Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations

Joshua D. Kertzer, Brian C. Rathbun, and Nina Srinivasan Rathbun

Abstract Canonical models of costly signaling in international relations (IR) tend to assume costly signals speak for themselves: a signal's costliness is typically understood to be a function of the signal, not the perceptions of the recipient. Integrating the study of signaling in IR with research on motivated skepticism and asymmetric updating from political psychology, we show that individuals' tendencies to embrace information consistent with their overarching belief systems (and dismiss information inconsistent with it) has important implications for how signals are interpreted. We test our theory in the context of the 2015 Joint Comprehensive Plan of Action (JCPOA) on Iran, combining two survey experiments fielded on members of the American mass public. We find patterns consistent with motivated skepticism: the individuals most likely to update their beliefs are those who need reassurance the least, such that costly signals cause polarization rather than convergence. Successful signaling therefore requires knowing something about the orientations of the signal's recipient.

Understanding international relations (IR) as a set of strategic interactions marked by incomplete information is one of the most important developments in the study of world politics in recent decades.[1] In the face of uncertainty about others' resolve or intentions, states are thought to signal information in an effort to either threaten or reassure.[2] Signaling is important both to cooperation and peace building where the challenge is to convey peaceful intentions or credibly commit not to defect on collaborative efforts, as well as in coercive bargaining situations in which actors are trying to demonstrate that they will stand firm.

The general presumption in this literature, of course, is that signals must be costly given strategic incentives to dissemble.[3] In cooperative scenarios, there is the temptation to defect, leaving one's potential partner with the sucker pay-off.[4] In crisis bargaining, it might pay to bluff about one's willingness to fight. Therefore states must take action that reveals their type—behavior that only a resolute state, or a peaceful state, or a trustworthy state, would exhibit.

1. See Keohane 1984; Lake and Powell 1999.
2. See Fearon 1994; Kydd 2005; Schultz 2001.
3. See Fearon 1994; Schelling 1966.
4. Axelrod 1984.

International Organization, 2019, page 1 of 24
© The IO Foundation, 2019
doi:10.1017/S0020818319000328

The literature on costly signals generally rests on the assumption that what is costly (and what is not) is obvious to the observer, the intended target of the informative behavior. Yet as Jervis writes, "it might seem that problems of interpretation do not arise with costly signals because costs are objective and will be seen the same way by all participants. But what the actor feels to be a cost, observers may not so categorize, or vice versa."[5] And it is not just senders and recipients who fail to see eye to eye. Often, the recipients themselves disagree on how to best interpret an actor's intentions, as disagreements between Brent Scowcroft and James Baker in the George H.W. Bush administration, George Shultz and Caspar Weinberger in the Reagan administration, and Joseph Davies and James Byrnes in the Truman administration show.

Unlike their rationalist counterparts, psychologically informed theories of foreign policy behavior have long problematized objective perception and information processing.[6] One of the recurring themes in psychological approaches to world politics is that individuals engage in belief assimilation and belief perseverance. Rather than continually adjusting their beliefs in light of objective evidence, they embrace the information that matches their prior attitudes and then explain away anomalies.[7] There are likely important individual differences in the perception of signals as a result. Similarly, studies in political behavior, meant to highlight the problem of ideological polarization, have shown that we are "motivated skeptics," dismissing evidence that fails to align with our preexisting beliefs. Insomuch as we change our beliefs it is generally in the direction of strengthening them, not reevaluating them, a phenomenon called "asymmetric updating." This is highly germane for the study of costly signaling, with implications at both the micro and macro level. To know whether a signal is credible enough, we have to know something about the preexisting beliefs of the perceiver, and those most likely to change their beliefs are those who already want to do so.[8]

We seek to bring these notions of motivated skepticism and asymmetric updating into the study of costly signaling in IR, studying a real-world example of reassurance, rather than resolve. We offer a theory of *who* finds what types of signals, costly or otherwise, informative. We adapt existing theories of foreign policy attitudes into theories of updating and assimilation, hypothesizing that political ideology, foreign policy postures such as cooperative internationalism and militant internationalism, as well as preexisting images of Iran likely influence how individuals process signals in foreign policy.

In the run-up to the 2015 Joint Comprehensive Plan of Action (JCPOA) seeking to limit the Iranian nuclear program, we fielded a survey experiment asking respondents questions about potential outcomes of negotiations between the United States and

5. Jervis 1970, 301.
6. See Breslauer and Tetlock 1991; Jervis 1976; Jervis, Lebow, and Stein 1989; Kertzer 2016.
7. See Jervis 1976; Mercer 1996; Tetlock 1998; Yarhi-Milo 2014.
8. Yarhi-Milo, Kertzer, and Renshon 2018.

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Iran. Manipulating the signals of reassurance that Iran might send to the United States along two dimensions—limiting uranium enrichment and allowing inspections with different degrees of intrusiveness—we look to see whether such signals, represented in the form of hypothetical negotiating positions, change estimates of Iran's trustworthiness, the threat it poses, and support for a nuclear deal in which economic sanctions are lifted. In other words, what is the price of peace that respondents charge for an improvement in relations with Iran?

Consistent with canonical models of costly signaling, we find that, on average, respondents change their beliefs in response to costly signals of reassurance. This is an important—and reassuring—finding for a research program that is generally purely deductive or tested on large-*N* data sets that do not permit us to observe behavioral microfoundations.[9] However, we also find substantial and theoretically important evidence of heterogeneity. While some respondents change their views of Iran in response to costly signals, others do not, and this is a function of prior dispositions: cooperative internationalists and liberals embrace corrective information that bolsters their beliefs, while others largely dismiss such signals. In other words, *it is precisely those who are motivated to find evidence of a costly signal* who act as classic signaling models would expect, while those motivated not to update their beliefs do not respond to the treatments to the same degree, and sometimes not at all. Even though, on average, individuals update in response to costly signals, their beliefs polarize, rather than converge. We find similar results in a supplementary experiment fielded three years later in a very different context: the lead-up to the Trump administration's decision to withdraw from the deal.

Motivated Skepticism and the Logic of Costly Signals

There has been an explosion of interest in recent years in signaling processes in international relations. Following Schelling, the focus has primarily been on the signaling of resolve in crises.[10] Before the recent surge of interest in crisis bargaining, however, IR scholars were more preoccupied with questions of cooperation, such as sending signals of reassurance to escape mutually detrimental security dilemma dynamics.[11] The presumption in all of this literature, however, is that costly signals speak for themselves. Whether a signal is costly or not is inherent in the signal, not to those who are interpreting it. Indeed the point of costly signals is that they help differentiate between two different potential states of the world (an actor is either trustworthy or untrustworthy, resolute or irresolute, and so on) in a way that cheap talk does not. Costly signals are those that have little room for interpretation; it is this quality that allows for updating and convergence on the part of those who observe them.

9. See Fearon 1994; Kertzer 2017; Schultz 2001.
10. See Schelling 1966; Carson and Yarhi-Milo 2017; Fearon 1994; Schultz 2001.
11. See Axelrod 1984; Glaser 2010; Jervis 1978; Keohane 1984; Kydd 2005.

Although these canonical models of signaling have made many valuable contributions, we believe there are reasons to revisit the assumption that signals speak for themselves. There is a rich literature on cognition in IR, much of which rests on the fundamental assumption that perception should not be taken for granted.[12] This logic has been extended to signaling as well. Robert Jervis, one of the foundational theorists of costly signaling, argues that signals are rarely perceived as the sender intended, which contributes to the frequent (and momentous) misperceptions in the history of international relations.[13] Information is interpreted through ideological filters, such that "knowing how theorists read a signal does not tell us how the perceiver does."[14]

We take our cue from several decades of studies in psychology and political behavior that emphasize how difficult it is to induce individuals to revise their preexisting beliefs. While both psychologists and political scientists have been familiar with phenomena like motivated reasoning, confirmation bias, and disconfirmation bias for some time, these dynamics have taken on particular resonance in our contemporary era of pronounced ideological polarization.[15] Scholars in this tradition typically focus on one of two mechanisms that are related in practice, but which we can treat as conceptually distinct. The first is what we refer to as *motivated skepticism*, which involves how individuals evaluate information; the second we refer to as *selective attention*, which involves the types of information individuals seek out in the first place.

