Persuasion and Backlash from International Law in Global Swing States

Stephen Chaudoin and Taegyun Lim

January 08, 2026

Abstract

When one country or international organization makes accusations about violations of international law, the intended audience is often third-party states who might support punishing the offender. When do these accusations persuade publics in those countries and when do they trigger backlash? We show that reactions to accusations about international law violations are consistent with a theoretical model that allows for both types of responses – persuasion and backlash – depending on the audience member's prior beliefs and trust in the information source. We provide evidence from large survey experiments in four global swing states: India, South Africa, Turkey, and Indonesia. The publics in swing states are not strongly aligned with any geopolitical bloc. Persuasion or backlash among those audiences matters greatly, since allegations about international law could tilt their support toward the accuser or the accused. In our survey experiments, when the International Criminal Court makes accusations that Russia violated international law, this persuades and backfires among the theoretically expected subsets of respondents. When the United States makes an identical accusation, it fails to persuade—and often backfires—because publics in swing states hold lower trust in the United States as a credible source. We further show how accusations affect perceptions of the *accuser*, not just the accused. We show a feedback effect, where an accusation can increase or decrease views of the credibility of the information source. Accusations by the ICC improve respondents' views of the Court's credibility, while accusations by the United States further undermine its credibility.

1 Introduction

Accusations that a state has broken international law are a prominent rhetorical argument in international relations.[1] To name just a few, the United States and the International Criminal Court have both accused Russia of committing war crimes in Ukraine. South Africa and a subsequent ruling from the International Court of Justice accused Israel of committing genocide in Gaza. The very first thing the Iranian foreign minister said in condemnation of U.S. airstrikes in 2025 was that they were "a grave and unprecedented violation of… international law."[2] Accusations that a state has violated international law constitute a message from a sender – such as another state or an international organization – to an audience – such as third-party states. The audience then decides whether to change its opinions about or policies toward the accused state.

When a state or international organization (IO) makes accusations about violations of international law, it hopes to convince audiences to help punish the offender. We take direct aim at a fundamental question: are these accusations persuasive, and if so, for whom? When do accusations change third parties' beliefs about the target state's guilt or shift public support for punishing the accused? Additionally, how do accusations shape attitudes about the *accuser* as well as the accused? We focus on audiences in third party states, since a key goal of accusations is to persuade those countries to impose costs on the accused country.

Current scholarship finds a dichotomy between persuasion and backlash.[3] Persuasion occurs when an accusation changes the audience's beliefs about the target's guilt and increases support for punishment. Backlash occurs when the audience shifts their beliefs or support in the opposite direction from that intended by the accuser. We describe a parsimonious theoretical model that accommodates *both* persuasion and backlash. Audience members differ in their prior beliefs about the state of the world and their perceptions of the trustworthiness of an information source. A prominent feature of the model is that it gives clear predictions for both persuasion and backlash among different audience

[1] Morse and Pratt (2022), Morse and Pratt (2025).

[2] "The Latest: US claims strikes on Iran's nuclear sites caused severe damage but full impact unclear," *AP News*, June 22, 2025. https://apnews.com/article/israel-palestinians-iran-war-latest-06-22-2025-7ab46578cb56feecc16f4e4940a46e0a

[3] Chilton and Linos (2021), Madsen et al. (2022).

members. The effects of an accusation depend jointly on the audience member's prior beliefs about the state of the world and perceived credibility of the information source. Persuasion is most likely for an audience member who trusts the credibility of the information source and does not already have prior beliefs that agree with the piece of information sent. Backlash is most likely when the audience distrusts the source but does not already staunchly disagree with the information.

The model also describes how accusations can change audience beliefs about the *accuser*, as well as the accused. Accusations can change audience beliefs about the credibility of the information source itself, in addition to the target. Messages that reinforce the audience member's prior beliefs upgrade her views of the information source, and vice versa.

We show that both persuasion and backlash occur among a critical set of third-party audiences: citizens in "global swing states." By global swing states, we mean countries where public opinion is neither firmly aligned with nor overwhelmingly against "the West" or its adversaries. Swing states are in contrast with bloc members, countries whose polities strongly support the United States and oppose Russia (or vice versa). The citizens of swing states vary greatly in their prior beliefs about the targets of many accusations and trust in the information source making the accusation, like United States officials or prominent international organizations. They are an important place to study persuasion and backlash, because an accusation could be pivotal for moving their citizens from opposing to supporting consequences for the accused, or vice versa. Again in contrast, an accusation against Russia won't change many minds in non-swing states, like Norway or Belarus. Their stances on Russia are firmly entrenched. The swing states of the world could potentially go either way, so the marginal effect of an accusation may be greatest in swing states.

We test the predictions from the model using large survey experiments in four geopolitically important swing states: India, Indonesia, South Africa, and Turkey (N = 6,742). The experiments describe a critical case: how accusations against Russia for alleged war crimes in Ukraine shape public opinion about Russian guilt and support for sanctions or aid to Ukraine. We included pre-treatment measures of respondents' prior beliefs about Russian actions and various measures of trust in the International Criminal Court (ICC) and the United States as information sources. Respondents were

then randomly assigned to receive a prompt with and accusation made by either the ICC or the United States, with a control group receiving no such prompt.

We first examine whether accusations shift citizens' beliefs and policy preferences at both the aggregate and subgroup levels. At the aggregate level, U.S. accusations backfired. Overall, U.S. accusations *reduced* respondents' beliefs that Russia committed war crimes and lowered their support for policies to punish Russia, such as sanctions. In contrast, accusations by the ICC had more muted impacts. ICC accusations had little effect on respondents' beliefs about Russian guilt and their support for punishment.

We then show how aggregate analyses alone obscure how different groups of citizens – depending on their prior beliefs and trust in information sources – update their views because of accusations. When accounting for heterogeneity in prior beliefs about Russian guilt and the credibility of the accuser, we find evidence of *both* persuasion and backlash. The patterns of persuasion and backlash are generally consistent with the predictions of the theoretical model.

Second, we show how accusations alter citizens' trust in information sources. The very act of making an accusation affected respondents' views of the accusers themselves. U.S. accusations diminished respondents' trust in the United States as a credible source of information. In contrast, ICC accusations increased trust in the ICC as a source of information. These effects were also moderated by respondents' prior beliefs, as predicted by the model. Those skeptical of Russian guilt were more likely to downgrade their perceptions of the accusers, and vice versa.

The broader implication of the first set of findings is that accusations from untrusted sources can do more harm than good. Accusations from the United States were more likely to backfire than persuade. While ICC accusations are only persuasive to a certain subset of the audience, they are nonetheless more persuasive than U.S. accusations, with significant differences between the two. The consequences of their weak credibility go beyond feelings toward the U.S or soft power. A lack of credibility affects whether the U.S. can rally foreign support for sanctions and aid for allies - things which are tied directly to the hard power and material consequences of geopolitical conflicts. Foreign governments like the United States would benefit from channeling accusations about international

law through IOs to be more persuasive and avoid backlash.

The broader implication of the second set of findings is to show a feedback effect. Accusations can shape the audience's perceptions of the IO or state making the accusation, which can help or harm that actor's credibility. An accusation today can potentially affect future persuasion by altering trust in the source. Accusations that are well-received could convince the target to support the source's goals and increase their trust in the source, which will make the source even more persuasive the next time it seeks to sway opinions. Accusations that fall on deaf ears lower trust in the source, making their messaging even less effective the next time around. The credibility of the accuser is itself endogenous, shaped by the signals sent by the information source. ICC accusations can potentially build a foundation for trust, which may make it easier to persuade audiences to support its goals. Successes towards its goals likely increases its credibility, which shapes future reactions, making future successes more likely. Conversely, the United States might erode its own credibility even further with poorly-received accusations, making success less likely, and in turn, decreasing its credibility the next time around. Here too, from the perspective of rallying foreign support, it would be better for the United States to avoid bluster when it lacks credibility.

Beyond our specific substantive context, and international relations research more broadly, our approach to modeling audience reactions unites many common arguments in experimental work about heterogeneous treatment effects under a common theoretical framework. Many existing arguments can be classified as arguments about priors about the state of the world,[4] perceptions of sources,[5] or some combination of the two. We show how measuring prior beliefs and trust in an information source can provide direct evidence of the mechanisms underlying arguments about persuasion and backlash. Our approach also directly matches its empirical designs with a theoretical model that yields clear predictions about the heterogeneous effects of messages. Recent work on the effects of IOs and international law has emphasized contestation.[6] Our approach demonstrates a theoretical model and experimental approach for making and testing crisp predictions about how contestation can play out

[4] E.g. Chaudoin (2014).
[5] E.g. Bearce and Cook (2018).
[6] Chaudoin (2016), Morse and Pratt (2022), Morse and Pratt (2025).

across and within important audiences.

Finally, our research highlights the importance of global swing states. Like the opinions of their polities, the policies chosen by swing state governments do not always align fully with one bloc or another. The last decade has seen a resurgence of great power competition, pitting the United States and its allies against Russia and China. The United States has increasingly abdicated its leadership role in the international order, choosing instead to engage in bilateral negotiations or direct coercion on many trade and security issues. States that are less strictly aligned with either bloc face pressure to choose sides, adding to the geopolitical importance of swing states.[7] Understanding the conditions under which messaging from leader states and IOs can persuade swing states will be critical to predicting the future directions of their foreign policies.

2 Information from IOs and Governments

A deep and wide literature on the effects of messages from IOs, naming and shaming, and public diplomacy shares a common structure: a sender tries to persuade an audience. A sender transmits a piece of information to an audience. The sender hopes to change the audience's beliefs about the state of the world and the appropriate action they should take. For example, when an Indian citizen learns that the ICC has accused Russia of breaking international law, the Indian citizen is the audience, the ICC is the sender of this message, and Russia is the target. The citizen has prior beliefs about the true state of the world: whether Russia has broken international law. The citizen also has beliefs about the trustworthiness of the messenger: whether the source's information correctly matches the state of the world. The information she receives potentially changes her posterior beliefs about this state of the world and whether she should therefore support some action, like sanctioning Russia.

A striking feature of this literature is that there are myriad findings where messages lead to persuasion and backlash, with a wide array of proposed factors that moderate the effect of messages.[8] In-

[7] Fontaine and McKinley (2025).

[8] There is of course an even wider literature on signaling and messaging across subfields and disciplines. We focus here on applications most directly related to this setting.

formation that a policy violates international law generally decreases support for that policy.[9] When a specific sender, like an IO or government, transmits information or makes an accusation, this can increase or decrease support for the policy or target of the accusation. For example, information disseminated by IOs can have its intended persuasive effect on public support for human rights, migration policy, military coalitions, or the use of force.[10] However, other studies find that signals from IOs or pertaining to legal standards can trigger backlash, prompting individuals to reject the information or shift their beliefs in the opposite direction.[11] Research on public diplomacy is similarly mixed, with some studies finding positive effects[12] and others finding backlash.[13]

Recent advances in this literature have emphasized the importance of other actors in shaping whether messages persuade. Chaudoin (2023) and Mikulaschek and Parizek (2025) show how an IO's actions are filtered through the media. The former shows how the ICC shifted the content of media coverage of the war on drugs to more greatly emphasize human rights, but it also increased the degree to which contestation about the war received coverage. The latter show how a UNGA resolution condemning the Russian invasion of Ukraine decreased approval of Russian leadership in foreign countries where the media coverage of the UN was one-sided. In other countries, it had a more muted effect because of the uneven quantity and content of coverage across countries. IOs also now use new media communication tools to attempt to reach mass audiences.[14] Accusations are also just the first steps in a longer dance of messaging and counter-messaging between accusers and the accused.[15] Accused governments can avoid domestic political costs with image management.[16] Morse and Pratt (2025) measure support for punishing a country accused of violating international law in a sample of U.S. respondents and global elites. Retorts by the accused generally decrease respondents' willingness to punish them, while IO rebuttals blunt some of these retorts.

[9] Eg Wallace (2013). See Chilton and Linos (2021) for a summary.
[10] Anjum, Chilton, and Usman (2021), Mikulaschek (2023), Recchia and Chu (2021), Suong, Desposato, and Gartzke (2024).
[11] For example, Madsen et al. (2022), Cope and Crabtree (2020), and Efrat and Yair (2023).
[12] Eg Goldsmith, Horiuchi, and Matush (2021), Wang et al. (2023), Choi et al. (2023).
[13] Eg Goldsmith and Horiuchi (2009), Rhee, Crabtree, and Horiuchi (2024), and Mattingly and Sundquist (2023).
[14] Carnegie, Clark, and Fan (2024).
[15] Zvobgo (2019), Chow and Levin (2024).
[16] Morse and Pratt (2022).

Our study focuses on arguments about moderators. Existing work gives many explanations for things that can moderate – meaning magnify, mute, or reverse – the effects of signals from IOs and governments. Certain audience characteristics could make them more or less responsive to certain signals. One set of moderators emphasizes how audiences triangulate their own preferences with those of the signal sender with respect to the issue at hand. The trustworthiness of a signal arises from its relation to the underlying preferences (or "type" or incentives) of the signal-sender. If the sender and audience's preferences are aligned, the audience trusts the signal more. For example, UN Security Council authorization for a use of force can persuade audiences who worry that the use of force is hegemonic adventurism, and the signal sent by authorization is particularly trustworthy when it diverges from usual inclinations of powerful member states.[17] Conversely, when the signal sender is the "wronged country" complaining about another country's actions, this is less persuasive, since the wronged country has an incentive to overstate its grievance.[18]

Similarly, the audience's broader, overall disposition towards international law, IOs, or international relations also make signals more or less persuasive. If the audience has favorable attitudes or confidence in the sender, then they are more responsive to its signals.[19] Audiences with "cooperative internationalist" dispositions may be more receptive to signals advocating for cooperation, multilateralism, or international law.[20] Other studies highlight the importance of relational dynamics between the sender and the audience. The effects of a message depend on geopolitical alignment and perceptions of the sender as part of the in-group or out-group.[21] Both ideology and partisanship are correlated with preferences over specific policies and general dispositions towards different messages and messengers. They, too, can also moderate the persuasiveness of a signal. For example, Cope and Crabtree (2020) find that prompts about international law obligations regarding refugees backfire in Turkey, especially among incumbent party supporters.[22]

[17] Eg Chapman (2007) and Fang (2008). See also Thompson (2006) and Thompson (2015). Mikulaschek (2023) shows how unanimity makes these signals particularly persuasive.

[18] Cohen and Powers (2024).

[19] Bearce and Cook (2018), Anjum, Chilton, and Usman (2021), Grieco et al. (2011).

[20] Kertzer, Rathbun, and Rathbun (2020).

[21] Chu (2025), Terman (2023), Pauselli (2023).

[22] See Brutger (2021), Chaudoin (2023), and Kertzer, Rathbun, and Rathbun (2020) for other examples of partisan moderation.

Few of these arguments are mutually exclusive, and there is substantial overlap between many of them. This overlap is theoretical: they describe similar concepts. These moderators describe the credibility of a messenger, which conditions how the audience reacts. These arguments diverge in the microfoundation they provide for why the message is considered credible when delivered by a particular source. The overlap in moderators is also empirical: in observed data, many of these attributes are correlated.

The audience's prior beliefs about the state of the world also moderate the effects of a message. If the audience already agrees with the message, then its *marginal* effect is minimal. The sender is "preaching to the choir." If the audience already disapproves of the target, then condemnation about its illegality cannot push approval further downward. For example, Chaudoin (2014) finds that treatments about a possible WTO dispute have the largest effect for those who are neither strongly supporting of or opposed to free trade, *ex ante*. Spilker, Nguyen, and Bernauer (2020) also find that the effect of new information about a trade agreement is less impactful for those with stronger priors.[23]

Though many arguments about prior beliefs or credibility are not mutually exclusive, the direction of their moderation effects are often in tension. Consider whether someone holds a favorable view of the United Nations or holds cooperative internationalist (CI) foreign policy attitudes. If an IO says "the target violated international law and should be sanctioned," then these moderators have conflicting effects on the overall effect of the message. Both moderators may make the audience more trusting of a message from an IO, which should magnify the persuasive effects of the message. But that same audience member is also likely to *already* believe what the IO is telling them and condemn the behavior in question, which blunts the persuasive effects of the message. Their net effect is theoretically unclear. Which effect dominates would be difficult to know *ex ante*.

