New Information, and Audience Costs. *Journal of Politics* 74 (2):323–38.
Levy, Jack S. 1994. Learning and Foreign Policy: Sweeping a Conceptual Minefield. *International Organization* 48 (2):279–312.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

Levy, Jack S., Michael K. McKoy, Paul Poast, and Geoffrey P.R. Wallace. 2015. Backing Out or Backing In? Commitment and Consistency in Audience Costs Theory. *American Journal of Political Science* 59 (4):988–1001.

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

McDermott, Rose. 2004. The Feeling of Rationality: The Meaning of Neuroscience for Political Science. *Perspectives on Politics* 2 (4):691–706.

McDermott, Rose, Dustin Tingley, Jonathan Cowden, Giovanni Frazzetto, and Dominic D.P. Johnson. 2009. Monoamine Oxidase: A Gene (MAOA) Predicts Behavioral Aggression Following Provocation. *Proceedings of the National Academy of Sciences* 106 (7):2118–23.

McGillivray, Fiona, and Alastair Smith. 2008. *Punishing the Prince: A Theory of Interstate Relations, Political Institutions, and Leader Change*. Princeton, NJ: Princeton University Press.

McGovern, George S., and William R. Polk. 2006. The Way Out of War: A Blueprint for Leaving Iraq Now. *Harper's* October: 31–39.

McMahon, Robert J. 1991. Credibility and World Power: Exploring the Psychological Dimension in Postwar American Diplomacy. *Diplomatic History* 15 (4):455–72.

Mead, Walter Russell. 2002. *Special Providence: American Foreign Policy and How It Changed the World*. New York: Routledge.

Mellers, B.A., A. Schwartz, and A.D.J. Cooke. 1998. Judgment and Decision Making. *Annual Review of Psychology* 49:447–77.

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

———. 2005. Rationality and Psychology in International Politics. *International Organization* 59 (1):77–106.

Mildenberger, Matto, and Dustin Tingley. 2017. Beliefs About Beliefs: The Importance of Second-Order Opinions for Climate Politics. *British Journal of Political Science*. Published online 4 December <https://doi.org/10.1017/S0007123417000321>.

Miller, Warren E., and Donald E. Stokes. 1963. Constituency Influence in Congress. *American Political Science Review* 57 (1):45–56.

Milliken, Jennifer L. 1996. Metaphors of Prestige and Reputation in American Foreign Policy and American Realism. In *Post-Realism: The Rhetorical Turn in International Relations*, edited by Francis A. Beer and Robert Hariman, 217–39. East Lansing: Michigan State University Press.

Mitzen, Jennifer. 2005. Reading Habermas in Anarchy: Multilateral Diplomacy and Global Public Spheres. *American Political Science Review* 99 (3):401–17.

Modigliani, Andre. 1972. Hawks and Doves, Isolationism and Political Distrust: An Analysis of Public Opinion on Military Policy. *The American Political Science Review* 66 (3):960–78.

Morgan, Patrick. 1985. Saving Face for the Sake of Deterrence. In *Psychology and Deterrence*, edited by Robert Jervis, Richard Ned Lebow, and Janice Gross Stein, 125–52. Baltimore, MD: John Hopkins University Press.

Morgenthau, Hans J. 1985. *Politics Among Nations: The Struggle for Power and Peace*. Boston, MA: McGraw-Hill.

Mueller, John. 1973. *War, Presidents, and Public Opinion*. New York: John Wiley.

Müller, Harald. 2004. Arguing, Bargaining and All That: Communicative Action, Rationalist Theory and the Logic of Appropriateness in International Relations. *European Journal of International Relations* 10 (3):395–435.

Murdie, Amanda M., and David R. Davis. 2012. Shaming and Blaming: Using Events Data to Assess the Impact of Human Rights INGOs. *International Studies Quarterly* 56 (1):1–16.