The logic of motivated skepticism is simple. Taber and Lodge explain how "ideally, one's prior beliefs and attitudes—whether scientific or social—should 'anchor' the evaluation of new information and then, depending on how credible is some piece of evidence, impressions should be adjusted upward or downward."[16] However, as Kunda argues, when we receive information that seems to confound our beliefs, we instead become "motivated skeptics."[17] As Nyhan and Reifler summarize, "humans are goal-directed information processors who tend to evaluate information with a directional bias toward reinforcing their pre-existing views … Specifically people tend to display bias in evaluating political arguments and evidence, favoring those that reinforce their existing views and disparaging those that contradict their views."[18] These ideas are so widely held in psychology that some cognitive scientists have started viewing these phenomena as features rather than bugs, arguing that human reasoning faculties were designed precisely to make us better at doubling down on our preconceptions.[19]

12. See Jervis 1976; Tetlock 1998.
13. Jervis, Lebow, and Stein 1989; see also Shifrinson 2019.
14. See Jervis 2002, 297–98; Albuyeh and Paradis 2018; Quek 2016; Yarhi-Milo, Kertzer, and Renshon 2018.
15. See Nyhan and Reifler 2010; Redlawsk 2002; Taber and Lodge 2006.
16. Taber and Lodge 2006, 755.
17. Kunda 1990.
18. Nyhan and Reifler 2010, 307.
19. Mercier and Sperber 2011.

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

These studies show that the updating of beliefs in light of new information does occur, but is asymmetric in nature. Those for whom new facts are less threatening, or confirm preexisting beliefs, are found to strengthen their attitudes, while others are slower to update, or resist altogether.[20] These asymmetries at the evaluation phase are further compounded by asymmetries in terms of what kinds of information individuals seek out in the first place. As Yarhi-Milo puts it, individuals pay "selective attention" to signals, often seeking out information consistent with their prior attitudes, a confirmation bias.[21] Thus, individuals not only downweight signals inconsistent with their beliefs, but also seek out signals consistent with them.[22] And, importantly, these tendencies do not decrease with cognitive sophistication.[23]

Motivated skepticism suggests that individuals might interpret signals, even costly ones, in ways that buttress or reinforce their existing beliefs so that signals induce polarization rather than convergence among recipients. We focus on the effects of three kinds of preexisting beliefs on updating in response to costly signals. Our theoretical intuition is that these predispositions will influence how and whether individuals change their beliefs in light of costly signals of reassurance. That is, how much an individual is reassured by a signal from Iran is partially a function of whether they want to cooperate with Iran in the first place.

Dispositions, Ideology, and Images in Foreign Policy: Individual-level Moderators of Costly Signaling

While psychological approaches to decision making are common in IR, there have been few systematic studies of how individuals react to costly signals in the foreign policy arena with specific hypotheses about who is more or less likely to update.[24] Who are the types of individuals who are the most motivated to find evidence of reassurance? We construct a deductive theory of motivated skepticism that, in light of prior research, draws on the ideological axes we see as most likely to affect signals' interpretation. We consider three such kinds of predispositions here.

First, the most prominent tradition in research on foreign policy beliefs consistently finds that American foreign policy attitudes at both the mass and elite levels are organized along two main related but distinct dimensions: cooperative internationalism (CI) and militant internationalism (MI).[25] CI and MI operate as what are called "postures" or "orientations"—general approaches toward foreign policy that structure

20. See Nyhan and Reifler 2010; Taber and Lodge 2006; Sunstein et al. 2017.
21. Yarhi-Milo 2014.
22. We differentiate between motivated skepticism and selective attention more extensively in Appendix section 1, where we also model motivated skepticism in a Bayesian framework, in which the likelihood is a function of an individual's prior beliefs.
23. Baekgaard et al. 2017; Taber and Lodge 2006.
24. For an exception, see Albuyeh and Paradis 2018.
25. See Holsti and Rosenau 1988; Wittkopf 1990.

foreign policy attitudes. Even for elite decision makers, foreign policy attitudes on specific subjects are generally reducible to broader orientations.[26]

Cooperative internationalism (CI) is an orientation toward international affairs that stresses concern for others abroad with whom one should work toward common goals.[27] Global solidarity is therefore a key element of cooperative internationalism. However, cosmopolitanism is about more than just self-sacrifice and service to others since cooperative internationalists also believe that cooperation leads to mutual gains. Accordingly, previous work has found that support for international institutions, multilateralism, and international collaboration all load on the same CI dimension.[28] We expect that this desire for cooperation will make cooperative internationalists highly receptive to signals of reassurance from Iran. A deal on its nuclear weapons program would be a victory for international cooperation. Cooperative internationalists are looking for credible signals of cooperative intent.

Militant internationalism (MI), on the other hand, is generally thought to mark the ubiquitous division between hawks and doves over the importance, effectiveness, and desirability of using coercion or force to reach foreign policy objectives. Hurwitz and Peffley posit a "dimension of militarism … anchored, on the one end, by a desire that the government assume an assertive, militant foreign-policy posture through military strength and on the other by a desire for a more flexible and accommodating stance through negotiations."[29] Hawkish and dovish postures are thought to rest on different cognitive "models" about the effectiveness of force.[30] Hawks embrace the "deterrence model," in which peace is best achieved through strength and the demonstration of resolve. In this world view, lack of credibility and signs of weakness invite challenges by aggressive foes in a dangerous environment. Doves, in contrast, point out the often self-defeating nature of such displays because they can incite fear on the other side and escalate hostilities in such a way that leaves both sides worse off.

Those high in militant internationalism see the world as a more threatening place and believe that force and coercion are the most effective way of securing national interest. We expect that those high in MI will therefore be less receptive to signals of reassurance from Iran that suggest the possibility of a less threatening Iran and a cooperative, peaceful outcome to the standoff over its nuclear program.

Political ideology also likely influences how individuals react to new information on foreign policy issues. Liberals and conservatives in the United States have been shown to differ remarkably on issues of international politics, both at the mass and the elite levels.[31] Liberals generally have a more benign view of the international environment that allows them more room to consider multilateralism and pursue a

26. Yarhi-Milo, Kertzer, and Renshon 2018.
27. See Chittick, Billingsley, and Travis 1995; Nincic and Ramos 2010.
28. See Holsti and Rosenau 1988; Wittkopf 1990.
29. Hurwitz and Peffley 1987, 1107.
30. See Jervis 1976; Tetlock 1983.
31. See Gries 2014; Holsti and Rosenau 1988.

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

diplomatic style of mutual gain seeking.[32] In contrast, conservatives see a more threatening environment both at home and abroad. Jost and colleagues identify this "motivational goal" of security as the essence of conservatism.[33] At home this makes the right more authoritarian, in favor of coercive means to maintain societal order. Conservatives aim to "protect" whereas liberals aim to "provide."[34]

Finally, we expect that preexisting "images" of Iran will influence the degree to which individuals respond to Iranian behavior. Psychological theories of IR argue that once a particular image of another has formed it becomes highly resistant to reinterpretation, such that national self-images tend to be self-perpetuating.[35] Iran is often regarded negatively in the United States as inherently threatening, dangerous, and untrustworthy. We expect that those with warmer feelings toward Iran will be more receptive to costly signals of reassurance, whereas those who dislike Iran will be more closed off.

Research Design

To examine the microfoundations of costly signaling, we conducted two survey experiments based on a real-world security issue—US efforts to strike a deal with Iran to limit its nuclear program. In both cases, our interest was less in establishing the state of public opinion on the matter, but rather, to investigate whether signals a country might send during an effort to strike a cooperative deal with an adversary generally regarded as untrustworthy and adversarial affect respondents' attitudes about the nature of that other. In other words, we want to know whether, when, and who updates their beliefs about another country's "type" based on costly signals. Our primary experiment was fielded immediately before the details of the Joint Comprehensive Plan of Action (JCPOA) was announced in July 2015. Our follow-up experiment, which we turn to briefly afterwards, and discuss in greater detail in Appendix section 4, was fielded immediately before the Trump administration announced its plans to pull out of the deal in April 2018.