The theoretical model below contributes to this literature by giving a framework that can explain how messages can generate both persuasion and backlash among audiences.[24] When the audience receives information about an accusation, the effect on their posterior beliefs about the accusation

[23] Búzás and Bassan-Nygate (2024) and Cope (2023) also make floor or ceiling effect arguments. Arias et al. (2022) show this for malfeasance audits.

[24] We certainly do not claim to be the first to argue that prior beliefs or perceptions of sources matter. See for example Lupia and McCubbins (1998).

depends *jointly* on their prior views and the perceived trustworthiness of the information source. It therefore builds on arguments about the aggregate effect of accusations by making clear predictions about the heterogeneity in how audiences respond. We also show how this framework is versatile. It accommodates many of the disparate mechanisms for heterogeneous treatment effects described above. Our framework shows how to make precise predictions about the moderating effects of these attributes and how measuring both is necessary to test many arguments.

Furthermore, the theory shows how accusations reshape perceptions of the messenger, not just accused. We build on and apply a growing literature of models of persuasion, where audiences update their beliefs about information sources, in addition to the state of the world.[25] When accusations are perceived as credible, they enhance trust in the sender over time, creating a positive feedback effect. When perceived as biased, they erode trust and diminish the sender's future persuasive power. In the context of international relations, the trustworthiness of a messenger is an endogenous attribute, shaped by the messages they have previously sent. Brutger and Strezhnev (2022) and Chung (2025) show that information about disputes involving a respondent's country can erode public attitudes toward the IO associated with the dispute. Relatedly, a large body of research examines the features of IO legitimacy.[26] Institutional characteristics – encompassing both procedural and performance-based qualities – can broadly shape public perceptions of the legitimacy of IOs. We show how the trustworthiness of an IO's messages is affected a function of what it has previously said.

Among all possible third-party audiences, we focus on those in "global swing states." Since the onset of the Cold War, there have loosely been two blocs: the West (the United States, Western Europe, and like-minded democratic allies such as Japan and South Korea) and its adversaries. The West's primary adversaries have shifted in identity and salience, but they have generally included the major autocratic countries, Russia and China.[27] Global swing states are those where public opinion is generally more balanced in support or opposition to one bloc or the other.[28] Countries can differ in the

[25] Cheng and Hsiaw (2022), Gentzkow, Wong, and Zhang (Forthcoming). We are not aware of any existing applications to international relations.

[26] Lisa Maria Dellmuth, Scholte, and Tallberg (2019), Lisa M. Dellmuth and Tallberg (2021), Ecker-Ehrhardt, Dellmuth, and Tallberg (2024), Ghassim (2024).

[27] The term "third world" originally referred to countries neither aligned with the U.S. nor Soviet bloc.

[28] To draw an analogy with United States politics, (U.S.) states like Alabama and Massachusetts are not swing states;

degree to which they swing and this can, of course, vary across issue areas. But a handful of countries emerge as consistently inconsistent. The most commonly mentioned global swing states are: India, Brazil, Indonesia, Turkey, and South Africa. Each may lean towards supporting the West, but they do not follow lock-step.[29]

Global swing states are important for studying messaging, in particular, because their polities consist of audience members with a wide range of views on any particular issue. Their citizens' prior beliefs about issues like uses of force and international law are more dispersed than in countries where strong majorities already hold strong views. Their citizens' views are less monolithic about the state of the world and trustworthiness of many messengers. Swing states take on greater geopolitical importance by not being perfectly aligned with any bloc of countries. As hegemons or bloc leaders build coalitions to pursue their goals, they understandably focus a large amount of effort and attention on persuading swing state countries to choose their side.[30]

3 Theory

We will continue with the example of an accusation about Russian war crimes for simplicity and to match the subsequent experimental setup. We assume there is a binary state of the world that is unknown to the audience. Denote the state of the world as $S \in \{0, 1\}$, where $S = 1$ describes a situation where the accused is guilty. They have, in fact, committed war crimes. $S = 0$ denotes that they aren't guilty.[31] Each audience member, *i*, believes that the state of the world is drawn from a Bernoulli distribution, where the probability that $S = 1$ is $\pi_i \in [0, 1]$.

The audience members all receive a common signal about the state of the world from a source. Let s_1 indicate that the source has sent a signal that $S = 1$, i.e. the source says "Russia is guilty." Audience

their voters overwhelmingly support one party or the other. Pennsylvania is a swing state because support for the two parties is more balanced.

[29] We detail their stance on Russia and Ukraine in the experimental section below.

[30] Fontaine and Kliman (2013), Fontaine and McKinley (2025). To continue the analogy to U.S. politics, campaigns focus much more of their resources on swing states, rather than fighting losing battles in opposition strongholds or wasting money on states they know they will win.

[31] Note, this can incorporate "guilt" meaning "the accused did the act and it is illegal" and innocence as "they didn't do it" *or* "they did it, but it wasn't illegal.'' Our framework fits with either.

members have heterogeneous prior beliefs about the trustworthiness of the source: $\sigma_i = Pr(s_1|S = 1) = Pr(s_0|S = 0)$. In other words, σ_i denotes the audience member's prior beliefs that the source will send a signal that correctly matches the state of the world. For individual i, her priors are that σ_i is distributed according to a Beta distribution with parameters α_i and β_i.[32]

We are interested in the *treatment effect* of a signal, s_1, on the audience member's posterior beliefs about two things: (1) the state of the world and (2) source trustworthiness. Applying Bayes rule, her posteriors about the state of the world are $Pr(S = 1|s_1) = \frac{\pi_i \alpha_i}{\pi_i \alpha_i + (1-\pi_i)\beta_i}$. Her posteriors about the source have a Beta distribution.[33] We define the treatment effect for the state of the world for audience member i is: $\Pi_i = \Pr(S = 1|s_1) - \pi_i$. The treatment effect for source trustworthiness is: $\Sigma_i = \mathbb{E}[\sigma_i|s_1] - \mathbb{E}[\sigma_i]$. We use capital Greek letters to denote treatment effects.

Defining treatment effects in these ways is critical, because they describe how posteriors move relative to the audience member's priors. In other words, we want to think about the difference between her priors and her posteriors, not just her posteriors. This has a natural mapping to experimental work about the effects of signals. We want to compare beliefs in a world where the audience member receives a signal, compared to a control condition where she does not receive a signal. In the latter case, without any signal, her posteriors are simply her priors.[34] In a between-subjects experimental design, researchers compare posteriors from a group that has been treated with some piece of information to a control group that has not received it, and therefore retains their prior beliefs. In a within-subject design, researchers analyze the aggregate differences between pre-treatment (prior) beliefs and post-

[32] Beta distributions are bounded between zero and one. They also have an intuitive link to prior beliefs about source quality. The expectation of source trustworthiness or accuracy for an audience member is the proportion of signals from that source that correctly match the state of the world: $\mathbb{E}[\sigma_i] = \frac{\alpha_i}{\alpha_i + \beta_i}$. This is equivalent to an audience member who counts up the number of times the source has been correct in the past (α_i) and the number of times the source has been wrong (β_i), and then uses the proportion of correct past signals as her prior for beliefs about source accuracy.

[33] The fact that her posteriors about σ are distributed Beta follows from the Beta-Bernoulli conjugacy. The expectation of her posteriors about source trustworthiness are $\mathbb{E}[\sigma_i \mid s_1] = \Pr(S = 1 \mid s_1) \cdot \frac{\alpha_i + 1}{\alpha_i + \beta_i + 1} + \Pr(S = 0 \mid s_1) \cdot \frac{\alpha_i}{\alpha_i + \beta_i + 1}$. Proofs for all derivations are in Appendix A.

[34] It is worth noting that this description of a treatment effect does not capture updating in the absence of a signal. In other words, the audience member does not say "I haven't heard the source say anything about the state of the world, and the absence of that information is itself informative about the state of the world." Our exclusion of this directly matches experimental settings, where the researcher can strictly control the absence of a signal and the respondent does not know the information she could have received but did not receive. It is also a good first approximation of what happens outside of an experimental laboratory. News consumers have their preferred outlets and they consume the news that source provides to them each day. But they are not often thinking about all of the articles the source could have chosen to write but did not write.

treatment (posterior) beliefs. In other words, our theoretical definitions of "treatment effects" match exactly the quantity of interest that is implied by the vast majority of experimental applications.[35]

3.1 Treatment effects for the state of the world

Figure 1 shows the predicted treatment effects for posteriors about the state of the world (Π_i). The horizontal axis shows respondent i's prior beliefs about source trustworthiness. The vertical axis show her prior beliefs about the state of the world. For each cell in the plot, we calculate Π_i and the heatmap shows the magnitude and direction of the treatment effect.[36] Blue cells on the right hand side indicate persuasion, where $\Pi_i > 0$. The audience member's posteriors have moved in the sender's intended direction. The bottom right quadrant is where persuasion is most powerful. Individuals in the quadrant didn't think Russia was guilty and they trust the source of the signal. They therefore show the greatest positive movement from their priors to their posteriors. The top right quadrant shows a ceiling effect, where the sender is "preaching to the choir." These individuals trust the signal, but they already thought Russia was guilty, so their posteriors are only a small increase over their priors. Red regions indicate backlash, where $\Pi_i < 0$. These individuals think the signal is worse than uninformative. They think the signal should be interpreted in the opposite direction from the sender's intended effect. In the top left, these individuals thought Russia was guilty but distrust the signal, so they are the most "dissuaded" to believe the sender. In the bottom left, the sender's signal has "fallen on deaf ears." They distrust the signal, but there is a floor effect because they already didn't think Russia was guilty.

[35] For a comparison between our model and motivated reasoning models, see Appendix A.

[36] This is equivalent to the treatment effect if we simulated many respondents in each cell, randomly assigned them to treatment or control, and then regressed their posterior beliefs on an indicator for treatment assignment. It other words, the heatmap also shows the predicted regression coefficient desired in empirical studies.

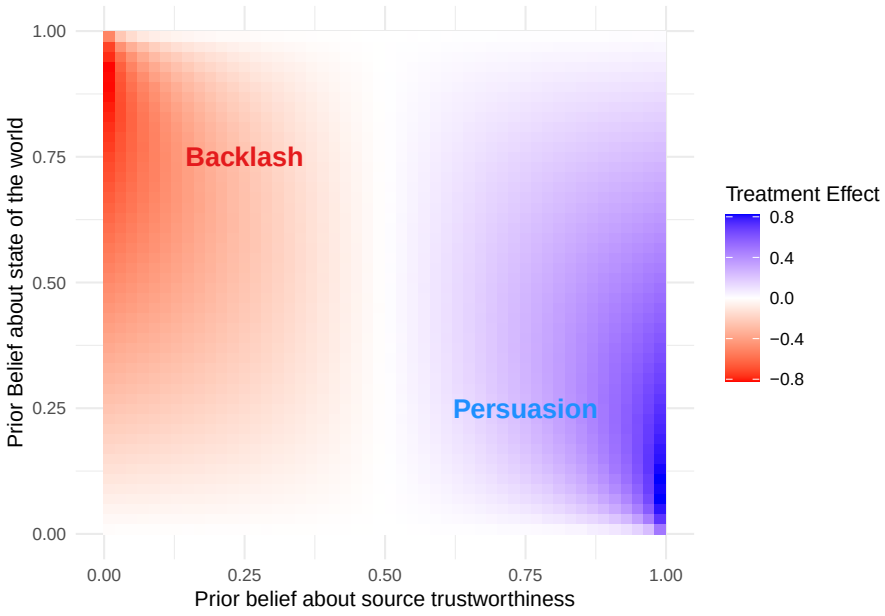


Figure 1: Predicted treatment effects on beliefs about the state of the world.

If audience members update their beliefs in a manner consistent with the model, then the magnitude and direction of belief change depend on both their prior beliefs and their trust of information sources. Some will be persuaded and others will move in the opposite direction to the signal. Our hypothesis summarizes the predictions shown in Figure 1.

H1 (Treatment effects on posteriors about the state of the world):

H1a (Persuasion): Accusations will be most persuasive for individuals with high trust in the source and who do not already have prior beliefs that are strongly aligned with the signal.

H1b (Backlash): Accusations lead to backlash the most for individuals with low trust in the source and who have prior beliefs that are strongly aligned with the signal.

The difficulty of testing aggregate hypotheses, like "the signal persuades," without measurements of priors about the state of the world *and* source trustworthiness is apparent in Figure 1. If the sample included people evenly spread across all four quadrants, and a researcher regressed posterior beliefs on whether the individual got the signal, the coefficient would equal zero, *despite these individuals responding exactly as predicted in the model.* We might conclude that the signal was ignored, even if every individual was a "complier" with the treatment and reacted in the exact way predicted by the model.

Figure 1 also shows why prior beliefs and trustworthiness matter *jointly* in predicting and assessing moderation effects. Varying prior beliefs generally has a non-monotonic effect on the magnitudes of predicted treatment effects. For a given perception of trustworthiness, the treatment effects increase and then decrease when we move from the bottom of the figure to the top. And the gradient of the treatment effect as we vary priors also depends on perceptions of trustworthiness. On the left hand side (signal is trustworthy), moving from bottom to top makes treatment effects more negative, and then less negative as priors converge to the limit. On the right hand side (signal is untrustworthy), the opposite is true.

Going from left to right in the figure, increasing trust of the signal unambiguously raises the treatment effect. But the magnitude of this change depends greatly on prior beliefs. When the receiver believes the target to be guilty (upper half), increasing trust changes the treatment effect from strongly negative to weakly positive. When the receiver believes the target to be innocent (lower half), increasing trust changes the treatment from weakly negative to strongly positive. The top right of Figure 1 is the canonical "ceiling effect", where treatments matter less because the receivers priors are already aligned with the message. The bottom left is the canonical "floor effect."

As we describe more extensively below, our empirical approach will aggressively measure respondent priors and their perceptions of information sources. This enables us to examine reactions across the prior and trustworthiness space, and compare treatment effects with a concrete prediction for each type of respondent. Figure 1 also makes it apparent why some moderators from existing work have ambiguous implications. Scoring higher on a cooperative internationalist (CI) scale, for example, might make a respondent more trusting of an IO's information, but it can also blunt treatment effects if it means that the respondent already believes what the IO is telling them.[37]

3.2 Treatment effects for source trustworthiness

The signal also affects beliefs about the trustworthiness of the *source* itself. The impact of information extends beyond its immediate persuasive effects, influencing long-term perceptions of the source it-

[37] We show this ambiguity theoretically and empirically in the appendix.

self. When the source provides information that matches the audience's priors, they audience is more likely to increase their support for and trust in that source. Conversely, when information conflicts with their priors, not only is the content of the message rejected, but this leads to a further erosion of trust. This loss of credibility can have significant downstream effects, diminishing the source's ability to shape public opinion or mobilize support in future interventions.

Figure 2 shows these predicted treatment effects for source trustworthiness (Σ_i).[38] The treatment effects are monotonically increasing in the audience's prior beliefs that "guilty" is the state of the world. The intuition is that, if an audience member starts more convinced about the state of the world, and they receive a signal that comports with that prior belief, then they update favorably about the quality of the source. "If the source tells me what I think is already true, then I trust the source more," and vice versa. The contours of the treatment effects are different from the predicted treatment effects about the state of the world. In the bottom right region, the audience member says "I thought you were trustworthy, but then you told me something that really contradicts my priors, so I lower my beliefs about your trustworthiness." In the upper left, the audience member says "I thought you were a terrible source, but then you said something that matched my priors, so I upgraded my beliefs about you as a source." Hypothesis 2 describes these key features of the predicted treatment effects that we analyze empirically below.

[38] The expression is: $\Sigma_i = \frac{\alpha_i}{\alpha_i+\beta_i+1} \cdot \left[1 + \frac{\pi_i}{\pi_i\alpha_i+(1-\pi_i)\beta_i}\right] - \frac{\alpha_i}{\alpha_i+\beta_i}$. See Appendix A.