Murray, Shoon Kathleen. 1996. *Anchors Against Change: American Opinion Leaders' Beliefs After the Cold War*. Ann Arbor: University of Michigan Press.

Nelson, Stephen C., and Peter J. Katzenstein. 2014. Uncertainty, Risk, and the Financial Crisis of 2008. *International Organization* 68 (2):361–92.

Nyhan, Brendan, and Jason Reifler. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* 32 (2):303–30.

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

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

Potter, Philip B.K., and Matthew A. Baum. 2014. Looking for Audience Costs in All the Wrong Places: Electoral Institutions, Media Access and Dispute Reciprocation. *Journal of Politics* 76 (1): 167–81.

Poznansky, Michael. 2015. Statis or Decay? Reconciling Covert War and the Democratic Peace. *International Studies Quarterly* 59 (4):815–26.

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

Putnam, Robert D. 1988. Diplomacy and Domestic Politics: The Logic of Two-Level Games. *International Organization* 42 (3):427–60.

Rathbun, Brian C., Joshua D. Kertzer, and Mark Paradis. 2017. Homo Diplomaticus: Mixed-Method Evidence of Variation in Strategic Rationality. *International Organization* 71 (S1):S33–S60.

Rathbun, Brian C., Joshua D. Kertzer, Jason Reifler, Paul Goren, and Thomas J. Scotto. 2016. Taking Foreign Policy Personally: Personal Values and Foreign Policy Attitudes. *International Studies Quarterly* 60 (1):124–37.

Ravenal, Earl C. 1982. Counterforce and Alliance: The Ultimate Connection. *International Security* 6 (4): 26–43.

Renshon, Jonathan, Allan Dafoe, and Paul K. Huth. 2018. Leader Influence and Reputation Formation in World Politics. *American Journal of Political Science*. First published 24 January online at <https://doi.org/10.1111/ajps.12335>.

Risse, Thomas. 2000. "Let's Argue!" Communicative Action in World Politics. *International Organization* 54 (1):1–39.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2014. *STM: R Package for Structural Topic Models*. R package version 1.0.8.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science* 58 (4):1064–82.

Ross, Lee. 1977. The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process. In *Advances in Experimental Social Psychology*, Vol. 10, edited by Leonard Berkowitz, 173–220. New York: Academic Press.

Sartori, Anne E. 2005. *Deterrence by Dipomacy*. Princeton, NJ: Princeton University Press.

Saunders, Elizabeth N. 2011. *Leaders at War: How Presidents Shape Military Interventions*. Ithaca, NY: Cornell University Press.

———. 2015. War and the Inner Circle: Democratic Elites and the Politics of Using Force. *Security Studies* 24 (3):466–501.

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

———. 1966. *Arms and Influence*. New Haven, CT: Yale University Press.

Schimmelfennig, Frank. 2001. The Community Trap: Liberal Norms, Rhetorical Action, and the Eastern Enlargement of the European Union. *International Organization* 55 (1):47–80.

Schultz, Kenneth A. 2001. Looking for Audience Costs. *Journal of Conflict Resolution* 45 (1):32–60.

Sechser, Todd S. 2010. Goliath's Curse: Coercive Threats and Asymmetric Power. *International Organization* 64 (4):627–60.

Shannon, Vaughn P., and Michael Dennis. 2007. Militant Islam and the Futile Fight for Reputation. *Security Studies* 16 (2):287–317.

Sharman, J.C. 2007. Rationalist and Constructivist Perspectives on Reputation. *Political Studies* 55 (1): 20–37.

Simmons, Beth A. 2010. Treaty Compliance and Violation. *Annual Review of Political Science* 13:273–96.

Simon, Herbert A. 1973. The Structure of Ill Structured Problems. *Artificial Intelligence* 4 (3–4):181–201.

Snidal, Duncan. 2002. Rational Choice and International Relations. In *Handbook of International Relations*, edited by Walter Carlsnaes, Thomas Risse, and Beth A. Simmons, 73–94. London: Sage.