Following consultations with individuals personally involved in negotiations on Iranian nuclear proliferation issues in the past, we based our creation of experimental treatments in our primary experiment on the main issues tackled in the Parameters for a Joint Comprehensive Plan of Action, the framework deal agreed to by the P5 + 1, the European Union, and Iran in Lausanne on 2 April 2015.[36] The parameters, which

32. Rathbun 2014.

33. Jost et al. 2003.

34. Janoff-Bulman 2009. In focusing on foreign policy orientations and ideological predispositions, we differ from the study of motivated reasoning in US politics in that our focus is on *ideologies* as the predispositions of interest, rather than merely *partisanship* (although as we show in Appendix section 2, our results hold even when controlling for partisan identification). We return to this point later.

35. See Herrmann and Fischerkeller 1995; Hirshberg 1993.

36. Author interviews, May 2015. For more details on the eventual terms of the Iran Deal, see Appendix section 3.

set the groundwork for the JCPOA that was announced the following July, divided the issue into five key categories: (1) uranium enrichment, (2) inspections and transparency, (3) limitations on heavy-water reactors that can be used to produce plutonium and a requirement to ship abroad spent fuel that can be reprocessed to produce plutonium, (4) sanctions, and (5) phasing. The first three categories contained the meat of the issues for containing Iran's ability to break out of the NPT and produce a nuclear weapon.

For reasons of tractability, our primary experiment focuses on the first two issues of the framework deal. The experiment began by presenting respondents with the following introduction to the issue:

> There has been a lot of discussion in the news lately about Iran and its nuclear program. We'd like to ask you some questions about it. In case you have not been keeping up with the news, here is a quick summary of the main issues. The United States and its allies are concerned that Iran might be using its development of nuclear materials in order to develop nuclear weapons. The United Nations Security Council has imposed economic sanctions on Iran for violating its previous obligations regarding its nuclear program and to prevent United Nations members from providing assistance to Iran that might be used for nuclear weapons production. Iran claims that its nuclear program is solely for peaceful purposes, such as providing power or for use in medical treatments.
>
> The two sides are currently negotiating about just what Iran will do in order to end the sanctions. One issue is whether Iran will continue to be able to develop a material called enriched uranium, which is necessary both for a peaceful nuclear program AND a program aimed at developing nuclear weapons. Another issue is whether Iran will allow inspections by an outside international agency of the sites of its nuclear program to make sure Iran is living up to its agreement.

Survey respondents were then presented with a hypothetical Iranian proposal, randomly assigned along two dimensions: (1) the degree to which Iran would be willing to limit its production and enrichment of uranium (which we call the *fuel cycle* manipulation) and (2) the degree to which Iran would be willing to allow monitoring of its nuclear activities (which we call the *inspection* manipulation). These two sets of treatments were chosen for three reasons. First, they were two of the most important questions in the (at the time) ongoing negotiations between the two parties and were also relatively easy to explain, as opposed to questions of heavy water reactors, for instance. Second, they are well suited to the research question—they lent themselves to escalating degrees of peaceful reassurance by Iran that could be experimentally manipulated. More generally, by studying receptivity to different terms of the deal in a controlled experimental environment, we can bracket selective attention and focus on motivated skepticism instead.

Each treatment had three possible levels, producing a 3×3 factorial design, illustrated in Figure 1. In terms of uranium enrichment, respondents were told either that Iran would (1) continue to produce enriched uranium, (2) freeze its production of enriched

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

uranium, or (3) dismantle its capacity for producing enriched uranium. The lowest level of the treatment, continuing to produce enriched uranium, is clearly weaker than the framework, which would seriously limit the production of LEU in amounts, technology, concentration levels, and locations. The middle value of the treatment, freezing Iran's production of enriched uranium, was more similar to the framework agreement, with its focus on limiting and freezing the technology for enriching uranium at first-generation R-1 centrifuges. The framework went beyond the second treatment value by requiring the dismantling of two-thirds of existing centrifuges, the redesign of the Fordow enrichment facility to no longer enrich uranium, and the sealing of advanced generation centrifuges by the IAEA. The framework comes close to the highest value of the treatment, dismantling its capacity for producing enriched uranium, long a US goal, but which has not been on the negotiation table for more than a decade.

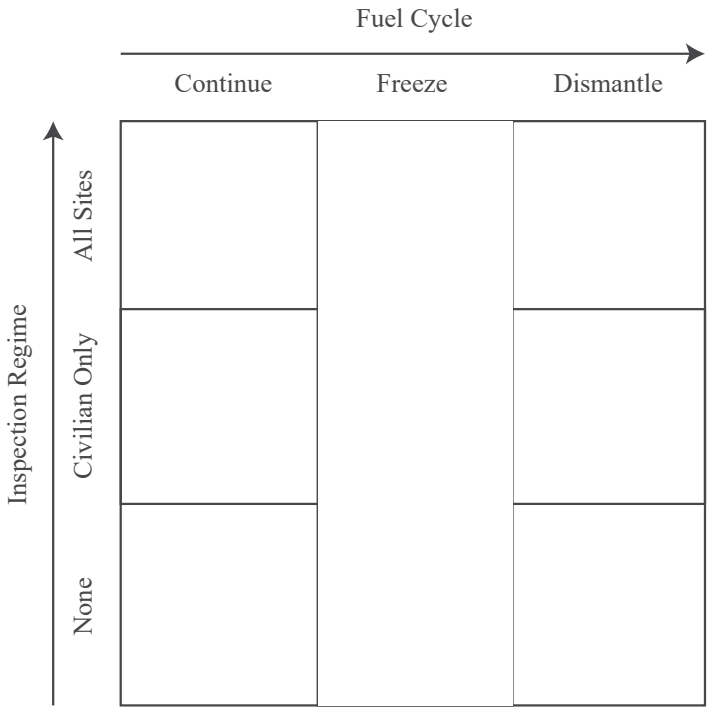

FIGURE 1. *Experimental design*

The inspection possibilities were that Iran proposed it would: (1) not allow the inspection by an international agency of any places where Iran has declared or is suspected of working with nuclear materials, regardless of whether they are military or nonmilitary sites; (2) allow the inspection by an international agency of all places where Iran has declared or is suspected of working with nuclear materials, with the

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

exception of military sites; or (3) allow the inspection by an international agency of all places where Iran has declared or is suspected of working with nuclear materials, including military sites. The lowest value, not allowing the inspection of declared or suspected sites of working with nuclear materials (the main basis for IAEA inspections), is clearly significantly weaker than the existing arrangement or the framework. The middle value of the treatment, to allow IAEA inspections in all declared and suspected sites except military ones, was closer to the status quo at the time. The framework comes closest to the highest treatment value in its emphasis on the IAEA's ability to investigate suspicious activities at any site in the country and its inclusion of measures to allow the IAEA to address concerns of the possible military dimensions of Iran's nuclear program.

After respondents were told of the terms that Iran was proposing, they were asked to rate both how dangerous and trustworthy they would perceive Iran to be if it made such an offer, on a scale of 0 to 10, with 10 being very trustworthy or very dangerous. Subjects also identified their degree of support for a deal on the terms they received with six response categories ranging from strongly oppose to strongly support. They were also asked how likely they felt it was that Iran would be willing to make a deal on those terms. Trustworthiness is of particular importance because it is at the crux of reassurance in IR.[37]

Hypotheses

Canonical models of costly signaling would expect the treatments to display significant effects: as Iran limits its production of uranium enrichment and/or allows more intrusive inspections, assessments of Iran's trustworthiness should increase and beliefs about its dangerousness should decrease. At the same time, however, given the centrality of enforcement in many rationalist models, classic signaling frameworks would likely argue that inspections are in general a more costly signal of reassurance than declarations to freeze or dismantle uranium enrichment because of the potential for commitment problems.[38] This gives rise to the following situational hypotheses, attributes of the structural environment that respondents are evaluating.