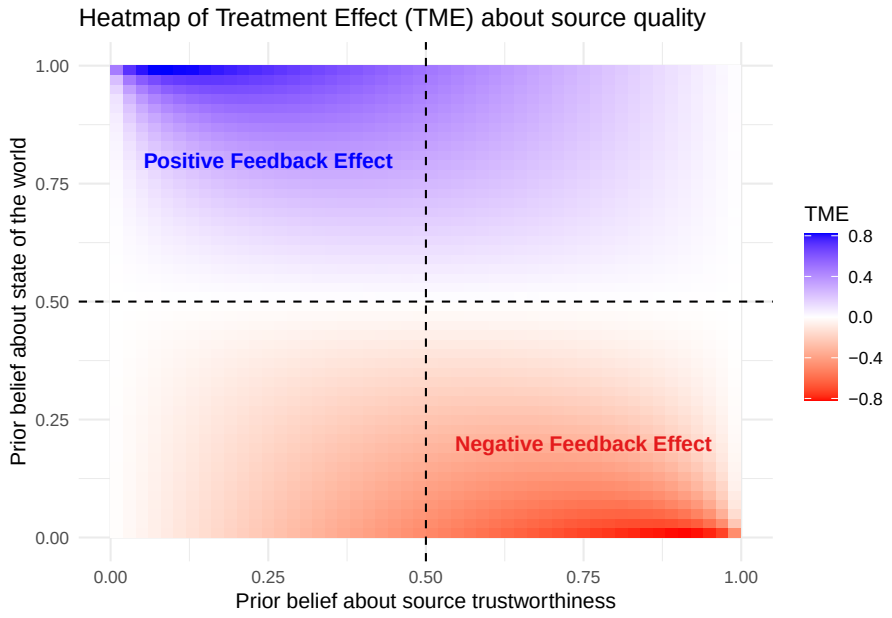


Figure 2: Predicted treatment effects on beliefs about the trustworthiness of the source

H2 (Conditioning effect of priors about the state of the world): The effect of an accusation on beliefs about the trustworthiness of the source are increasing in the individual's prior beliefs that "guilty" is the state of the world.

The model shows the possibility that IOs or diplomacy can create a positive feedback effect, if the public has at least some level of trust in the information source. Over time, sending information that matches the audience's priors enhances the source's ability to persuade and shape public opinion, even if short-term effects are more limited. An information source can become more credible with consistent engagement that bolsters their legitimacy and long-term persuasiveness.

4 Experimental Design

4.1 Background and sample

We chose accusations against Russian war crimes for the context of our survey because it represents a watershed event in which global swing states have played a critical role. In March 2023, the International Criminal Court (ICC) issued arrest warrants for Russian President Vladimir Putin, alleging

responsibility for war crimes committed during Russia's invasion of Ukraine in 2022. These accusations included the unlawful deportation of children from Ukraine, which the ICC classified as a violation of international law. Foreign governments, especially the United States, also condemned Russia for numerous instances of war crimes in Ukraine, ranging from indiscriminate attacks on civilian infrastructure to documented atrocities such as those committed in Bucha and Mariupol.

We conducted survey experiments in four global swing states— India (N = 1,704), Indonesia (N = 1,672), South Africa (N = 1,702), and Turkey (N = 1,664) —in collaboration with TGM Research in October 2024.[39] Public opinion in the countries we chose does not staunchly align with or against the Russia and the United States. Figure 3 uses information from the 2021 World Gallup Poll surveys to show where these countries generally lie in their attitudes towards Russia and the United States.[40] The vertical axis shows the mean number of respondents indicating that they disapprove of the leadership of Russia. The horizontal axis shows approval of U.S. leadership.[41] Each country is located toward the middle of the plot. None of our countries is on the extremes of both dimensions.

Reactions of swing states are especially important, since they play a critical role in forming coalitions to implement punishment against Russia. Additional condemnation from France or support from Iran is largely irrelevant for Russia. Russia received condemnation or support from those countries before ICC accusations and continued to do so afterwards. Condemnation from a swing state like India, however, has much larger implications. For example, the effectiveness of economic sanctions on Russia largely depends on these states' willingness to enforce trade restrictions, limit access to financial systems, or reduce dependency on Russian energy exports.[42] Without the participation of swing states, sanctions are more easily circumvented, weakening their impact.

In terms of government policies, these countries have also typified the "non-aligned" stance of countries that oppose Russia in some, but not all, ways. Table 1 summarizes each country's stance

[39] TGM recruits from existing panels of online respondents. Their samples reflect national distributions, in terms of demographics. Respondents passed an initial attention check before proceeding to the survey.

[40] The plot shows the top 25 countries by population in 2021.

[41] There is not a specific question about Russian war crimes, but approval of the leadership is likely correlated with views on the war in Ukraine. Views of U.S. leadership are also likely correlated with views on U.S. trustworthiness.

[42] Since our survey, these states have taken on even greater importance with respect to sanctions on Russia. The U.S. has threatened India and Brazil with huge tariffs in an attempt to curb their oil imports from Russia. See "Putting maximum pressure on Russia requires secondary sanctions on oil" Washington Post, August 2, 2025, for example.

on some of the key issues surrounding Ukraine. None of the four countries actively participate in the sanctions regime against Russia, though Indonesia stopped some arms imports from Russia and replaced them with French suppliers.[43] India and Turkey agreed to provide humanitarian or military aid to Ukraine, but their allocated amounts were the smallest among donors even after accounting for differences in economic size.[44] Each of the four countries either voted in favor of or abstained from the 2022 UN General Assembly resolutions condemning the latest Russian invasion of Ukraine.[45] The countries also took tepid, generally non-committal stances on the ICC's arrest warrant for Putin. South Africa, especially, has tread carefully about the arrest warrants, since their ICC membership legally obliges them to arrest Putin should he visit the country.

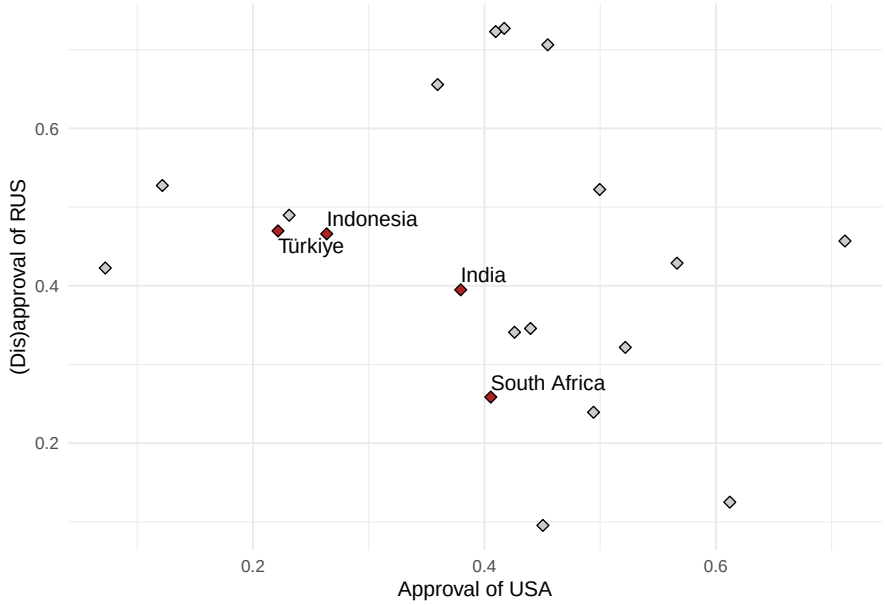


Figure 3: . Country-level attitudes towards Russia and the United States, Gallup 2021.

4.2 Pre-Treatment Measures

Pre-treatment, we measured respondents *prior beliefs* about whether Russia had violated international law. Our survey item read "Countries sometimes violate international laws of war that restrict attack-

[43] Chivvis, Noor, and Geaghan-Breiner (2023).

[44] Among 40 donors, India ranked 40th in assistance as a share of GDP (USD 3.5 million, less than 0.001%) and Turkey ranked 38th (USD 83.1 million, 0.01%). Source: IFW Kiel Ukraine Support Tracker.

[45] United Nations General Assembly. Aggression against Ukraine. Resolution adopted at the 11th Emergency Special Session, A/RES/ES-11/1, 2 March 2022. https://digitallibrary.un.org/record/3959039.

Country	Aid to Ukraine	Sanctions Regime Participation	UNGA 2022 Resolutions	ICC Arrest Warrant
India	Yes, but limited*	No	Abstained	Non-member, no public stance
Indonesia	No	No	Voted in favor	Non-member, no public stance
South Africa	No	No	Abstained	Member, mixed/critical public stance
Turkey	Yes, but limited*	No	Voted in favor	Non-member, no public stance
*Among 40 donors, India ranked 40th in assistance as a share of GDP (USD 3.5 million, <0.001%) and Turkey ranked 38th (USD 83.1 million, 0.01%).				

Table 1: Positions of selected countries on Ukraine-related issues.

ing civilians and other acts. In your opinion, what is the percent chance that the countries below have violated international laws of war over the last 5 years?" Respondents chose from a sliding scale from 0-100. They answered for Russia, the United States and China.[46]

We also included three items that measure respondents' perceptions of the trustworthiness of a particular source of information. The first item asked *"There are many sources of information about international affairs. Some sources of information are trustworthy and others are not. On a scale of 1-100, with zero being the least trustworthy and 100 being the most trustworthy, where would you place the following sources of information?"* Respondents answered for the United States Government, the ICC, and the media.[47] The second item read *"[Countries/international organizations] criticize each other. Sometimes they are telling the truth and other times they have another motive. In your opinion, what is the percent chance that these [countries/international organizations] are telling the truth when they criticize another country?"*. Respondents again chose on a scale from 0-100. The list of countries included the United States, China, and France. The list of IOs included the ICC, the WHO, and the EU. Third, we used a simple feeling thermometer for the United States and the ICC.[48]

[46] We randomized the country order. Including the United States and China helps make this item not solely focused on Russia. We used the "percent chance" language since it has been used in previous studies conducted internationally that were focused specifically on measuring probabilities (Delavande (2014)).

[47] Again, we randomized the sub-items for this and all subsequent questions where applicable, and we included the media to avoid focusing solely on the two actors of interest.

[48] The item read *"We'd like to get your feelings toward certain countries and international organizations on a 'feeling thermometer.' A rating of zero degrees means you feel as cold and negative as possible. A rating of 100 degrees means you feel as warm and positive as possible. You would rate the country or organization at 50 degrees if you don't feel particularly positively or negatively*

In practice, for both sources of information (the ICC and the United States), all three measures of prior beliefs about the source — trustworthy, telling truth, and feeling thermometer—are strongly correlated. For the ICC, pairwise correlations range from 0.67 to 0.76, and for the United States, they range from 0.71 to 0.74. Given this high degree of internal consistency, we construct a single index for each source by taking the simple average of the three measures.[49]

4.3 Treatment and Outcome Measures

Respondents assigned to the control group read the following sentence - *"As you may or may not know, Russia invaded Ukraine in 2022."* Respondents assigned to the U.S. or ICC treatment groups read the control group sentence, followed by an additional declaratory statement: *"The [United States/International Criminal Court] has accused Russian leaders of committing war crimes during the invasion."* We chose this treatment design because it is simple and minimal: the only new information it conveys is that a particular source has made an accusation.

We selected the United States as the individual state treatment because it is widely regarded as one of the most influential countries in the world, both in material or soft power. The potential effects of its accusations are intrinsically important. Among international organizations that condemned Russia, we selected the ICC for the IO treatment because it is an independent legal institution, distinct from the signaling of individual states. The ICC opened a formal investigation in 2022 and issued arrest warrants for President Putin and another offical in March 2023. Unlike other high-profile IOs such as the UN or the EU—whose statements often reflect the collective positions of member states—the ICC operates through independent legal bodies, including a judiciary and an Office of the Prosecutor, which are not strongly associated with any single national interest. This makes the ICC an important contrast for isolating the effects of signaling from an ostensibly neutral, norm-enforcing IO versus an individual state actor.

We included three types of outcome measures: the respondents' posterior beliefs about Russian guilt, their preferences over policies toward Russia that their country could adopt, and their beliefs

toward them. How do you feel about following countries or international organizations?". The list also included Russia and Israel.

[49] For summaries of the distributions of responses, see Appendix.

about information sources themselves.[50] For posterior beliefs about Russian guilt, we asked *"How likely is it that Russian leaders have committed war crimes in Ukraine?"* and respondents used a 100-point scale.

For policy responses the respondent's government could adopt, we asked three agree/disagree questions about whether the respondent's government should: (1) *"impose sanctions on the Russian government, companies, and individuals?"*, (2) *"provide non-military aid to Ukraine?"* and (3) *"provide military aid to Ukraine?"*. Respondents chose from a five-point scale (Strongly agree/disagree, somewhat agree/disagree, neither agree nor disagree).

The overall level of support for the policy responses in each country was consistent with our characterization of them as swing states. If we assign numerical values to the 5-point agreement scale, 1-5, the mean of the responses across all four countries was 3.3 for non-military aid, 2.9 for military aid, and 3.0 for sanctions.[51] Indian respondents had the strongest support for non-military and military aid (means of 3.4 and 3.2 respectively). South Africa had the strongest support for sanctions (mean of 3.2). Indonesia had the lowest means for all three policies (3.3, 2.6, and 2.8).

To assess post-treatment beliefs about information sources, we measured two outcomes: trust in the information source and perceived legitimacy of the source. For trust, respondents were asked: *"Some sources of information are biased and others are not. On a scale from 0 to 100, with 100 being the most biased, where would you place the following sources of information?"* They evaluated the International Criminal Court (ICC), the U.S. government, and the news media. We then reverse-coded these bias ratings to construct a trust measure, so higher values indicate greater trust of information sources.[52]

[50] We randomized the order of these items across respondents, following Chaudoin, Gaines, and Livny (2021).

[51] These numbers are calculated from control group respondents, since these measures were post-treatment. See appendix for country breakdowns.

[52] Before turning to results, we would note that this experiment was not pre-registered. For readers who are, understandably, vigilant about mining for heterogeneous treatment effects (HTE), we would note two interrelated things. First, the analysis of HTE that follows is derived directly from our formal model. The theoretical model gives predictions about how prior beliefs about the state of the world and trustworthiness of sources should shape treatment effects. Second, the survey instrument itself was sparse. It focused on measuring priors and trustworthiness and did not include an extensive buffet of possible moderators from which we could search for HTE. Appendix B describes the entire survey instrument, summary statistics, and balance assessments.

5 Results

We first analyze the aggregate effect of treatment on beliefs about Russian guilt and support for policy responses. We then show how heterogeneous treatment effects are consistent with our theoretical predictions (Hypothesis 1). We then analyze the aggregate and heterogeneous effects of treatment on perceptions of the information source (Hypothesis 2).

5.1 Treatment effects on aggregate posterior beliefs and policy support

Did U.S. and ICC accusations influence aggregate opinions about Russian guilt and possible governmental responses? The left pane of Figure 4 shows the effect of the U.S. and ICC treatment on agreement with the statement that Russia committed war crimes. For these estimates, we regressed (OLS) responses to the question about whether Russia committed war crimes on an indicator for which treatment the respondent received. The estimates compare a particular treatment group to the control group, excluding the other treatment group.[53] Aggregate treatment effects are generally modest. The aggregate effect of the U.S. treatment is actually *negative*. Overall, the U.S. accusation moved respondents' beliefs in the unintended direction. The ICC treatment effect is also negative, but it is very close to zero.

The right pane of Figure 4 describes the difference in the two treatment effects.[54] While the ICC treatment did not generate a statistically significant change on its own, public belief in Russian war crimes under the ICC treatment is at least significantly less prone to backlash compared to the U.S. treatment. The right panel of the figure illustrates this contrast, showing that beliefs about Russian war crimes are stronger among those exposed to the ICC treatment than among those who received the U.S. treatment.

Figure 5 shows the effects of treatment on the downstream policy responses: support for non-military aid to Ukraine, military aid, and the sanctions regime.[55] The results are very similar, with the

[53] These specifications have a single intercept. Results are very similar with country-specific intercepts and other alternate specifications. See appendix.

[54] These estimates describe the effect of the ICC treatment, excluding control observations, i.e. they compare the ICC and U.S. treatment groups.

[55] The estimates are from the same regressions as above, just with different outcome measures.