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

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press

Snyder, Jack, and Erica D. Borghard. 2011. The Cost of Empty Threats: A Penny, Not a Pound. *American Political Science Review* 105 (3):437–56.
Stein, Janice. 1994. Political Learning by Doing: Gorbachev as Uncommitted Thinker and Motivated Learner. *International Organization* 48 (2):155–83.
———. 2002. Psychological Explanations of International Conflict. In *Handbook of International Relations*, edited by Walter Carlsnaes, Thomas Risse, and Beth A. Simmons, 292–308. London: Sage.
Sylvan, Donald A., and James F. Voss. 1998. *Problem Representation in Foreign Policy Decision-Making*. New York: Cambridge University Press.
Tang, Shiping. 2005. Reputation, Cult of Reputation, and International Conflict. *Security Studies* 14 (1): 34–62.
Tetlock, Philip E. 1983. Policy-makers' Images of International Conict. *Journal of Social Issues* 39 (1):67–86.
Tetlock, Philip E. 1999. Theory-Driven Reasoning About Plausible Pasts and Probable Futures in World Politics: Are We Prisoners of Our Preconceptions? *American Journal of Political Science* 43 (2):335–66.
Tingley, Dustin. 2017. Rising Power on the Mind. *International Organization* 71 (S1):S165–88.
Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. 2014. Mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software* 59 (5):1–38. Available at <https://www.jstatsoft.org/article/view/v059i05>.
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.
Tomz, Michael. 2007a. Domestic Audience Costs in International Relations: An Experimental Approach. *International Organization* 61 (4):821–40.
———. 2007b. *Reputation and International Cooperation: Sovereign Debt Across Three Centuries*. Princeton, NJ: Princeton 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.
Voss, James F., and Timothy A. Post. 1988. On the Solving of Ill-structured Problems. In *The Nature of Expertise*, edited by Michelene H. Chi, Robert Glaser, and Marshall J. Farr, 261–85. Hillsdale, NJ: Erlbaum.
Wallace, Geoffrey P.R. 2015. The Reputational Consequences of International Legal Commitments. Working paper.
Walt, Stephen M. 2015. The Credibility Addiction. *Foreign Policy*, 6 January. Available at <http://foreign-policy.com/2015/01/06/the-credibility-addiction-us-iraq-afghanistan-unwinnable-war/>.
Walter, Barbara F. 2006. Building Reputation: Why Governments Fight Some Separatists but Not Others. *American Journal of Political Science* 50 (2):313–30.
Waltz, Kenneth N. 1979. *Theory of International Politics*. Boston: McGraw-Hill.
Weeks, Jessica L. 2008. Autocratic Audience Costs: Regime Type and Signaling Resolve. *International Organization* 62 (1):35–64.
Weisiger, Alex, and Keren Yarhi-Milo. 2015. Revisiting Reputation: How Past Actions Matter In International Politics. *International Organization* 69 (2):473–95.
Wight, Colin. 2006. *Agents, Structures and International Relations: Politics as Ontology*. Cambridge, UK: Cambridge University Press.
Wittkopf, Eugene R. 1990. *Faces of Internationalism: Public Opinion and American Foreign Policy*. Durham, NC: Duke University Press.
Yarhi-Milo, Keren. 2018. *Who Fights for Reputation in International Politics? Leaders, Resolve, and the Use of Force*. Princeton, NJ: Princeton University Press.
Yarhi-Milo, Keren, and Marcus Holmes. 2017. The Psychological Logic of Peace Summits: How Empathy Shapes Outcomes of Diplomatic Negotiations. *International Studies Quarterly* 61 (1):107–22.
Zaller, John R. 1992. *The Nature and Origins of Mass Public Opinion*. New York: Cambridge University Press.

https://doi.org/10.1017/S0020818318000188 Published online by Cambridge University Press