H1a: *The more Iran limits its fuel cycle capabilities, the more trustworthy and less dangerous respondents will perceive Iran to be.*

H1b: *The more extensive the inspections regime, the more trustworthy and less dangerous respondents will perceive Iran to be.*

Our model relies on hypotheses that expect a between-subjects difference in the effect of the survey's treatments—in other words, heterogeneous treatment effects in the form of an interaction with individual-level variables and experimental

37. See Booth and Wheeler 2007; Rathbun 2011.
38. See, for example, Koremenos, Lipson, and Snidal 2001.

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

conditions. Those who are most motivated to believe that Iran has benign intentions should respond most to the escalation in costly signals on the two dimensions.

Cooperative internationalists want and believe in multilateral cooperation and peaceful resolution. Militant internationalists are more attuned to threat and believe that force is more efficacious in securing American interests. Conservatives see the international (and domestic) environment as more dangerous than liberals perceive.

Therefore we hypothesize that, when it comes to costly signals of reassurance from Iran:

H2a: *Individuals high in CI will respond more than individuals low in CI.*

H2b: *Individuals high in MI will respond less than individuals low in MI.*

H2c: *Conservatives will respond less than liberals.*

H2d: *Those with more positive feelings toward Iran will respond more than individuals with more negative feelings.*

We would further expect that these individual differences will also be associated with evaluations of Iranian trustworthiness and dangerousness:

H3a: *Individuals high in CI will perceive Iran as more trustworthy and less dangerous.*

H3b: *Individuals high in MI will perceive Iran as less trustworthy and more dangerous.*

H3c: *Conservatives will perceive Iran as less trustworthy and more dangerous.*

H3d: *Those with more positive feelings toward Iran will perceive Iran as more trustworthy and less dangerous.*

This set of hypotheses is relevant given our interest in motivated skepticism. Because of the experimental design, any asymmetric patterns we recover will be a result of asymmetries at the evaluation stage, rather than the information-seeking stage, since all respondents within each treatment condition are presented with the same information. The experiment thus enables a more direct test of motivated skepticism than would be possible with observational data. In particular, if individuals high in CI see Iran as more trustworthy regardless of the signal it sends (H3a), and individuals who are high in CI are also more responsive to costly signals (H2a), this is consistent with the model of motivated skepticism we sketch out in Appendix section 1, in which those who respond the most to signals of reassurance are the ones who need the signals the least. If how individuals respond to costly signals of reassurance is partially a function of foreign policy orientations (H3), then signals will have polarizing effects: the costlier the signal of reassurance, the more respondents should disagree with one another about Iran's level of trustworthiness.

Results

We fielded our primary survey experiment on 1,816 American adults recruited through Amazon Mechanical Turk (MTurk) from 18 to 21 April 2015.[39] The survey was fielded following the conclusion of a general framework of an agreement with Iran but before final details of the pact were concluded in June 2015. This was an optimal time to collect data because the issue was in the public eye but not widely discussed in the media. In addition to their responses on Iran, subjects also completed a battery of items measuring three foreign policy dispositions—cooperative internationalism, militant internationalism, and isolationism—based on previous work in this tradition.[40]

TABLE 1. *Trustworthiness of Iran*

	(1)	*(2)*	*(3)*	*(4)*
INSPECTIONS: NONE	−1.839***	−2.014***	−2.062***	−2.071***
	(0.142)	(0.131)	(0.123)	(0.123)
INSPECTIONS: ALL	0.433***	0.336**	0.383***	0.375***
	(0.142)	(0.132)	(0.123)	(0.123)
FUEL CYCLE: CONTINUE	−0.557***	−0.427***	−0.416***	−0.420***
	(0.142)	(0.132)	(0.123)	(0.123)
FUEL CYCLE: DISMANTLE	0.014	0.097	0.048	0.050
	(0.141)	(0.131)	(0.122)	(0.122)
MILITANT INTERNATIONALISM		−2.734***	−1.321***	−1.348***
		(0.337)	(0.326)	(0.326)
COOPERATIVE INTERNATIONALISM		2.519***	1.908***	2.013***
		(0.316)	(0.297)	(0.300)
ISOLATIONISM		0.938***	1.061***	1.054***
		(0.312)	(0.291)	(0.291)
FEELINGS TOWARD IRAN			0.041***	0.041***
			(0.003)	(0.003)
NEED FOR COGNITION				−0.566**
				(0.230)
MALE		−0.026	−0.028	0.008
		(0.108)	(0.101)	(0.102)
LOG(AGE)		−0.891***	−0.702***	−0.688***
		(0.186)	(0.174)	(0.174)
WHITE		−0.229*	−0.131	−0.107
		(0.134)	(0.126)	(0.126)
EDUCATION		0.017	0.037	0.055
		(0.043)	(0.040)	(0.041)
IDEOLOGY		−0.270	−0.219	−0.230
		(0.252)	(0.235)	(0.235)
CONSTANT	4.583***	7.130***	4.729***	4.883***
	(0.129)	(0.839)	(0.797)	(0.799)
N	1,816	1,816	1,816	1,816
Adjusted R^2	0.145	0.272	0.365	0.367

Notes: Reference categories for treatments are the middle categories in each factor (INSPECTIONS: CIVILIAN ONLY, and FUEL CYCLE: FREEZE, respectively). *$p < .10$; **$p < .05$; ***$p < .01$.

39. See Appendix section 2 for further details.
40. See Kertzer et al. 2014; Rathbun et al. 2016.

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Model 1 in Table 1 presents a linear regression model estimating the average treatment effects (that is, the effects of different signals sent by Iran in the negotiations). Subsequent models in the table also control for a number of individual-level attributes to show the results remain unchanged. The effect of each treatment is measured with dummy variables. On both dimensions, the excluded category that provides the baseline by which to compare the treatment effect is the middle category: for uranium enrichment, this is a freeze on production; for inspection it is allowing access to all sites other than military facilities.

The results generally support H1 about the importance of situational factors—in this case, the different signals that Iran sends through its hypothetical negotiating offers. Three of the four treatment dummy variables are highly statistically significant. They explain about 15 percent of the variance in assessments of Iran's trustworthiness and they remain significant with similar substantive effects as we add in dispositional factors. We also see that the effect of inspections has a much stronger effect than offers on uranium enrichment on perceptions of trustworthiness. While there is a statistically significant difference between the effect of continuing to enrich uranium as opposed to freezing, there is no real difference between freezing and the more permanent step of dismantling production.

We find similar patterns in terms of the treatments' effects on participants' beliefs about the danger posed by Iran, and their support for the deal, shown in Table A1 of Appendix section 2. The inspection remains most important for determining respondents' attitudes about the danger Iran poses and overall support for a deal, and there is the same pattern of results regarding the fuel cycle. There is no difference between freezing and dismantling, but there is a difference between no limitations at all and stopping the enrichment of uranium. However, the treatments generally have a much stronger effect on respondents' assessments of Iran's trustworthiness than they do on their evaluations of how dangerous Iran is. The treatments account for only 7 percent of the variance in the evaluation of danger.

Figure 2 offers a clearer visualization of these average treatment effects. We see a general upward trend in evaluations of Iranian trustworthiness as inspections become more intrusive and a downward trend in evaluations of Iranian threat. The effect of inspection signals on trust is particularly striking. There is at least a two-point jump on a ten-point scale for those in all fuel cycle treatment groups of moving from no inspections to stringent inspections. Support for the deal also rises with greater monitoring. Respondents did not respond to the likelihood of Iran making different offers in a systematically different way depending on which treatments they receive, and were equally pessimistic that Iran would offer a relatively good deal or a bad deal to the United States.