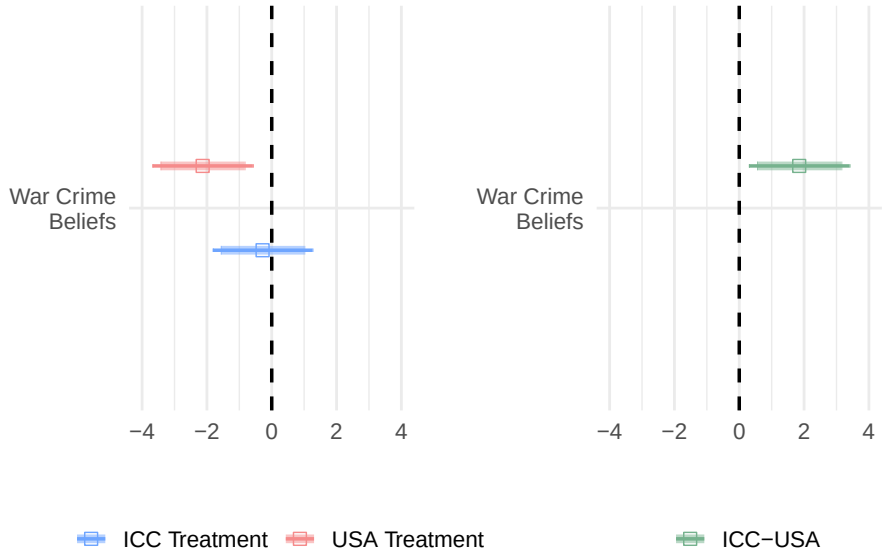


Figure 4: Aggregate treatment effects on posterior beliefs about Russian war crimes.

U.S. treatment having negative effects and the ICC treatment having positive effects. The differences in the two treatment effects are also similar.

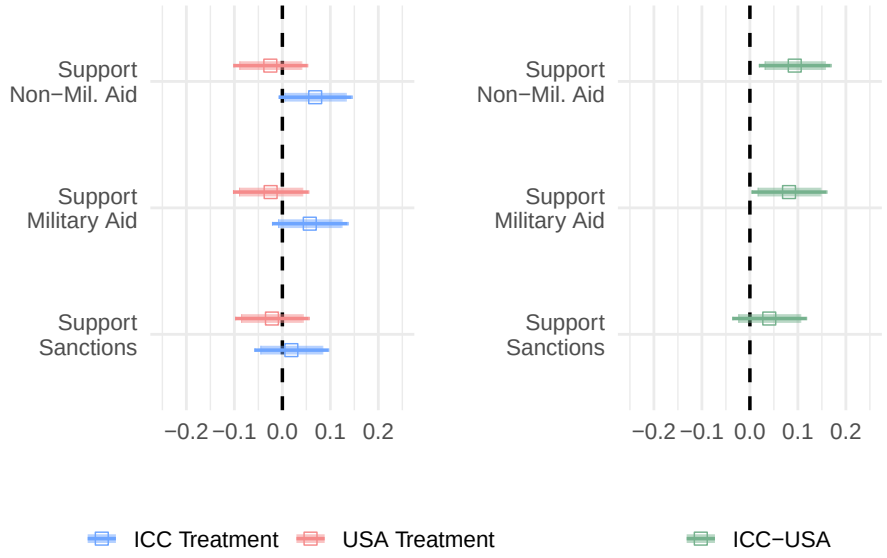


Figure 5: Aggregate treatment effects on support for policy responses.

5.2 Hypothesis 1 Results

Why did the U.S. treatment have more negative effects than the ICC treatment? Which respondents were most affected by treatment? The distributions of trust in each source across respondents give

the first clue. Figure 6 shows the smoothed distributions of responses to the question of trust in the United States and ICC, by country. The vertical lines show the sample means. In all four countries, the ICC is viewed as much more trustworthy than the United States. The largest difference was in Turkey, where the mean for trust in the ICC was 54.6 compared to 36.3 for the United States, a gap of 18.3 points ($p < 0.001$). In Indonesia, the mean for trust in the ICC was 59.3 and trust in the United States was 43.9, a difference of 15.4 points ($p < 0.001$). In South Africa, the mean for trust in the ICC was 62.6 compared to 52.5 for the United States, a gap of 10.2 points ($p < 0.001$). The smallest difference was in India, where the mean for trust in the ICC was 72.8 and trust in the United States was 67.7, a difference of 5.1 points ($p < 0.001$).

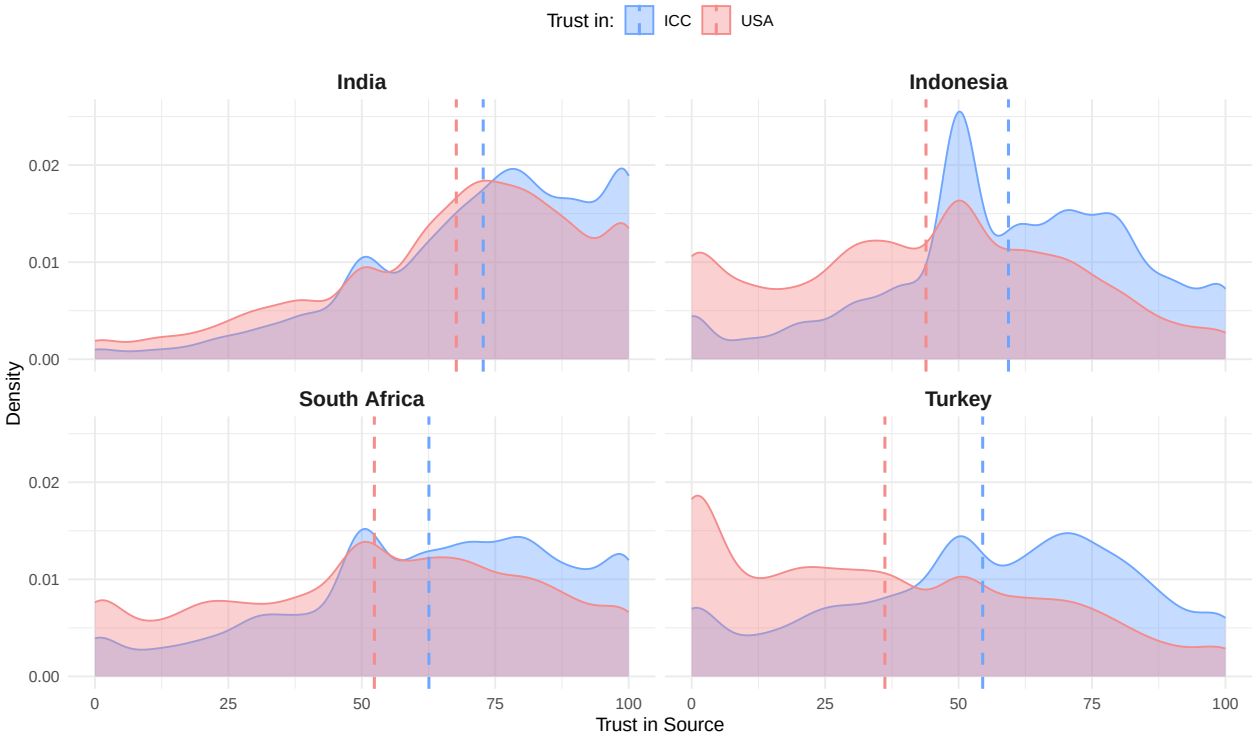


Figure 6: Distribution of trust in sources by country.

To assess Hypotheses 1a and 1b we classified respondents according to whether they were above or below the sample medians for the measures of prior likelihood of Russian guilt and pre-treatment measures of the trustworthiness of a source of information.[56] We then estimated the effect of the U.S.

[56]The appendix shows the sample sizes for each box and alternate specifications. We used the country-specific medians, but results are similar with a variety of other specifications, like using global medians instead of country-specific medians, using means instead of medians, including/excluding respondent characteristic controls and/or country-specific

and ICC treatments for the subsamples based on above/below median for priors and above/below median based on measures of source trustworthiness.

Hypothesis 1a predicts the greatest persuasive effects for respondents with low prior probabilities of Russian guilt and high trust in the source. Hypothesis 1b predicts backlash among respondents with high prior probabilities of Russian guilt and low beliefs about source trustworthiness. Figure 7 shows the ICC treatment effects by subgroup. The vertical axis shows whether the respondent was above or below the median in their prior beliefs about Russian guilt. The horizontal axis shows whether the respondent was above or below the median in their prior beliefs about the trustworthiness of the ICC. In other words, its layout matches that of Figure 1. Each cell shows the estimated treatment effect and the p value for a test of whether the treatment effect is different from zero.[57]

Respondents in the bottom right cell – with priors that Russia was innocent and who also had higher prior trust in the ICC were *most* persuaded by the treatment. Their posterior beliefs about Russian guilt were approximately 3% higher than their priors about Russian guilt. In this cell, for respondents in the control group, the outcome measure for Russian guilt was approximately 62 out of 100. In the treatment group, this outcome measure was approximately 65. Respondents in the top left cell – who thought Russia was guilty but did not trust the ICC – moved their beliefs in the opposite direction, as expected. They *lowered* their posterior probability of Russian guilt by approximately 3%, from 77 to 74. This pattern matched that predicted by the theoretical model and Hypothesis 1.

It is most important to establish that the treatment effects in the top left and bottom right are different from one another. The diagonal arrow shows the p-value for a statistical test of whether the top left effect (backlash group) differs from the bottom right effect (persuasion group). The treatment effects differ significantly between the two groups at the p = 0.01 level.

These results are especially striking because, recall, that the aggregate effects were near zero and insignificant. Those aggregate analyses obscured substantial heterogeneity in treatment effects - heterogeneity that closely matched that predicted by our model. For a respondent in the top left, the ICC

intercepts. See appendix.

[57] We interacted treatment with indicator variables for which cell a respondent was in, including cell-specific intercepts. Standard errors and test statistics were calculated with the *emmeans* package in R.

treatment lowered her agreement with the statement that Russia was guilty by over three points. For a respondent in the bottom right, treatment persuaded her and increased her belief that Russia was guilty by almost three points.

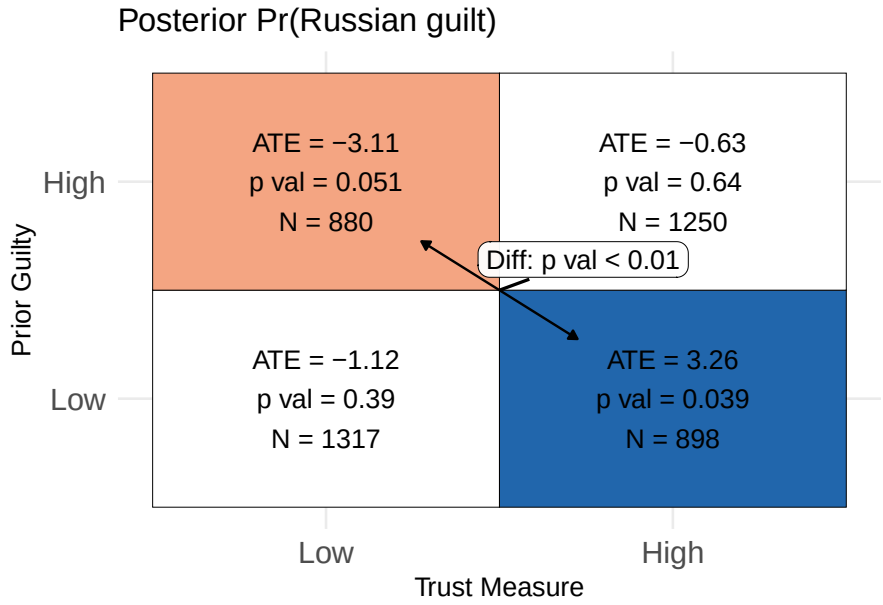


Figure 7: **Effect of ICC treatment on posteriors of Russian guilt.** The ATE captures the effect of treatment among respondents within each subgroup by prior beliefs about Russia and the ICC. The diagonal arrow shows the difference in treatment effects between the backlash and persuasion groups predicted by our model.

Figure 8 shows the ICC treatment effects, in this same way, for all four outcome variables. The top left pane, matches Figure 7. The other panes show treatment effects for sanctions on Russia, non-military aid for Ukraine, and military aid for Ukraine. The patterns are similar. Respondents in the bottom right are those most moved to support sanctions or aid to Ukraine. Backlash is less common, though it tends to be among respondents in the top left. For all four outcome measures, the diagonal arrow shows the treatment effects differ significantly between the backlash and persuasion groups at conventional significance levels. These differences show that respondents in different subgroups react differently to the same information, in a way that generally aligns with our theoretical model.

Figure 9 shows effects of the U.S. treatment in the same format. There are important similarities and differences, relative to both the ICC treatment effects and the model's predictions. As expected, backlash is most intense in the top left cells. For those respondents, U.S. accusations lowered their

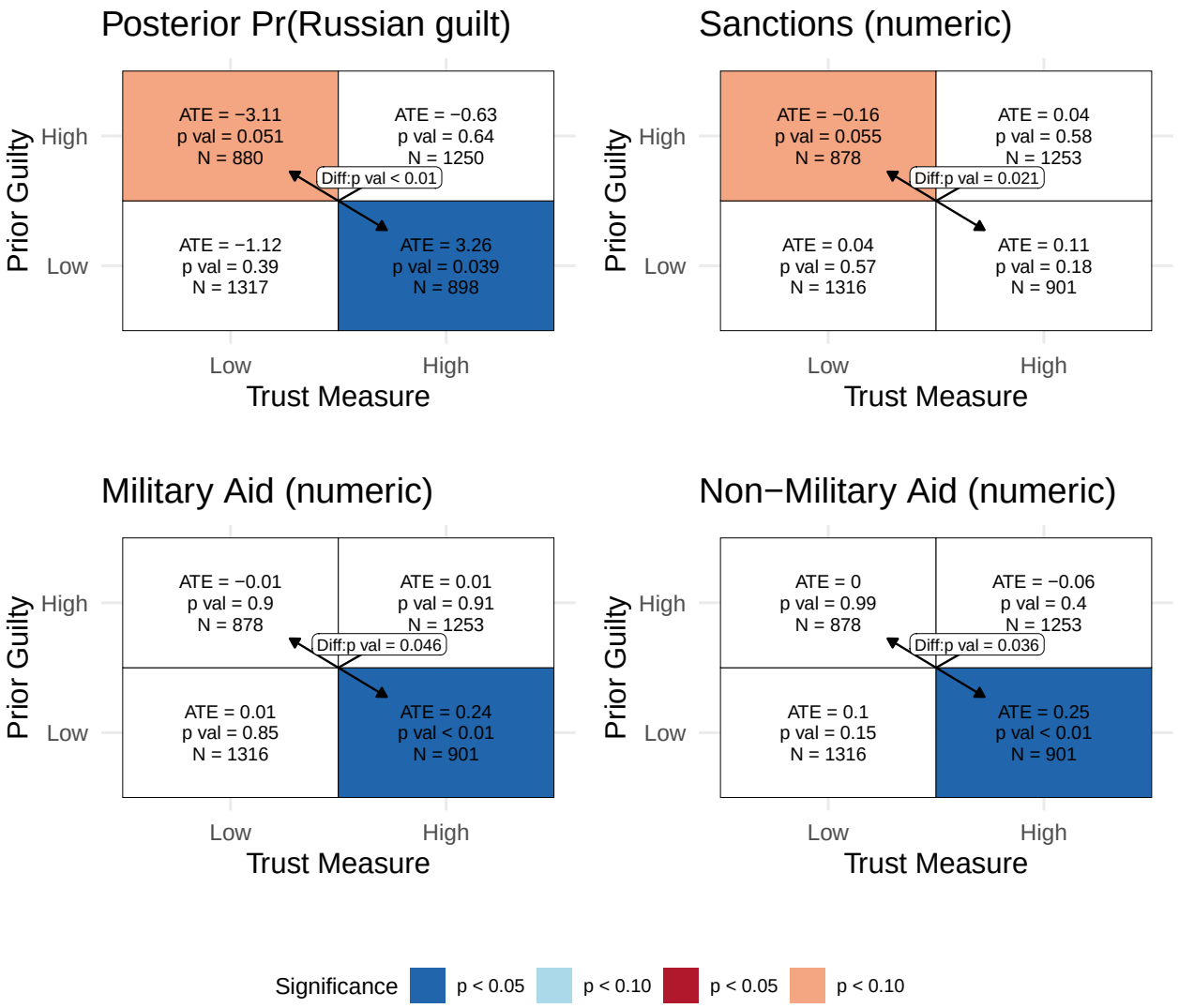


Figure 8: **Effect of ICC treatment on posteriors of all four outcome measure.** The ATE captures the effect of treatment among respondents within each subgroup by prior beliefs about Russia and the ICC. The diagonal arrow shows the difference in treatment effects between the backlash and persuasion groups predicted by our model.

posterior beliefs that Russia was guilty by over 3%.