The regression results in Table 1 and Table A1 in Appendix section 2 show that dispositional variables are associated with changes in perceptions of Iran's trustworthiness and dangerousness, in a manner consistent with our hypotheses: individuals high in CI, or with warmer feelings toward Iran, see it as more trustworthy, and less dangerous, regardless of what signal Iran sends, and are more likely to support a deal. In contrast, hawkish individuals high in MI see Iran as significantly

less trustworthy and more dangerous, and are less likely to support a deal. To ascertain whether these effects are consistent with motivated skepticism, however, we need to look at the role of these dispositions as moderators of treatment effects: how do they affect how signals are interpreted? As Table 2 shows, there is a strong interaction between cooperative internationalism and treatments of both kinds. The significant interaction terms indicate that those higher in cooperative internationalism are much more sensitive to Iranian willingness to limit uranium enrichment and allow inspections than those who are lower in CI, offering strong support for H3.

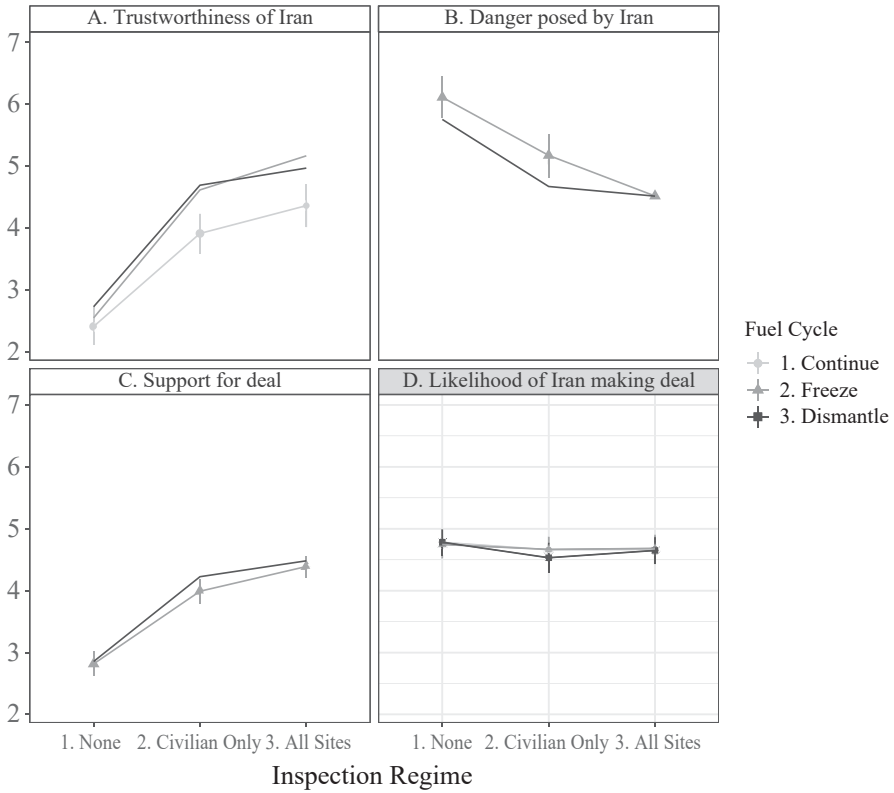

FIGURE 2. *The effect of costly signals*

Table 2 and Figure 3 present conditional average treatment effects in which the conditional effect of moving from the cheapest to the costliest signal in terms of both enrichment and inspection on beliefs about Iran is shown for varying levels of CI. The left-hand panel in Figure 3(a) shows that while those lowest in CI do not respond at all to the inspection treatments when it comes to their perceptions

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

of Iran, those highest in CI move up three points in their estimates of Iranian trustworthiness as Iran accepts the stringent inspections. We see the same pattern in the conditional effect of dismantling uranium enrichment in the right-hand panel. For those lowest in CI, the treatment effects are not statistically distinguishable from 0; for those highest, moving from the cheapest to the costliest signal about the fuel cycle leads to a one-point increase in perception of Iranian trustworthiness.

TABLE 2. *CI moderates the impact of costly signals*

	Trustworthy (1)	*Dangerous* (2)	*Approval of Deal* (3)
INSPECTIONS: NONE	−1.208***	0.356	−0.464*
	(0.405)	(0.421)	(0.245)
INSPECTIONS: ALL	−0.080	−0.592	0.135
	(0.388)	(0.403)	(0.235)
INSPECTIONS: NONE × CI	−1.274**	0.910	−1.170***
	(0.577)	(0.599)	(0.349)
INSPECTIONS: ALL × CI	0.659	0.305	0.219
	(0.560)	(0.582)	(0.339)
FUEL CYCLE: CONTINUE	0.299	0.630	−0.053
	(0.405)	(0.421)	(0.245)
FUEL CYCLE: DISMANTLE	0.017	−0.099	0.007
	(0.401)	(0.416)	(0.242)
FUEL CYCLE: CONTINUE × CI	−1.086*	0.160	−0.626*
	(0.578)	(0.600)	(0.350)
FUEL CYCLE: DISMANTLE × CI	0.045	−0.284	0.192
	(0.568)	(0.590)	(0.344)
COOPERATIVE INTERNATIONALISM	2.524***	−0.306	1.131***
	(0.547)	(0.569)	(0.331)
MILITANT INTERNATIONALISM	−1.423***	3.154***	−0.677***
	(0.326)	(0.339)	(0.197)
ISOLATIONISM	1.039***	−0.973***	0.736***
	(0.290)	(0.301)	(0.176)
FEELINGS TOWARD IRAN	0.041***	−0.023***	0.012***
	(0.003)	(0.003)	(0.002)
NEED FOR COGNITION	−0.556**	−0.192	−0.076
	(0.230)	(0.239)	(0.139)
MALE	0.012	−0.579***	−0.038
	(0.102)	(0.106)	(0.062)
LOG(AGE)	−0.686***	0.522***	−0.346***
	(0.173)	(0.180)	(0.105)
WHITE	−0.116	−0.233*	−0.004
	(0.125)	(0.130)	(0.076)
EDUCATION	0.051	−0.060	0.040
	(0.041)	(0.042)	(0.025)
IDEOLOGY	−0.205	1.012***	−0.420***
	(0.234)	(0.244)	(0.142)
CONSTANT	4.612***	3.551***	4.134***
	(0.851)	(0.884)	(0.515)
N	1,816	1,815	1,815
Adjusted R^2	0.371	0.287	0.318

Notes: Reference categories for treatments are the middle categories in each factor (INSPECTIONS: CIVILIAN ONLY, and FUEL CYCLE: FREEZE, respectively). See Table A2 in Appendix section 2 for equivalent results from a fully saturated three-way interaction model. *$p < .10$; **$p < .05$; ***$p < .01$.

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

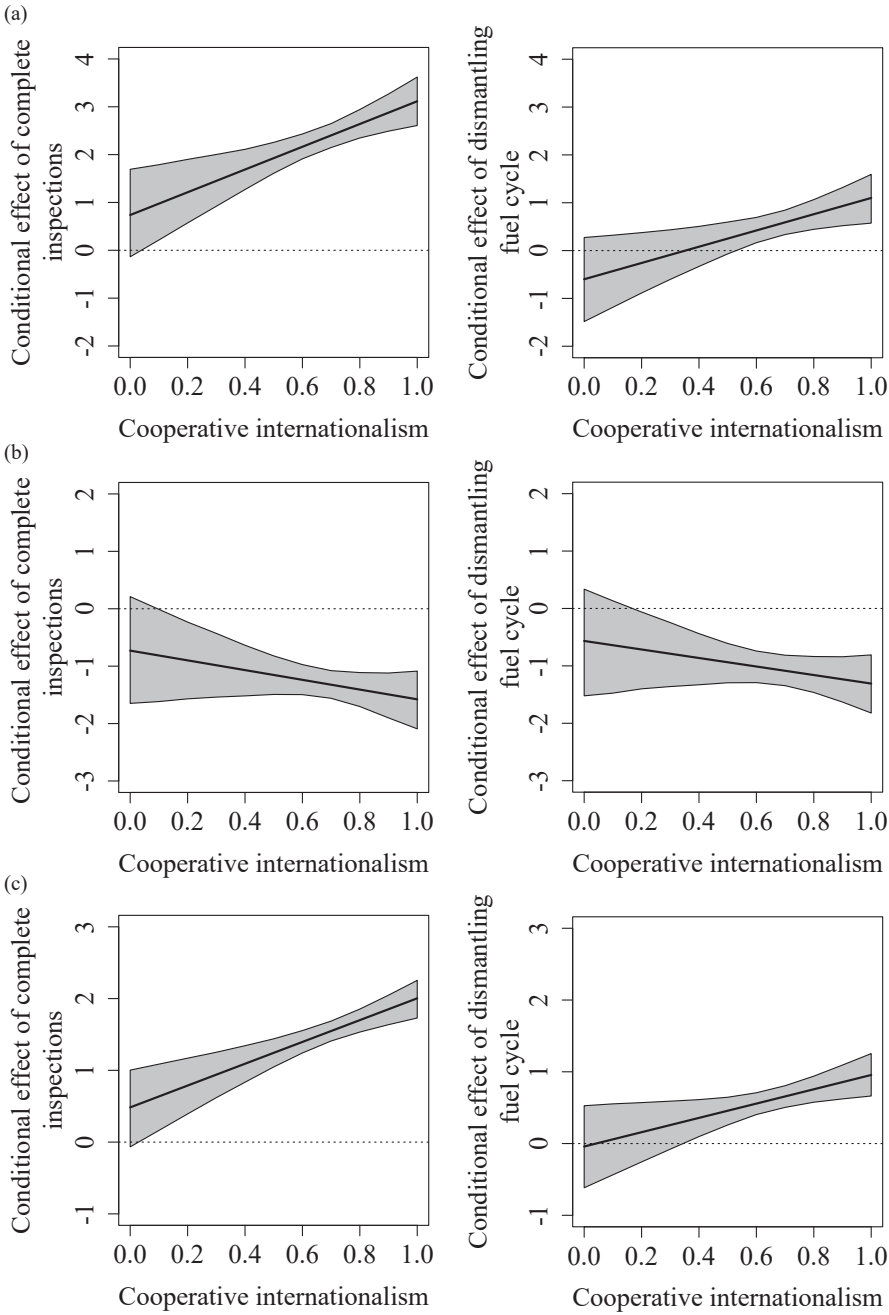

FIGURE 3. *Conditional effects of signals by cooperative internationalism*

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Figure 3(b) shows the conditional effect of the treatments by levels of cooperative internationalism on perceptions of the danger that Iran opposes. Here, we see a negative interaction. The conditional effect is not as strong as was the case with assessments of trustworthiness, but in the case of both the inspections and the fuel cycle, those highest in CI experience a one-point drop in their estimates of the threat posed by Iran as it makes the costliest signal in the experiment, whereas those lowest in CI do not seem to respond to these signals. Finally, Figure 3(c) shows that while for those who are lowest in CI, the steps that Iran is willing to take on the fuel cycle and on monitoring have relatively weak effects on their support for a deal, there are very strong effects for cooperative internationalists. A willingness to allow complete inspections is associated with a two-point increase on a six-point scale among those highest in CI for supporting a deal.

In Appendix section 2 we conduct a large number of robustness checks, including replicating these analyses using political ideology, militant internationalism, and preexisting attitudes about Iran as moderators instead. The ideology and MI results nicely mirror the CI results: liberals and doves update the most, even though they need the signals the least. We also find similar, though weaker, results using feelings toward Iran as a moderator. Looking across our findings, we find evidence that variations in foreign policy orientations have a substantial effect on both judgments of Iran's trustworthiness, and receptivity to costly signaling. This combination illustrates the asymmetry frequently found in the motivated skepticism literature. However, we should also stress that we also find a general trend toward updating for the sample as a whole. Our findings thus simultaneously offer reassuring evidence in favor of canonical models of signaling, while also revealing heterogeneity that is important to take into account. In Appendix section 2, we consider and find no evidence that our results reflect unobserved heterogeneous beliefs that might be correlated with our predispositions of interest and which might affect the weight attached to the information conveyed in the costly signal, such as knowledge about Iran or beliefs about the costliness of diplomatic concessions. Similarly, since the Iran deal was a highly partisan issue in 2015, our supplementary analysis controls for partisan identification, showing that partisanship cannot be responsible for the results we report here.[41] While not dismissing the role that partisanship plays in foreign policy questions, our results show the extent to which motivated reasoning has relevance for IR scholars far beyond the confines of partisanship.

We also conducted a follow-up experiment in April 2018, immediately before the Trump administration announced the United States was withdrawing from the JCPOA; in addition to fielding an experiment immediately before the JCPOA's

41. Figure A5 in Appendix section 2 also estimates heterogeneous treatment effects by party ID, finding similar, though slightly weaker, patterns to that displayed by political ideology; as we show in Figure A6 in the appendix, foreign policy orientations generally perform partisan identity, consistent with bottom-up theories of public opinion in foreign policy. Kertzer and Zeitzoff 2017.

birth, we also do so before its death. Unlike its predecessor, this supplementary experiment incorporates a within-subject component, measuring respondents' prior beliefs about Iranian trustworthiness before randomly presenting them with three different sets of signals from Iran, and then measuring respondents' posterior beliefs. In the compliance condition, respondents were told that the Iranian government had complied with the deal's terms. In the ballistic condition, respondents were told that the Iranian government had abided by the deal's terms, but was also behaving in an untrustworthy manner in terms not covered by the original agreement. In this sense, unlike the other treatment conditions, which should reassure respondents, the ballistic condition sends a mixed signal about Iranian trustworthiness. In the Pledge condition, respondents were presented with the same information as before about Iranian ballistic missile testing, but were also told that Iran had pledged to halt its testing.

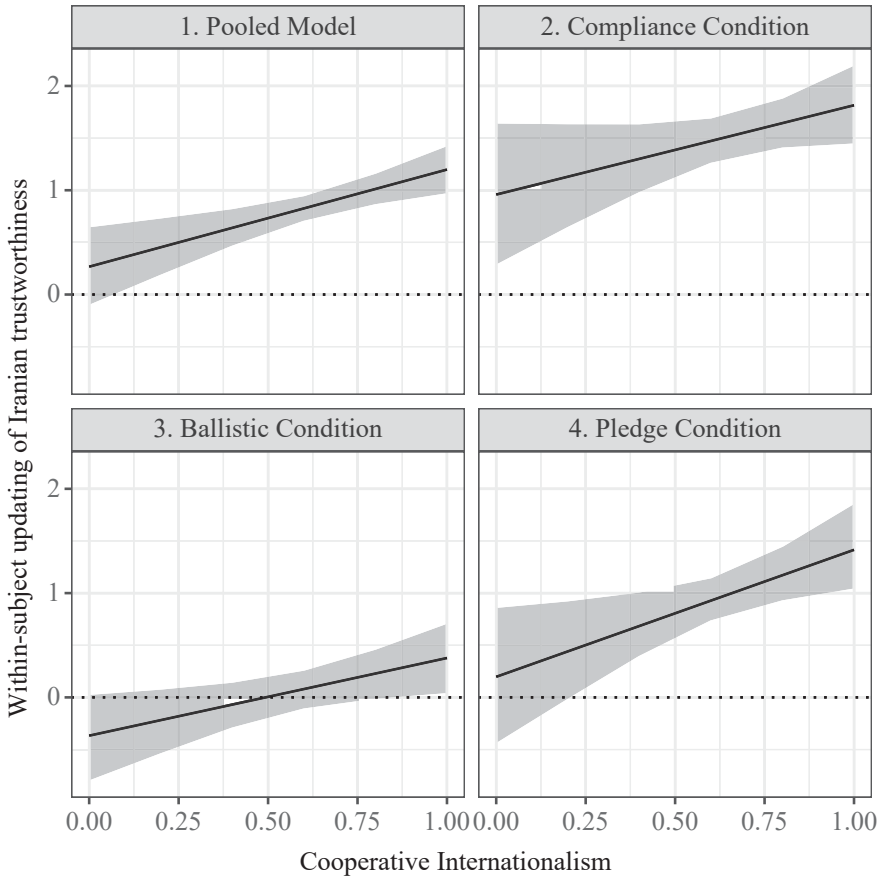

FIGURE 4. *Within-subject effects of signals on Iranian trustworthiness*

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Word limits prohibit us from delving into these supplementary results in detail (see Appendix section 4 for a full analysis), but Figure 4 presents the main results: predicted values from sets of regression models where the dependent variable is the within-subject difference in respondents' perceptions of Iranian trustworthiness before and after receiving a positive signal about Iranian compliance, estimated as a function of each costly signaling treatment, as well as foreign policy orientations like cooperative internationalism. We find, as before, that costly signals work in the aggregate (although the ballistic condition, a mixed signal where Iran complies with the terms of the nuclear deal but also pursues a separate ballistic missile program, has a relatively weak effect), but that there is substantial heterogeneity, with high-CI respondents being more likely to update. Supplementary analyses in Appendix section 4 use this follow-up experiment to test for and rule out an alternative mechanism in which respondents differ not in how they evaluate the signals but in the variance of their priors.