Yet, backlash is much more prevalent, overall, for the U.S. treatment than expected by our theoretical model. In most cells, for all four outcome measures, the U.S. treatment has a negative effect lowering posteriors of Russian guilt or lowering support for policy responses against Russia. Among respondents in the bottom right, there is not evidence of persuasion where we would have expected it. For those respondents, treatment effects are generally negative and statistically insignificant. We can also only reject the null of no difference along the diagonals for posteriors about Russian guilt at the 0.1 level and reject the null for non-military aid at the 0.05 level.[58]

One possible explanation for the non-persuasiveness of the U.S. treatment, even among bottom right respondents, is that their general trust in the United States may be higher than their trust on issues related to Russia or war crimes. If respondents thought the United States told the truth in general, but not about Russia, then that would make respondents across different cells backlash against the U.S. accusation. Respondents' high levels of backlash against the U.S. treatment could also indicate that part of their negative responses are an expression of disapproval of the United States itself, separate from the intended effect of their beliefs about Russia.[59]

The ICC and U.S. results are also evidence that the effects are as predicted by the theoretical model and not simply floor and ceiling effects. Floor and ceiling effects would show red effects on the top row and blue effects on the bottom row. But this is not the observed pattern. Beliefs and support for policies are not simply being moved away from floors or ceilings. They are being moved in ways that depend *jointly* on prior beliefs and trust in the source.

5.3 Aggregate Effects on Perceptions of Information Sources

Figure 10 shows the effect of treatment on perceptions of the trustworthiness of each source. When the United States made accusations, respondents' trust in the U.S. declined. It reinforced perceptions

[58] It's possible that attitudes are harder to move on sanctions and military aid, since those options may be perceived as too costly in this context, even in the face of an accusation. Distribution of baseline support for each policy is summarized in Figure B.3. While exploring perceptions of these costs is beyond the scope of this paper, it remains an important question for future research.

[59] We return to this in the conclusion.

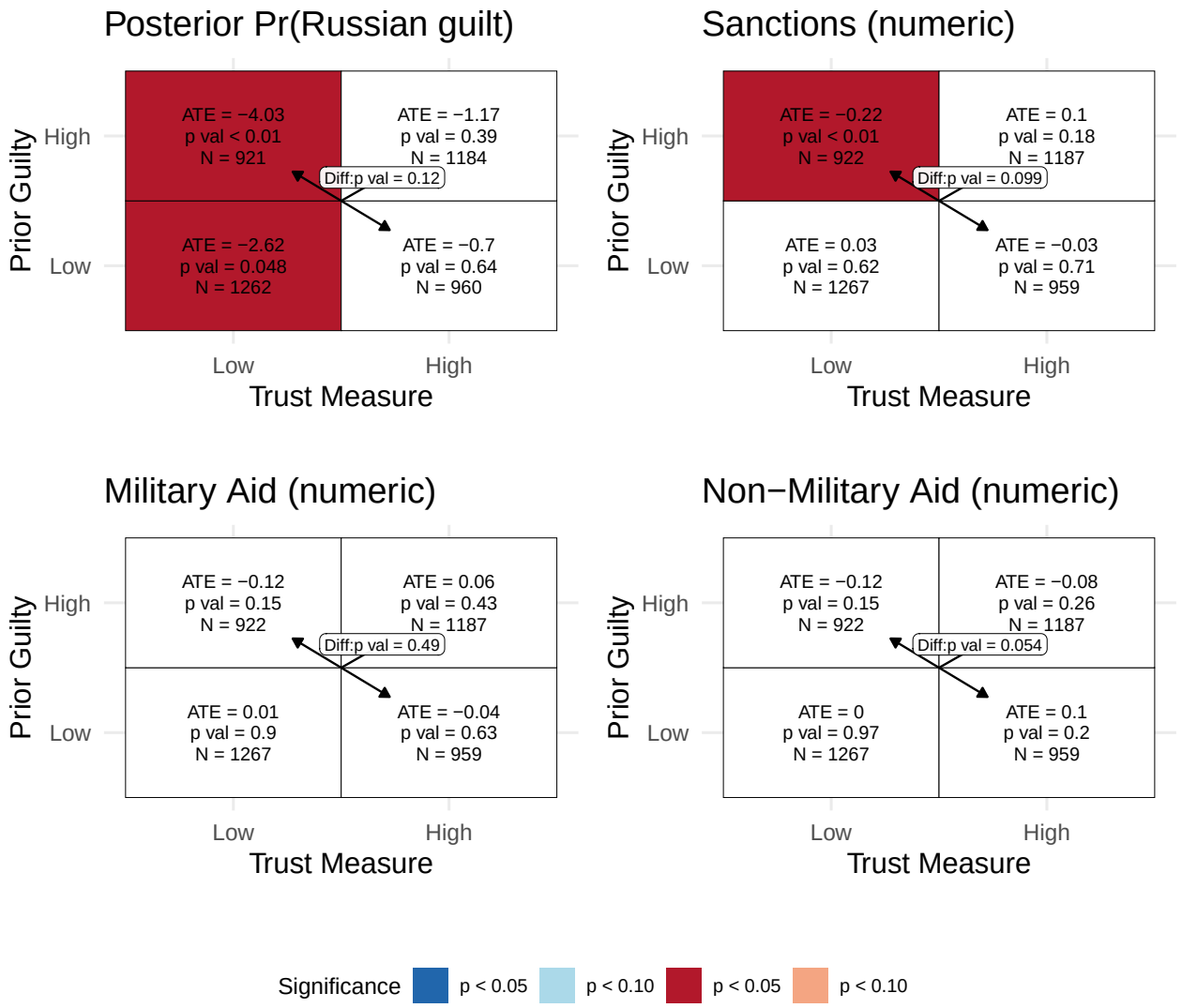


Figure 9: **Effect of USA treatment on posteriors of all four outcome measures** The ATE captures the effect of treatment among respondents within each subgroup by prior beliefs about Russia and the US. The diagonal arrow shows the difference in treatment effects between the backlash and persuasion groups predicted by our model.

of the United States as a biased source of information. Trust in the U.S. decreased by approximately 1.5 points on a 100-point scale, a decline of about 4.2%. In contrast, when the same accusation came from the ICC, trust in the ICC increased, strengthening beliefs in the ICC's impartiality. Trust in the ICC rose by approximately 1.8 points, corresponding to a 4.3% increase. The ICC's signal strengthened public trust of the source.[60]

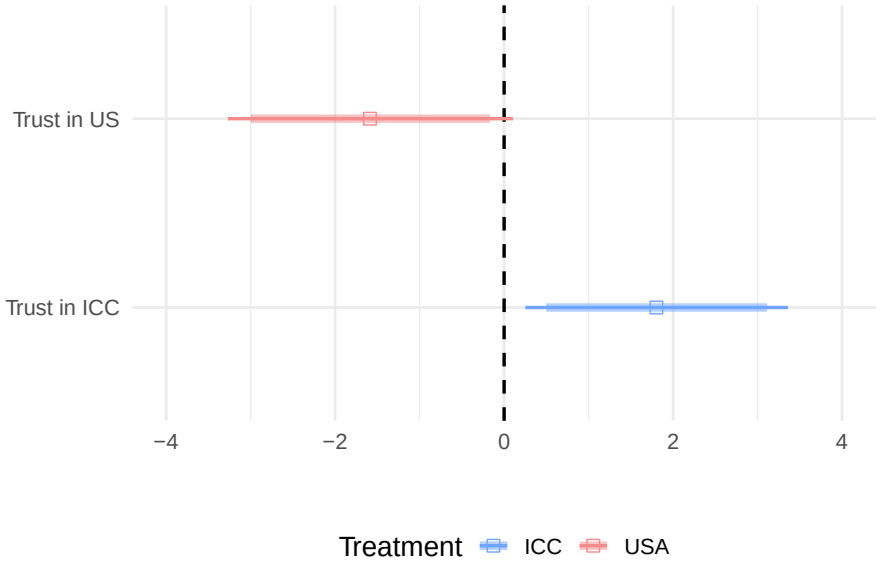


Figure 10: Effect of treatment on trustworthiness of source

5.4 Hypothesis 2 Results

For Hypothesis 2, the results are consistent with the prediction that treatment effects are increasing in priors about Russian guilt. Figure 11 shows how treatment effects vary with prior beliefs about Russian guilt. These are estimates from a linear interaction term model, interacting treatment with priors about Russian guilt. As expected, the lines are upward sloping. When the source gives information that matches the respondent's priors, the respondent increases their trust in the source.

Figure 12 shows 2x2 style boxes that we used to estimate treatment effects on posterior beliefs about Russian guilt. We again have priors about trust on the horizontal axis and priors about Russian guilt on the vertical axis. The outcome variable is now posteriors beliefs about a source's trustworthi-

[60] We also tested whether treatment affected perceptions of ICC legitimacy, as a related outcome measure. Treatment increased perceptions of ICC legitimacy. See appendix.

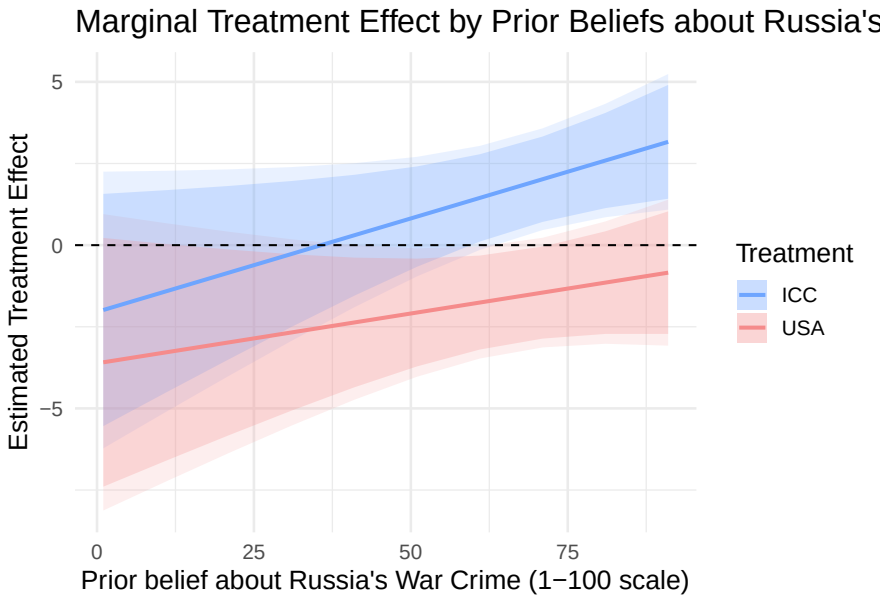


Figure 11: Effect of treatment on posteriors about source quality, as prior beliefs vary.

ness. Some features of the results match expectations from the model. We would expect the largest, positive effects to be in the top left of each figure. This is true for both the ICC and the United States. For the ICC, the respondents who had low initial trust in the Court, but then were treated with information that matched their priors, showed a significant increase in trust in the ICC. The only positive treatment effects for the United States were also found in the upper left quadrant, though these effects were insignificant. We would also expect the largest negative effects in the bottom right, which is true for the U.S. treatment. Respondents in this quadrant had lower post-treatment views of U.S. credibility. Though the same is not true for the ICC treatment, where there were positive effects in each quadrant

However, the differences between the results and other parts of the model's predictions are also interesting. Within each treatment group, the general patterns of treatments effects are similar to the model's predictions. But the positive aggregate effect of the ICC treatment, compared to the negative aggregate effect for the United States is surprising, because respondents generally started with more trusting views of the ICC. If more respondents start out with positive views of the ICC's trustworthiness and beliefs that Russia is guilty, then the model would have expected it to be harder to move their beliefs even further. The marginal effect on beliefs about the source's trustworthiness should be

smaller. This is similar to the "preaching to the choir" effect, where treatment has a weaker effect if respondents already believe what a source's message tells them.

Yet, the effect of the ICC on source trustworthiness is consistently stronger than the effect of the U.S. treatment (as seen in Figure 10). The estimated treatment effects of the ICC are more positive/less negative for respondents all along the range of prior beliefs about Russian guilt (as see in Figure 11). The treatment effects for the ICC are positive in all four quadrants and the U.S. treatment effects are negative in three out of four quadrants (as seen in Figure 12). Even respondents that believe strongly in Russian guilt *ex ante* still have negative estimated treatment effects for the U.S. treatment. The model would have expected those respondents to feel slightly more trusting of the United States when it made an accusation that matched their strongly held priors.

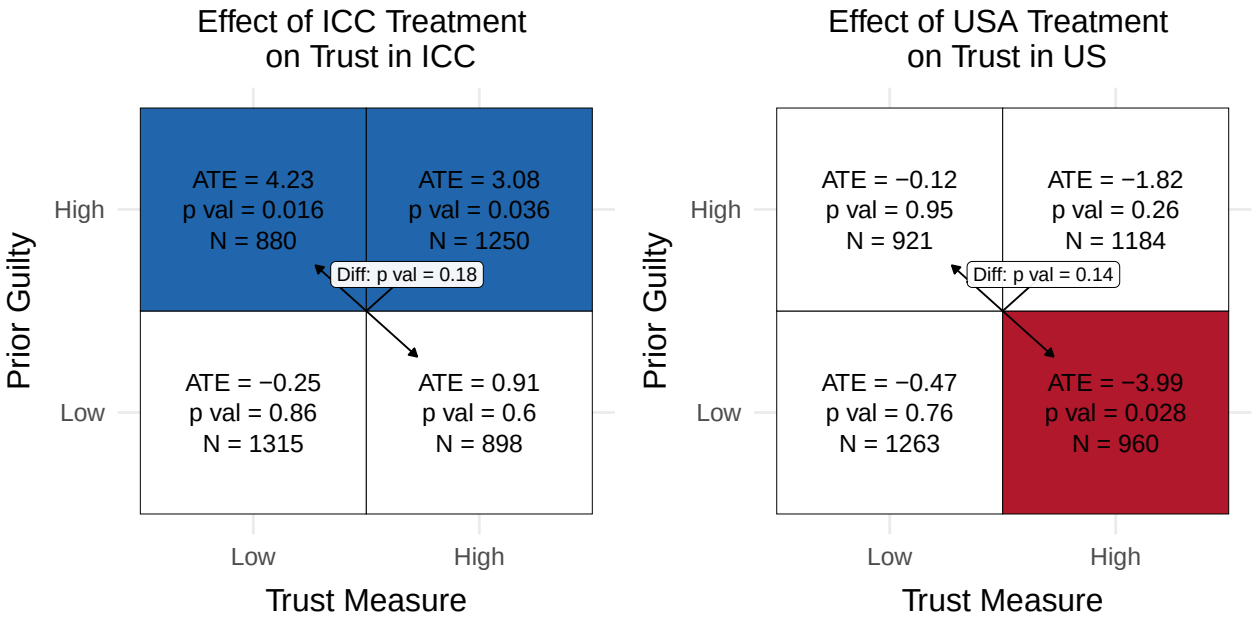


Figure 12: Effect of treatment on beliefs about the source, by prior beliefs about Russia and the source with arrows

This aspect of the results – with treatment improving views of the ICC's trustworthiness and decreasing views of the United States' trustworthiness – is striking, because it shows respondents diverging in their beliefs about source trustworthiness, despite both sources sending the same signal. Most models of persuasion generally predict convergence, when two sources say the same thing.[61]

[61]For a more extended discussion of the conditions under which convergence in beliefs about the sender may not converge, see Acemoglu, Chernozhukov, and Yildiz (2016) and Cheng and Hsiaw (2022).

This difference between the theoretical model's prediction and the findings is especially interesting because it suggests that there is something distinct about the signalling advantage of the ICC over the United States. Despite giving the same signal, and despite the deck being stacked against finding positive effects of the ICC treatment, the Court is *still* better able to persuade audiences about its credibility. The skepticism that respondents held towards the United States, *ex ante*, was reinforced and deepened, even though our treatment had the United States give respondents the same piece of information as the information given by ICC.

One possible explanation is that respondents infer different information from the treatment, even though the wording is identical, and the post-treatment measure of trust elicits something about this additional information. When the ICC accuses Russia of war crimes, it is possible that this conveys information that a legal body has evaluated evidence following a particular legal procedure and come to a corresponding conclusion about war crimes. Perhaps this conveys the idea of an investigation, with evidence weighed and debated in open court. And the Court only sends its signal that Russia has committed war crimes after this careful process. When the United States accuses Russia of war crimes, it is possible that the respondent does not infer anything about deliberation or weighing of evidence.

If the post-treatment measurement of trust captures more than just a posterior belief about the accuracy of a signal – i.e. it also captures beliefs about the quality of the process to arrive at that signal, beyond the statement of the signal itself – then that could explain this unexpected aspect of the treatment effects. The treatment effect is therefore capturing an updated view about the source's process, not just that the source's output was a signal that conveys the right answer about the state of the world. If so, that could explain this unexpected aspect of treatment effects. To assess these possibilities, we would need different outcome measures that captured these potential unintended effects.

Another related possibility is that respondents express disapproval of the United States any time it issues a negative judgement about other states because they view the United States as hypocritical or hubristic. Chow and Levin (2024) highlight the power of negative "whataboutism" in international relations. It is possible that respondents express negative post-treatment attitudes about trust in the

United States because they believe that Americans have also committed similarly objectionable acts. Chow and Levin (2024) explain how whataboutism is an effective, defensive rhetorical strategy employed by the targets of criticism. Here, perceptions of U.S. hypocrisy might have made third party audiences more skeptical of their accusations, as well. Respondents may be saying that the U.S. is untrustworthy as part of their own broader condemnation of the United States.

6 Conclusions

Accusations about violations of international law generate both persuasion and backlash among publics in global swing states. Whether accusations persuade or backfire depends on who sends the message and whom they are trying to convince. Using survey experiments in four swing states – India, Indonesia, Turkey, and South Africa – we show that identical accusations against Russia lead to divergent reactions when attributed to different sources. Accusations from the International Criminal Court produce modest persuasion among those who trust the ICC and do not already hold strong prior beliefs about Russian guilt. In contrast, accusations from the United States often backfire, undermining both belief in the accusation and trust in the United States as a sender. We offer a theoretical model to explain these patterns, showing how belief updating is jointly shaped by priors and perceived source credibility.

Our results offer cautious optimism about the potential for international organizations to build persuasive power through consistent and credible engagement. All but one of our surveyed countries has refused to join the Court. Yet, even a Court that does not have universal support, especially among global swing states, was able to persuade some subsets of respondents. Even more encouragingly, its messages also increased perceptions of the Court's own trustworthiness and legitimacy, even among respondents that doubted the Court's message. A hopeful aspect of this finding is that the Court may be building a well of legitimacy that makes it even more persuasive in the future, despite many of its decisions being met with disagreement or ambivalence.

However, the contrast between the ICC findings and those for the United States are ominous

with respect to U.S. credibility. Our results suggest that the United States' messaging is more harmful for its agenda than remaining silent, at least with respect to public opinion in critical swing states. U.S. messaging was less effective even than the often-maligned ICC, and it backfired for a plurality of respondents. Notably, our surveys were implemented *before* the 2024 U.S. Presidential elections. Global perceptions of the United States have further declined after the survey. According to surveys conducted in over 100 countries, China's net favorability rating is now 19 points higher than the United States' and Russia's net rating is only 4 points behind that of the United States.[62] The ICC can at least hope that its messaging triggers positive feedback effects, where accusations enhance credibility which makes future accusations more effective. The United States, on the other hand, needs to worry about a doom spiral, where its lack of credibility causes accusations to backfire and make its future messaging less credible. U.S. policymakers that discount the importance of soft power would do well to remember that credibility helps persuade others to back concrete punishments and hard power coercion against U.S. adversaries, like sanctions or arms transfers to allies.

Our arguments speak to a broader literature on how sources of information shape public opinion. We demonstrate a theoretical framework that generates testable predictions and match it with a survey experimental design that allows for precisely testing its predictions. Our results show that such messages can either persuade or provoke backlash, depending on who delivers the message, how audiences evaluate the trustworthiness of the sender, and the respondent's prior beliefs. This framework makes it clear how aggregate effects can obscure important heterogeneity. Without measurements of priors and beliefs about sources, most experimental designs can't discriminate between or detect possible heterogeneous effects based on how the respondent views the state of the world and the source of the information. Similarly, many often used moderating variables may conflate cross-cutting effects, which complicates their analysis as well. We hope that our approach, while more demanding of experimental designs and data, gives a tractable approach to modelling and assessing how people react to signals.

Our findings point towards several areas for further research. Our theoretical model and empirical

[62] Nira Democracy Perception Index 2025 Report. https://www.niradata.com/dpi.

approach give a template for how to alter any of those items, make testable predictions, and assess them with an experiment. In our paper, the analysis examines a specific pair of information sources (the ICC, USA), making an accusation about an important, but also specific event (alleged Russian war crimes), to public audiences in specific swing states (India, Indonesia, Turkey, and South Africa). To evaluate the generalizability of these expectations, it is essential to examine whether similar patterns of persuasion and backlash arise with different signal senders, audiences, and accusations.

Future research could vary the type of messenger, including both international organizations and individual states. Institutions such as the International Court of Justice, the United Nations, and the European Union also seek to shape public opinion through the dissemination of information. Whether such signals from these actors elicit similar patterns of persuasion or backlash represents a valuable avenue for future studies. Likewise, individual states engage in efforts to influence foreign publics – campaigns described as public diplomacy or propaganda. The signals in our application came from foreign sources, like international organizations, public diplomacy efforts, or other states' naming and shaming. But our approach is portable to domestic sources, like the media or political elites. There, too, we would expect priors about the message and the messenger condition reactions. Future research could also vary the audience, beyond publics in swing states. The same audience may also have priors distributed in significantly different ways, depending on the messenger and the message. Varying any one (or more) of these dimensions has implications for the reactions we would expect.

Varying the issue or event at hand is also an important exploration of the scope conditions of our argument. For instance, there are many other important issues where the international court makes similar accusations to similar audiences, but about other alleged perpetrators and events. In 2024, the ICC issued arrest warrants for both Hamas and Israeli leaders in connection with the Israel–Palestine conflict. The Court's involvement in this case sparked significant public debate and controversy across countries and political groups. Some governments, especially member states of the Rome Statute, welcomed the investigation, while other governments expressed strong opposition.[63] The United States, for example, condemned the Court's actions and ultimately imposed sanctions on the ICC in

[63] https://www.justsecurity.org/105064/arrest-warrants-state-reactions-icc/.

2025.[64] Our approach gives a starting point for assessing the effects of these accusations, which is intrinsically valuable since these are important, real-world events.

[64] https://www.whitehouse.gov/presidential-actions/2025/02/imposing-sanctions-on-the-international-criminal-court/.

7 References

Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz. 2016. "Fragility of Asymptotic Agreement Under Bayesian Learning." *Theoretical Economics* 11 (1): 187–225.

Anjum, Gulnaz, Adam Chilton, and Zahid Usman. 2021. "United Nations Endorsement and Support for Human Rights: An Experiment on Women's Rights in Pakistan." *Journal of Peace Research* 58 (3): 462–78.

Arias, Eric, Horacio Larreguy, John Marshall, and Pablo Querubin. 2022. "Priors Rule: When Do Malfeasance Revelations Help or Hurt Incumbent Parties?" *Journal of the European Economic Association* 20 (4): 1433–77.

Bearce, David H, and Thomas R Cook. 2018. "The First Image Reversed: IGO Signals and Mass Political Attitudes." *The Review of International Organizations* 13 (4): 595–619.

Brutger, Ryan. 2021. "The Power of Compromise: Proposal Power, Partisanship, and Public Support in International Bargaining." *World Politics* 73 (1): 128–66.

Brutger, Ryan, and Anton Strezhnev. 2022. "International Investment Disputes, Media Coverage, and Backlash Against International Law." *Journal of Conflict Resolution* 66 (6): 983–1009.

Búzás, Zoltán I, and Lotem Bassan-Nygate. 2024. "Race, Shaming, and International Human Rights." *American Journal of Political Science*.

Carnegie, Allison, Richard Clark, and Lisa Fan. 2024. "Multilateral Messaging: International Organizations, Populism, and Social Media."

Chapman, Terrence L. 2007. "International Security Institutions, Domestic Politics, and Institutional Legitimacy." *Journal of Conflict Resolution* 51 (1): 134–66.

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

———. 2016. "How Contestation Moderates the Effects of International Institutions: The International Criminal Court and Kenya." *The Journal of Politics* 78 (2): 557–71.

———. 2023. "How International Organizations Change National Media Coverage of Human Rights." *International Organization* 77 (1): 238–61.

Chaudoin, Stephen, Brian J Gaines, and Avital Livny. 2021. "Survey Design, Order Effects, and Causal Mediation Analysis." *The Journal of Politics* 83 (4): 1851–56.

Cheng, Haw, and Alice Hsiaw. 2022. "Distrust in Experts and the Origins of Disagreement." *Journal of Economic Theory* 200: 105401.

Chilton, Adam, and Katerina Linos. 2021. "Preferences and Compliance with International Law." *Theoretical Inquiries in Law* 22 (2): 247–98.

Chivvis, Christopher S., Elina Noor, and Beatrix Geaghan-Breiner. 2023. "Indonesia in the Emerging World Order." 2023. https://carnegieendowment.org/research/2023/11/indonesia-in-the-emerging-world-order?lang=en.

Choi, Ha Eun, JiHwan Jeong, Amanda Murdie, Byungwon Woo, and Hyunjin Yim. 2023. "UN Secretary-General Visits and Human Rights Diplomacy." In *Paper Presented at the 15th Annual Conference Political Economy of International Organization*.

Chow, Wilfred M, and Dov H Levin. 2024. "The Diplomacy of Whataboutism and US Foreign Policy Attitudes." *International Organization* 78 (1): 103–33.

Chu, Jonathan Art. 2025. *Social Cues: How the Liberal Community Legitimizes Humanitarian War*. Cambridge University Press.

Chung, Seowoo. 2025. "Strategic Censorship? Public Opinion, Authoritarian Politics, and the International Trade Regime."

Cohen, Harlan, and Ryan Powers. 2024. "Judicialization and Public Support for Compliance with International Commitments." *International Studies Quarterly* 68 (3): sqae078.

Cope, Kevin L. 2023. "Measuring Law's Normative Force." *Journal of Empirical Legal Studies* 20 (4): 1005–44.

Cope, Kevin L, and Charles Crabtree. 2020. "A Nationalist Backlash to International Refugee Law: Evidence from a Survey Experiment in Turkey." *Journal of Empirical Legal Studies* 17 (4): 752–88.

Coppock, Alexander. 2023. *Persuasion in Parallel: How Information Changes Minds about Politics*. University of Chicago Press.

Delavande, Adeline. 2014. "Probabilistic Expectations in Developing Countries." *Annu. Rev. Econ.* 6 (1): 1–20.

Dellmuth, Lisa Maria, Jan Aart Scholte, and Jonas Tallberg. 2019. "Institutional Sources of Legitimacy for International Organisations: Beyond Procedure Versus Performance." *Review of International Studies* 45 (4): 627–46.

Dellmuth, Lisa M, and Jonas Tallberg. 2021. "Elite Communication and the Popular Legitimacy of International Organizations." *British Journal of Political Science* 51 (3): 1292–1313.

Ecker-Ehrhardt, Matthias, Lisa Dellmuth, and Jonas Tallberg. 2024. "Ideology and Legitimacy in Global Governance." *International Organization* 78 (4): 731–65.

Efrat, Asif, and Omer Yair. 2023. "International Rankings and Public Opinion: Compliance, Dismissal, or Backlash?" *The Review of International Organizations* 18 (4): 607–29.

Fang, Songying. 2008. "The Informational Role of International Institutions and Domestic Politics." *American Journal of Political Science* 52 (2): 304–21.

Fontaine, Richard, and Daniel M Kliman. 2013. "International Order and Global Swing States." *The Washington Quarterly* 36 (1): 93–109.

Fontaine, Richard, and Gibbs McKinley. 2025. "Global Swing States and the New Great Power Competition." *The Washington Quarterly* 48 (2): 7–28.

Gentzkow, Matthew, Michael B Wong, and Allen T Zhang. Forthcoming. "Ideological Bias and Trust in Information Sources." *American Economic Journal: Microeconomics*, Forthcoming.

Ghassim, Farsan. 2024. "Effects of Self-Legitimation and Delegitimation on Public Attitudes Toward International Organizations: A Worldwide Survey Experiment." *International Studies Quarterly* 68 (2): sqae012.

Goldsmith, Benjamin E, and Yusaku Horiuchi. 2009. "Spinning the Globe? US Public Diplomacy and Foreign Public Opinion." *The Journal of Politics* 71 (3): 863–75.

Goldsmith, Benjamin E, Yusaku Horiuchi, and Kelly Matush. 2021. "Does Public Diplomacy Sway Foreign Public Opinion? Identifying the Effect of High-Level Visits." *American Political Science Review* 115 (4): 1342–57.

Grieco, Joseph M, Christopher Gelpi, Jason Reifler, and Peter D Feaver. 2011. "Let's Get a Second Opinion: International Institutions and American Public Support for War." *International Studies Quarterly* 55 (2): 563–83.

Hansen, Ben B, and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science*, 219–36.

Kertzer, Joshua D, Brian C Rathbun, and Nina Srinivasan Rathbun. 2020. "The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations." *International Organization* 74 (1): 95–118.

Little, Andrew T. 2025. "How to Distinguish Motivated Reasoning from Bayesian Updating." *Political Behavior*, 1–25.

Lupia, Arthur, and Mathew D McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge University Press.

Madsen, Mikael Rask, Juan A Mayoral, Anton Strezhnev, and Erik Voeten. 2022. "Sovereignty, Substance, and Public Support for European Courts' Human Rights Rulings." *American Political Science Review* 116 (2): 419–38.

Mattingly, Daniel, and James Sundquist. 2023. "When Does Public Diplomacy Work? Evidence from China's 'Wolf Warrior' Diplomats." *Political Science Research and Methods* 11 (4): 921–29.

Mikulaschek, Christoph. 2023. "The Responsive Public: How European Union Decisions Shape Public Opinion on Salient Policies." *European Union Politics* 24 (4): 645–65.

Mikulaschek, Christoph, and Michal Parizek. 2025. "How Media Coverage Shapes the Effect of IOs on Public Attitudes: Quasi-Experimental Evidence on Mass Opinion about Russia's Leadership in 49 Countries."

Morse, Julia C, and Tyler Pratt. 2022. "Strategies of Contestation: International Law, Domestic Audiences, and Image Management." *The Journal of Politics* 84 (4): 2080–93.

———. 2025. "Smoke and Mirrors: Strategic Messaging and the Politics of Noncompliance." *American Political Science Review*, 1–19.

Pauselli, Gino. 2023. "Look Who Is Talking: Direct and Indirect Effects of Criticism on LGBT Rights." *Available at SSRN 4317082*.

Recchia, Stefano, and Jonathan Chu. 2021. "Validating Threat: IO Approval and Public Support for Joining Military Counterterrorism Coalitions." *International Studies Quarterly* 65 (4): 919–28.

Rhee, Kasey, Charles Crabtree, and Yusaku Horiuchi. 2024. "Perceived Motives of Public Diplomacy Influence Foreign Public Opinion." *Political Behavior* 46 (1): 683–703.

Spilker, Gabriele, Quynh Nguyen, and Thomas Bernauer. 2020. "Trading Arguments: Opinion Updating in the Context of International Trade Agreements." *International Studies Quarterly* 64 (4): 929–38.

Suong, Clara H, Scott Desposato, and Erik Gartzke. 2024. "Ubiquitous but Heterogeneous: International Organizations' Influence on Public Opinion in China, Brazil, Japan, and Sweden." *International Relations of the Asia-Pacific*, lcae018.

Terman, Rochelle. 2023. "The Geopolitics of Shaming: When Human Rights Pressure Works and When It Backfires."

Thompson, Alexander. 2006. "Coercion Through IOs: The Security Council and the Logic of Information Transmission." *International Organization* 60 (1): 1–34.

———. 2015. *Channels of Power: The UN Security Council and US Statecraft in Iraq*. Cornell University Press.

Wallace, Geoffrey PR. 2013. "International Law and Public Attitudes Toward Torture: An Experimental Study." *International Organization* 67 (1): 105–40.

Wang, Austin Horng-En, Charles KS Wu, Yao-Yuan Yeh, and Fang-Yu Chen. 2023. "High-Level Visit and National Security Policy: Evidence from a Quasi-Experiment in Taiwan." *International Interactions* 49 (1): 132–46.