These micro-level findings about motivated reasoning also have important macro-level implications about polarization. If individuals respond to costly signals as motivated skeptics—with the ones who need the signals the least being the ones who update the most—then costly signals have the possibility of inducing polarization rather than convergence. As the violin plots in Figure 5(a) show, the variance of respondents' estimates of Iran's trustworthiness in the first experiment actually significantly *increases* as Iran sends costlier signals of reassurance, especially with respect to the inspections treatments; even though respondents update to costly signals in the aggregate, they do so without converging in their beliefs. Figure 5(b) shows the same patterns also hold in the supplementary experiment; the variance of respondents' prior beliefs about Iran's trustworthiness is always significantly smaller than the variance of their posterior beliefs after Iran sends a costly signal.

Conclusion: Seeing Signals, Charging for Peace

We find evidence both in support of and in opposition to canonical theories of costly signaling. At the aggregate level, experimental participants update as canonical models would expect. On average, signals of reassurance indeed reassure: participants presented with a broader inspections regime, for example, tend to perceive Iran as more trustworthy and less dangerous, and are more supportive of a deal. Our data therefore provide a stronger evidentiary basis for classic signaling models of reassurance. However, consistent with other studies of public opinion,[42] these average effects belie considerable ideological heterogeneity: the respondents who update the most are those who need signals the least. Cooperative internationalists tend to embrace corrective information that bolsters their beliefs, while those low in cooperative internationalism tend to fail to be receptive to signals of reassurance.

42. Page and Shapiro 1992.

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S0020818319000328

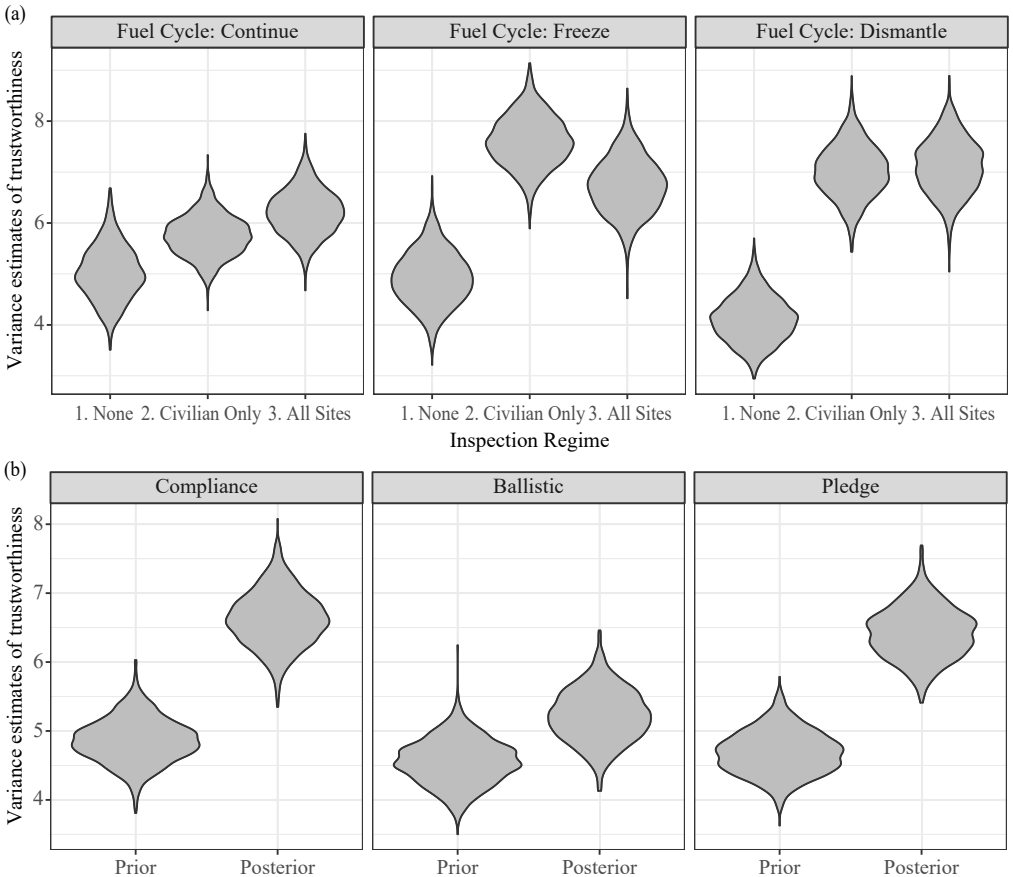

FIGURE 5. *Costly signals induce polarization*

As a result of this heterogeneity, costly signals have important polarizing effects: signals lead to updating, but not convergence.

What are the implications for international relations theory and practice? First, if the individuals who find signals of reassurance the most persuasive are the ones who need to be reassured the least, reassurance should be harder than canonical models would suggest. Second, and relatedly, as Americans and their political leaders become more ideologically polarized, we can expect great gaps in how to deal with important foreign policy problems, divides that are made worse not better by new information coming from the international environment. This implies that even if we observe a process consistent with canonical signaling models on average, the differences across individuals are pronounced enough to lead to significant variation across cases—with enormous consequences. We need look no further than the Iran deal's fate after the election of an administration low in CI and high in MI. Asymmetric updating led to a highly consequential policy turn. Secretary of State Pompeo put our findings into words when he said "We know they're cheating anyway—we're just not seeing it."[43] Despite evidence of Iranian compliance, he did not update his beliefs.

We might also expect that increased polarization of this type is likely to be particularly pernicious: those who have updated their beliefs based on costly signals will believe they have done so in a highly objective, deliberative, and rational manner. Those who have not will be considered irrational and unreasonable. If one side feels that it cannot reason with others based on self-evident facts, the incentives to reach already-difficult compromises on policy further disappears, exacerbating previous polarization.

Third, and relatedly, policymakers should be as aware as possible about the psychological dynamics of their adversary. Our results call for an objectivity about subjectivity. Practitioners must be honest with themselves about how others actually see them—an often painful process that Jervis has long reminded us is a difficult task and empirically rare. In a typical stylized signaling framework, a party interested in cooperation sends a costly signal of reassurance and, given the obvious intent behind the policy concession, can make a judgment about the recipient's cooperativeness based on its response. If it reciprocates, the adversary has underlying cooperative intentions; if it does not, it is belligerent. Our findings suggest that this interpretation would be premature given the tendency for some to engage in belief assimilation. At the same time, our results show that in the aggregate, opinion does shift in the direction that a canonical signaling model expects. This indicates that multiple, perhaps unreciprocated, costly signals of reassurance might be necessary to convince motivated skeptics but that they are capable of updating, albeit at a slower pace. Our findings therefore suggest that successful reassurance may require a series of larger costly

43. Chris Smith, "Trump's Cold War with the C.I.A. Could Derail the Iran Deal," *Vanity Fair*, 6 October 2017, retrieved from <https://www.vanityfair.com/news/2017/10/trumps-cold-war-with-the-cia-could-derail-the-iran-deal>.