Zvobgo, Kelebogile. 2019. "Human Rights Versus National Interests: Shifting US Public Attitudes on the International Criminal Court." *International Studies Quarterly* 63 (4): 1065–78.

A Theory

A.1 Expression for treatment effect about state of the world

The expression for the treatment effect for posteriors about the state of the world is a straightforward application of Bayes' rule. We omit the i subscripts for simplification.

$$\Pr(S = 1 \mid s_1) = \frac{\Pr(S = 1)\Pr(s_1 \mid S = 1)}{\Pr(S = 1)\Pr(s_1 \mid S = 1) + (1 - \Pr(S = 1))\Pr(s_1 \mid S = 0)}$$

We can compute the likelihood terms using expectations under the Beta distribution:

$$\Pr(s_1 \mid S = 1) = \int_0^1 f(\sigma) \cdot \sigma \, d\sigma = \mathbb{E}[\sigma] = \frac{\alpha}{\alpha + \beta}$$

$$\Pr(s_1 \mid S = 0) = \int_0^1 f(\sigma) \cdot (1 - \sigma) \, d\sigma = 1 - \mathbb{E}[\sigma] = \frac{\beta}{\alpha + \beta}$$

Substituting into Bayes' rule:

$$\Pr(S = 1 \mid s_1) = \frac{\Pr(S = 1) \cdot \frac{\alpha}{\alpha+\beta}}{\Pr(S = 1) \cdot \frac{\alpha}{\alpha+\beta} + (1 - \Pr(S = 1)) \cdot \frac{\beta}{\alpha+\beta}}$$

Simplifying:

$$\Pr(S = 1 \mid s_1) = \frac{\Pr(S = 1) \cdot \alpha}{\Pr(S = 1) \cdot \alpha + (1 - \Pr(S = 1)) \cdot \beta}$$

Letting $\pi = \Pr(S = 1)$:

$$\Pr(S = 1 \mid s_1) = \frac{\pi\alpha}{\pi\alpha + (1 - \pi)\beta}$$

So the treatment effect, with i subscripts reintroduced, is:

$$\Pi_i = \frac{\pi_i\alpha_i}{\pi_i\alpha_i + (1 - \pi_i)\beta_i} - \pi_i$$

A.2 Expression for treatment effect about source trustworthiness

Recall that the treatment effect for source trustworthiness is: $\Sigma_i = \mathbb{E}\left[\sigma_i | s_1\right] - \mathbb{E}\left[\sigma_i\right]$.

The first term, omitting i subscripts again, $\mathbb{E}\left[\sigma | s_1\right]$ can be written by breaking down the two possibilities - either the sender was right or they were wrong.

$$\mathbb{E}\left[\sigma | s_1\right] = \Pr(S = 1 \mid s_1) \cdot \mathbb{E}\left[\sigma | s_1, S = 1\right] + \Pr(S = 0 \mid s_1) \cdot \mathbb{E}\left[\sigma | s_1, S = 0\right]$$

Substituting the expression for $\Pr(S = 1 \mid s_1)$ from above...

$$\mathbb{E}\left[\sigma | s_1\right] = \frac{\pi\alpha}{\pi\alpha + (1 - \pi)\beta} \cdot \mathbb{E}\left[\sigma | s_1, S = 1\right] + (1 - \frac{\pi\alpha}{\pi\alpha + (1 - \pi)\beta}) \cdot \mathbb{E}\left[\sigma | s_1, S = 0\right]$$

$$\mathbb{E}\left[\sigma|s_1\right] = \frac{\pi\alpha}{\pi\alpha + (1-\pi)\beta} \cdot \mathbb{E}\left[\sigma|s_1, S=1\right] + \frac{(1-\pi)\beta}{\pi\alpha + (1-\pi)\beta} \cdot \mathbb{E}\left[\sigma|s_1, S=0\right]$$

For the term $\mathbb{E}\left[\sigma|s_1, S=1\right]$, this occurs when the signal sender gets it "right." Their signal correctly matched the state of the world. From the Beta-Binomial conjugacy, their "new" σ is distributed Beta with parameters $\alpha+1$ and β. The expectation of that new distribution is $\frac{\alpha+1}{\alpha+\beta+1}$. For the term $\mathbb{E}\left[\sigma|s_1, S=0\right]$, this occurs when the signal sender gets it "wrong." The expectation of that new distribution is $\frac{\alpha}{\alpha+\beta+1}$.

Substituting these expressions in…

$$\mathbb{E}\left[\sigma|s_1\right] = \frac{\pi\alpha}{\pi\alpha + (1-\pi)\beta} \cdot \frac{\alpha+1}{\alpha+\beta+1} + \frac{(1-\pi)\beta}{\pi\alpha + (1-\pi)\beta} \cdot \frac{\alpha}{\alpha+\beta+1}$$

Simplifying and re-adding i subscripts…

$$\mathbb{E}\left[\sigma_i|s_1\right] = \frac{\alpha_i}{\alpha_i+\beta_i+1} \cdot \left[1 + \frac{\pi_i}{\pi_i\alpha_i + (1-\pi_i)\beta_i}\right]$$

Note that this expression is increasing in π_i. By extension, the treatment effect expression is also increasing in π_i.

The full treatment effect expression is…

$$\Sigma_i = \frac{\alpha_i}{\alpha_i+\beta_i+1} \cdot \left[1 + \frac{\pi_i}{\pi_i\alpha_i + (1-\pi_i)\beta_i}\right] - \frac{\alpha_i}{\alpha_i+\beta_i}$$

A.3 Relation to motivated reasoning models

Plenty of research contrasts Bayesian models of belief updating with alternate models of belief formation, such as those based on motivated reasoning. In motivated reasoning models, individuals form posteriors based on accuracy and directional motives. They may want to get their posteriors "right" (an accuracy motive), but they may also like it when their posteriors are closer to a preferred point (the directional motive). Kertzer, Rathbun, and Rathbun (2020) is a good example from international relations research. They describe how motivated reasoning conditions individuals' reactions to information about costly signalling. "It is precisely those who are motivated to find evidence of a costly signal who act as classic signalling 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" (97). They predict, and find, that individuals with more cooperative internationalist attitudes and/or less militant internationalist attitudes will respond more to costly signals. In their particular application, liberals and those with more positive feelings toward Iran responded more to costly signals from Iran.

Coppock (2023) (ch 7) and Little (2025) both argue that Bayesian models and most motivated reasoning models are indistinguishable with most experimental designs. If a piece of information moves a respondent in a particular way, this could be because she had a particular configuration of priors and accuracy beliefs about the signal *or* because she had particular biases about the direction of her preferred posterior belief. In the example from Kertzer, Rathbun, and Rathbun (2020), those who responded most to a signal may have had directional motives or they may have had different beliefs about the likelihood function generating those signals. A cooperative internationalist may subconsciously think "I am responding to this treatment in the intended direction because it pushes me towards my preferred posterior" *or* they may think "I am responding to this treatment in the intended direction because signals like this are more credible." Source credibility is sometimes described as a likelihood ratio, e.g. (Pr(Iran is peaceful | signal) / Pr(Iran is peaceful | no signal)). Without measurements of priors and the respondent's beliefs about the signal's credibility, these alternatives are impossible to distinguish from one another.

We do not attempt to resolve this voluminous debate about Bayesian models versus their alternatives. Rather, we make two remarks. First, whatever model is used, it should make precise predictions about the direction and magnitude of treatment effects. If those predicted effects are moderated by receiver characteristics (like priors), then the model should make apparent what must be measured pre-treatment to test predictions about who will be most moved by a treatment. Making predictions based on Bayesian *or* motivated reasoning models generally requires measurements of priors and likelihood functions. Coppock argues that we can't tell Bayesian and motivated reasoning stories apart because "We would love to know if changing a likelihood changed a posterior, holding exposure to evidence constant, since that would provide direct evidence for the Bayesian model. But we can't, because likelihood functions are imaginary constructs whose existence in people's minds we can only posit." (137-8). However, just because likelihoods are hard to manipulate, this does not mean that they are impossible to measure. With measurements of priors and likelihoods, Bayes rule gives a predicted posterior, and by extension, a predicted treatment effect, that can be assessed against data. A motivated reasoning model would require those two measurements as well.[65]

[65] Little (2025) shows that these stories will still be indistinguishable, since the priors themselves could be generated from directional motives. We agree. However, our goal again is not to prove or disconfirm the existence of motivated reasoning models. Our goal is to say "conditional on observing priors and likelihoods, Bayes rule gives useful predictions about the types of individuals for whom treatment effects will be largest."

Second, in the context of diplomatic messaging and IO endorsements, it is important for any model of updating to accommodate persuasion *and* backlash. Existing work gives strong reasons to think that both phenomena occur in the real world.[66] Therefore, any model of the effects of diplomatic or IO messaging should be capable of yielding both types of effects. In most applications of motivated reasoning models, backlash does not occur. Predicted treatment effects may be muted, such as when a receiver chooses to discard information that does not match her priors. However, they generally do not generate predictions where information moves receivers in the opposite of its intended direction.

B Complete Survey Instrument, Summary Statistics

This section of the appendix describes every item on the survey. We include the entire instrument here for transparency and to hopefully reassure readers that we did not mine for moderating/heterogeneous treatment effects. Researchers have understandably become more worried about mining for results, particularly when investigating heterogeneous treatment effects - as is the focus of this paper. For readers worried that these HTE arguments are an example of mining, we wanted to show that this is the survey instrument in its entirety. The survey was designed to assess the predictions of the theoretical model – heterogeneous effects from prior beliefs about the state of the world and the trustworthiness of sources. The most likely candidate for an alternative HTE argument was cooperative internationalism. We also included it because the theoretical model makes it clear why the moderating effect of CI is ambiguous, which is what we find empirically. There aren't other moderators that we could potentially mine, or at least none that are tied directly to a formal model that makes precise predictions about what type of respondent should show what type of heterogeneous treatment effects.

We aren't against pre-registration. It is a useful check on mining for HTE and encourages researchers to have a high level of fidelity between their theory, instruments, and analysis. However, we think and hope that this study has a level of fidelity between the theory, its predicted heterogeneous treatment effects, and the accompanying survey instrument, that meets or exceeds the level of modern survey experimental work.

Pre-registration is also valuable for checking whether particular results are mined among different specifications. For this, Appendix C details the similarity of the main manuscript's results to estimates from a wide array of alternate specifications.

We first asked for informed consent. Respondents then had to pass a simple attention check that said "People are very busy these days and many do not have time to follow what goes on in the government. *We are testing whether people read questions.* To show that you've read this much, answer *both*"extremely interested" *and* "very interested."

We then presented the following six blocks, with their order randomized. We call them "blocks" but they are essentially one question, with responses for a small set of items. One block measured respondents' prior beliefs about whether countries had violated international law. We cared most about the item asking about Russia. The next four blocks measured beliefs about information sources. We cared most about the items asking about the United States and the ICC. The fifth block measured cooperative internationalism.

[66] Eg Terman (2023) on shaming and Goldsmith and Horiuchi (2009) on diplomacy.

B.1 Measuring prior beliefs

This block measured the respondent's prior beliefs that a country had broken international law. The underlined text below was not displayed to respondents. It is only here for readability. We included China and the United States to have other countries, but the key item here was the question about Russia.

- Prior beliefs about breaking international law Countries sometimes violate international laws of war that restrict attacking civilians and other acts. In your opinion, what is the percent chance that the countries below have violated international laws of war over the last 5 years? (0-100, order of items randomized)
  - Russia
  - United States
  - China

B.2 Measuring beliefs about information sources

The next set of blocks measured pre-treatment views about the trustworthiness of each source. Since respondents would not have been able to express their answers in terms of a likelihood function (e.g. "What's the probability the ICC says Russia is guilty if they are guilty?"), we used three different types of questions: about whether a country/international organization tells the truth, whether they are a trustworthy source of information, and a general feeling thermometer. The key items are those asking about the United States or the ICC. We again included other entities so that the entire focus was not just on the United States and ICC.

- Trustworthiness There are many sources of information about international affairs. Some sources of information are trustworthy and others are not. On a scale of 1-100, with zero being the least trustworthy and 100 being the most trustworthy, where would you place the following sources of information? (order of items randomized)
  - The International Criminal Court
  - The United States government
  - The media

- Countries Telling the Truth Countries criticize each other. Sometimes they are telling the truth and other times they have another motive. In your opinion, what is the percent chance that these countries are telling the truth when they criticize another country? (0-100, order of items randomized)
  - The United States
  - China
  - France

- IOs Telling the Truth International organizations accuse countries of breaking international rules. Sometimes they are telling the truth and other times they have another motive. In your opinion, what is the percent chance that these international organizations are telling the truth when they accuse countries of breaking international rules? (0-100, order of items randomized)

  - The International Criminal Court
  - The World Health Organization
  - The European Union

- Thermometer We'd like to get your feelings toward certain countries and international organizations on a "feeling thermometer." A rating of zero degrees means you feel as cold and negative as possible. A rating of 100 degrees means you feel as warm and positive as possible. You would rate the country or organization at 50 degrees if you don't feel particularly positively or negatively toward them. How do you feel about following countries or international organizations? (order of items randomized)
  - United States
  - The International Criminal Court
  - Russia
  - Israel

B.3 Cooperative internationalism

This block used the standard set of items for cooperative internationalism. This, too, was measured pre-treatment.

- Cooperative internationalism (agree/disagree, 5 point scale, order of items randomized)
  - It is essential for my country to work with other countries to solve problems such as overpopulation, hunger, and pollution.
  - It is important for countries to work together to tackle global challenges.
  - Countries should work together through international organizations.
  - Protecting the global environment is very important.
  - Helping to improve the standard of living in other countries is very important.

B.4 Post-treatment measures

The main manuscript already contains the exact treatment text. Post-treatment, we measured the respondent's posterior beliefs about Russian guilt and the trustworthiness of information sources. We randomized the order of the outcome measures and the order of items within outcome measures, where appropriate. For trustworthiness, we used the term "biased" to tap into the concept, without using the exact same words as the pre-treatment measures.

- Outcome: Russian Guilt How likely is it that Russian leaders have committed war crimes in Ukraine? (100 point scale)

- Outcome: Trustworthiness of Source Some sources of information are biased and others are not. On a scale of 0-100, with 100 being the most biased, where would you place the following sources of information?
  - The International Criminal Court
  - US government

- The media

After that, respondents answered two manipulation check questions and then demographic questions.

- Manipulation Checks
  - In one of the earlier questions, we asked about war crimes. Which country's leaders were accused of committing war crimes in that question? (Russia, USA, Guatemala)
  - In that same earlier question, who was accusing Russian leaders of war crimes? (The International Criminal Court, The US government, The Ukranian government)

- Demographics
  - Which political party do you feel most closely represents your views? (The lists varied by country.)
  - What is the highest level of education you have completed? (9 point scale, ranging from "No formal schooling" to "Post-graduate")
  - What is your current working status? (6 standard options)
  - What is your approximate monthly income? (12 point scale, currency and ranges varied by country)
  - In political matters, people talk of "the left" and "the right." Please tell me where would you place your views on a 10-point scale where 1 is the 'left' and 10 is the 'right'? (10 point scale)

B.5 Summary statistics

This figure shows the distribution of responses to the pre-treatment questions about prior Russian guilt and perceptions of the two sources. The two things are positively correlated, but not completely so. Figure B.1 shows the distribution of these two things, with all countries pooled. Respondents generally think Russia is guilty, as seen by more dots clustered in the top half of the figures. Respondents also tend to have greater trust in the ICC, as seen by more dots clustered on the right hand side for the ICC. Plenty of respondents choose "typical" answers, like 50 and 100, as seen by respondents located around the borders.