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

signals, rather than a graduated process of small, confidence-building measures, and a more forgiving strategy than tit for tat.

Supplementary Material

Supplementary material for this research note is available at <https://doi.org/10.1017/S0020818319000328>.

References

Albuyeh, Rod, and Mark Paradis. 2018. Thawing Rivalries and Fading Friendships: An Experimental Approach to Rapprochement and Alienation. *Political Psychology* 39 (4):811–27.

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

Baekgaard, Martin, Julian Christensen, Casper Mondrup Dahlmann, Asbjørn Mathiasen, and Niels Bjørn Grund Petersen. 2017. The Role of Evidence in Politics: Motivated Reasoning and Persuasion Among Politicians. *British Journal of Political Science* 49 (3):1117–40.

Booth, Ken, and Nicholas J. Wheeler. 2007. *The Security Dilemma: Fear, Cooperation, and Trust in World Politics*. Springer.

Breslauer, George W., and Philip E. Tetlock, eds. 1991. *Learning in US and Soviet Foreign Policy*. Westview Press.

Carson, Austin, and Keren Yarhi-Milo. 2017. Covert Communication: The Intelligibility and Credibility of Signaling in Secret. *Security Studies* 26 (1):124–56.

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.

Fearon, James D. 1994. Signaling Versus the Balance of Power and Interests: An Empirical Test of a Crisis Bargaining Model. *Journal of Conflict Resolution* 38 (2):236–69.

Glaser, Charles L. 2010. *Rational Theory of International Politics: The Logic of Competition and Cooperation*. Princeton University Press.

Gries, Peter Hays. 2014. *The Politics of American Foreign Policy: How Ideology Divides Liberals and Conservatives Over Foreign Affairs*. Stanford University Press.

Herrmann, Richard K., and Michael P. Fischerkeller. 1995. Beyond the Enemy Image and Spiral Model: Cognitive-Strategic Research After the Cold War. *International Organization* 49 (3):415–50.

Hirshberg, Matthew S. 1993. The Self-Perpetuating National Self-Image: Cognitive Biases in Perceptions of International Interventions. *Political Psychology* 14 (1):77–98.

Holsti, Ole R., and James N. Rosenau. 1988. The Domestic and Foreign Policy Beliefs of American Leaders. *Journal of Conflict Resolution* 32 (2):248–94.

Hurwitz, Jon, and Mark Peffley. 1987. How Are Foreign Policy Attitudes Structured? A Hierarchical Model. *American Political Science Review* 81 (4):1099–120.

Janoff-Bulman, Ronnie. 2009. To Provide–Protect: Motivational Bases of Political Liberalism and Conservatism. *Psychological Inquiry* 20 (2–3):120–28.

Jervis, Robert. 1970. *The Logic of Images in International Relations*. Princeton University Press.

———. 1976. *Perception and Misperception in International Politics*. Princeton University Press.

———. 1978. Cooperation Under the Security Dilemma. *World Politics* 30 (2):167–214.

———. 2002. Signaling and Perception: Drawing Inferences and Projecting Images. In *Political Psychology*, edited by K.R. Monroe, 304–405. Erlbaum.

Jervis, Robert, Richard Ned Lebow, and Janice Gross Stein. 1989. *Psychology and Deterrence*. Johns Hopkins University Press.

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Jost, John T., Jack Glaser, Arie W. Kruglanski, and Frank Sulloway. 2003. Political Conservatism as Motivated Social Cognition. *Psychological Bulletin* 129 (3):339–75.

Keohane, Robert O. 1984. *After Hegemony: Cooperation and Discord in the World Political Economy*. Princeton University Press.

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

———. 2017. Microfoundations in International Relations. *Conflict Management and Peace Science* 34 (1):81–97.

Kertzer, Joshua D., Kathleen E. Powers, Brian C. Rathbun, and Ravi Iyer. 2014. Moral Support: How Moral Values Shape Foreign Policy Attitudes. *The 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.

Koremenos, Barbara, Charles Lipson, and Duncan Snidal. 2001. The Rational Design of International Institutions. *International Organization* 55 (4):761–99.

Kunda, Ziva. 1990. The Case for Motivated Reasoning. *Psychological Bulletin* 108 (3):480–98.

Kydd, Andrew H. 2005. *Trust and Mistrust in International Relations*. Princeton University Press.

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 University Press.

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

Mercier, Hugo, and Dan Sperber. 2011. Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences* 34 (2):57–111.

Nincic, Miroslav, and Jennifer M. Ramos. 2010. Ideological Structure and Foreign Policy Preferences. *Journal of Political Ideologies* 15 (2):119–41.

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

Page, Benjamin I., and Robert Y. Shapiro. 1992. *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. University of Chicago Press.

Quek, Kai. 2016. Are Costly Signals More Credible? Evidence of Sender-Receiver Gaps. *Journal of Politics* 78 (3):925–40.

Rathbun, Brian C. 2011. *Trust in International Cooperation: International Security Institutions, Domestic Politics and American Multilateralism*. Cambridge University Press.

———. 2014. *Diplomacy's Value: Creating Security in 1920s Europe and the Contemporary Middle East*. Cornell University Press.

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.

Redlawsk, David P. 2002. Hot Cognition or Cool Consideration? Testing the Effects of Motivated Reasoning on Political Decision Making. *Journal of Politics* 64 (4):1021–44.

Schelling, Thomas C. 1966. *Arms and Influence*. Yale University Press.

Schultz, Kenneth A. 2001. *Democracy and Coercive Diplomacy*. Cambridge University Press.

Shifrinson, Joshua Iskowitz. 2019. A Signal Failure? The End of the Cold War, US Soviet Diplomacy, and Theories of Costly Signaling. Working paper.

Sunstein, Cass R. Sebastian Bobadilla-Suarez, Stephanie C. Lazzaro, and Tali Sharot. 2017. How People Update Beliefs About Climate Change: Good News and Bad News. *Cornell Law Review* 102 (6):1431–44.

Taber, Charles S., and Milton Lodge. 2006. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science* 50 (3):755–69.

Tetlock, Philip E. 1983. Policy-Makers' Images of International Conflict. *Journal of Social Issues* 39 (1): 67–86.

———. 1998. Social Psychology and World Politics. *Handbook of Social Psychology* 4:868–914.

Wittkopf, Eugene R. 1990. *Faces of Internationalism: Public Opinion and American Foreign Policy*. Duke University Press.

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328

Yarhi-Milo, Keren. 2014. *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations*. Princeton University Press.
Yarhi-Milo, Keren, Joshua D. Kertzer, and Jonathan Renshon. 2018. Tying Hands, Sinking Costs, and Leader Attributes. *Journal of Conflict Resolution* 62 (10):2150–79.

Authors

Joshua D. Kertzer is the Paul Sack Associate Professor of Political Economy at Harvard University. He can be reached at jkertzer@gov.harvard.edu.
Brian C. Rathbun is Professor of International Relations at University of Southern California. He can be reached at brathbun@usc.edu.
Nina Srinivasan Rathbun is Professor (Teaching) of International Relations at University of Southern California. She can be reached at nrathbun@usc.edu.

Acknowledgments

We are grateful to David Lindsey, Andrew Little, Jason Reifler, Josh Shifrinson, and the audience at the Norwegian University of Science and Technology for helpful comments and discussions; to Therese Anders for helpful research assistance; and to the *IO* editors and reviewers. This is one of several joint articles by the authors; the ordering of names reflects a principle of rotation.

Funding

Kertzer acknowledges the support of the Niehaus Center for Globalization and Governance at Princeton University.

Key Words

Motivated reasoning; Iran deal; costly signaling; reassurance; political psychology

Date received: July 29, 2017; Date accepted: May 2, 2019

Downloaded from https://www.cambridge.org/core. IP address: 76.217.27.203, on 06 Nov 2019 at 05:54:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0020818319000328