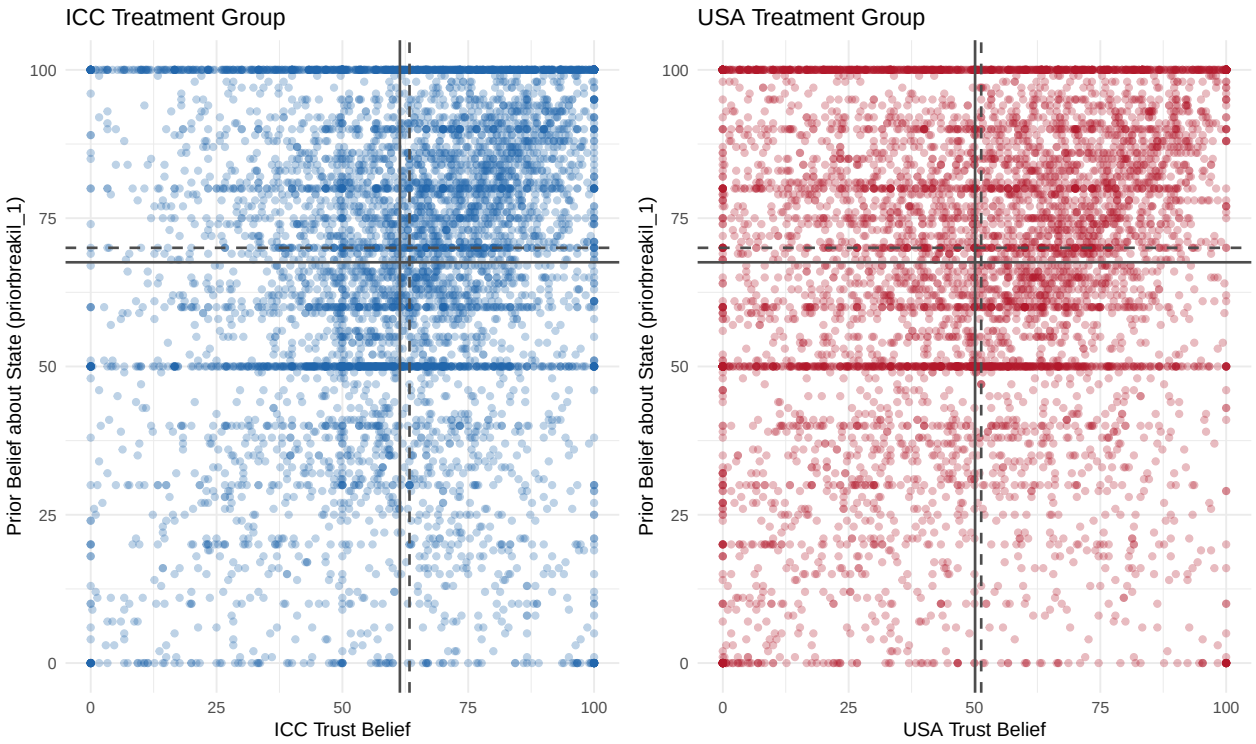


Figure B.1: Distribution of priors about Russian guilt and trust in the source, all countries.

B.6 Country specific summary information

Figure B.2 shows how countries differed in their pre-treatment beliefs about Russian guilt.

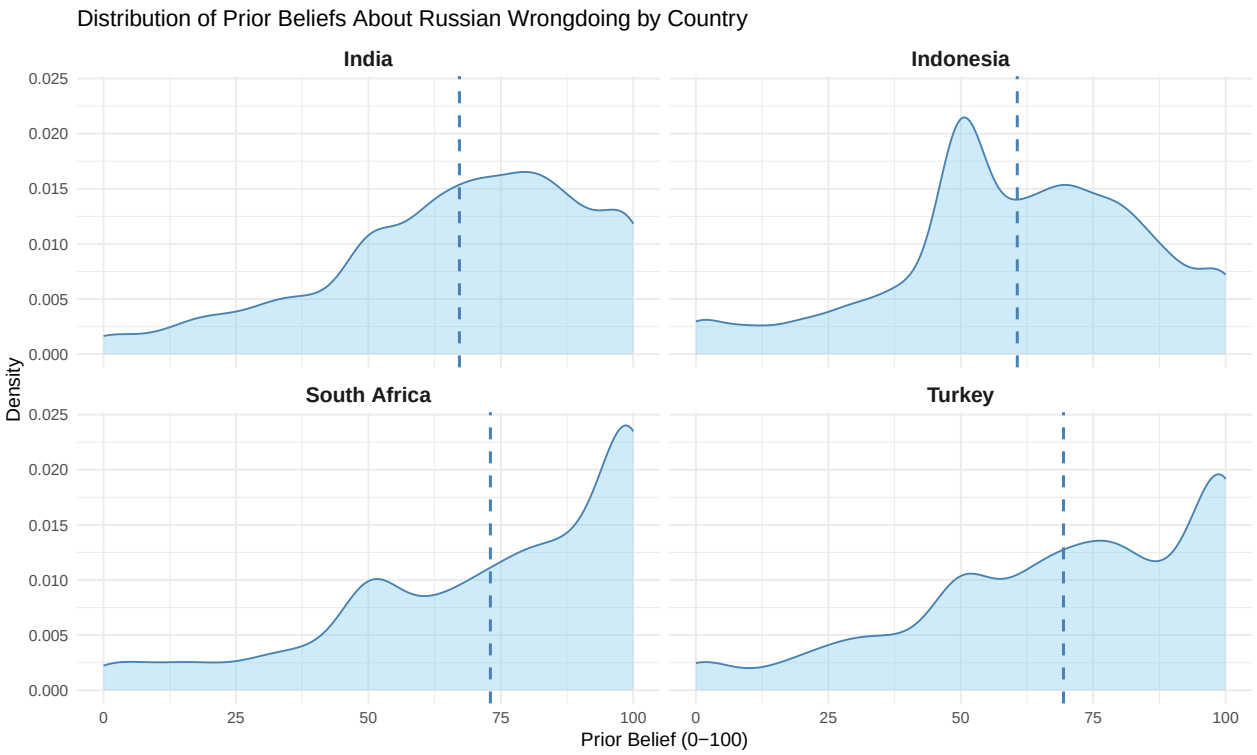


Figure B.2: Distribution of priors about Russian guilt by country.

B.7 Support for policy options

Figure B.3 and Figure B.4 show baseline levels of support for the different policy responses. These summary statistics are for respondents in the control group. Baseline support was generally higher for non-military aid, whereas support for military aid and sanctions appeared more split.

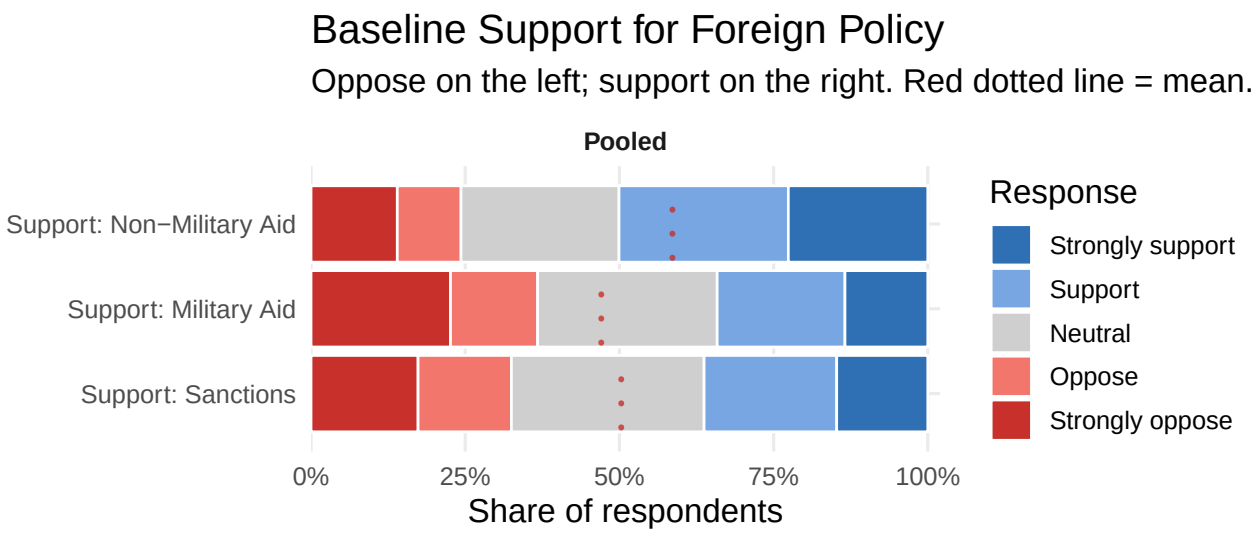


Figure B.3: Baseline Support for Foreign Policy (pooled)

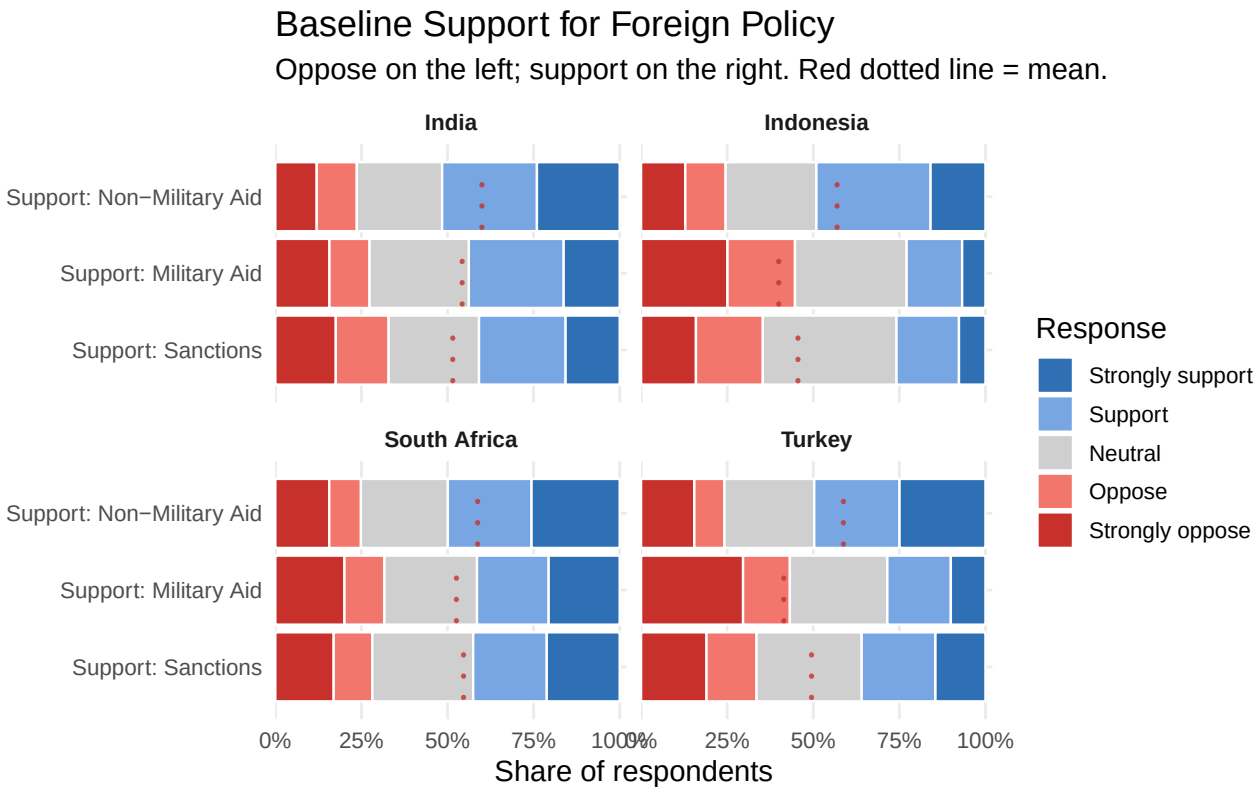


Figure B.4: Baseline Support for Foreign Policy (by country)

B.8 Manipulation Checks

We implemented two manipulation checks to assess whether the treatments had their intended effects. The first check asks, *"Which country's leaders were accused of committing war crimes in that question?"* The second asks, *"In that same earlier question, who was accusing Russian leaders of war crimes?"*. The second question was actually quite hard because the choices were "The Ukrainian government, The International Criminal Court, and The US government." Many respondents chose the Ukrainian government.

Table B.1: Mean Manipulation Check Pass Rates

sample	mani_pass1_mean	mani_pass2_mean
Indonesia	0.454	0.553
India	0.750	0.474
Turkey	0.799	0.639
South Africa	0.926	0.614
Overall	0.732	0.570

The pass rate for the first and second manipulation check was 73.2% and 57%, respectively.

B.9 Hypocrisy and Trust

Since our pre-treatment items also asked whether the respondents thought the United States had broken international law, we also looked at whether this was correlated with perceptions of trustworthiness of the United States. They are correlated. Table B.2 shows results from regressing trust in the United States on the respondent's answer to the question about whether the United States had broken international law (numerical). Respondents who thought the United States had broken international law were less trusting of it as an information source.

Table B.2: Beliefs About U.S. Violation of International Law and Trust in the U.S.

	Dependent variable:	
	Trust in the United States	
	No Controls	With Controls
	(1)	(2)
U.S. Violated Intl. Law	−0.219***	−0.178***
	(0.012)	(0.013)
Age		−0.073**
		(0.032)
Female		4.931***
		(0.698)
Education		−1.093***
		(0.177)
Income		0.196
		(0.136)
Voted for Incumbent		13.080***
		(0.730)
Constant	65.012***	66.745***
	(0.877)	(2.007)
Observations	6,553	5,415
Note:	*p<0.1; **p<0.05; ***p<0.01	

B.10 Balance tests

These are balance tests using the approach in Hansen and Bowers (2008). Samples are generally well-balanced in key covariates, while there are some imbalances in individual covariates. For example, there were more women in the ICC treatment group in Turkey, compared to the control group. In Indonesia, respondents in the USA treatment group had slightly higher incomes than the control group. These are unlikely to have had major effects on treatment effects.

Table B.3: ICC Treatment Balance Test

country	covariate	Control	Treated	Std.Diff	Z	p.value
Indonesia	age	36.748	36.102	-0.060	-0.860	0.390
	education	8.222	8.367	0.060	0.863	0.388
	female	0.486	0.512	0.051	0.733	0.464
	income_numeric	6.057	6.415	0.111	1.601	0.109
	incumbent	0.187	0.206	0.048	0.689	0.491
South Africa	age	36.327	36.075	-0.023	-0.355	0.723
	education	5.681	5.617	-0.043	-0.679	0.497
	female	0.544	0.511	-0.065	-1.014	0.310
	income_numeric	6.744	6.634	-0.054	-0.850	0.395
	incumbent	0.325	0.312	-0.029	-0.444	0.657
India	age	35.735	35.583	-0.014	-0.222	0.824
	education	7.450	7.333	-0.082	-1.314	0.189
	female	0.466	0.481	0.031	0.500	0.617
	income_numeric	4.136	3.977	-0.063	-1.017	0.309
	incumbent	0.647	0.600	-0.096	-1.546	0.122
Turkey	age	38.437	38.106	-0.030	-0.427	0.669
	education	8.518	8.426	-0.078	-1.168	0.243
	female	0.445	0.527	0.165	2.335	0.020
	income_numeric	5.201	5.150	-0.033	-0.490	0.624
	incumbent	0.239	0.222	-0.041	-0.573	0.567

Table B.4: USA Treatment Balance Test

country	covariate	Control	Treated	Std.Diff	Z	p.value
Indonesia	age	36.748	36.649	-0.009	-0.133	0.894
	education	8.222	8.278	0.023	0.326	0.745
	female	0.486	0.452	-0.068	-0.973	0.330
	income_numeric	6.057	6.644	0.186	2.670	0.008
	incumbent	0.187	0.172	-0.039	-0.557	0.578
South Africa	age	36.327	36.021	-0.027	-0.424	0.671
	education	5.681	5.716	0.024	0.379	0.705
	female	0.544	0.514	-0.061	-0.947	0.344
	income_numeric	6.744	6.737	-0.004	-0.057	0.954
	incumbent	0.325	0.303	-0.049	-0.752	0.452
India	age	35.735	35.890	0.014	0.220	0.826
	education	7.450	7.419	-0.023	-0.369	0.712
	female	0.466	0.478	0.025	0.399	0.690
	income_numeric	4.136	4.156	0.008	0.121	0.904
	incumbent	0.647	0.675	0.059	0.945	0.344
Turkey	age	38.437	38.631	0.018	0.250	0.802
	education	8.518	8.304	-0.173	-2.509	0.012
	female	0.445	0.508	0.126	1.773	0.076
	income_numeric	5.201	5.118	-0.053	-0.780	0.435
	incumbent	0.239	0.231	-0.018	-0.251	0.802

C Robustness

This section describes robustness checks for the main results and the regression tables for places where we reported or plotted coefficients.

C.1 Aggregate effect on posteriors: regressions from main figures

The main manuscript showed coefficient plots for aggregate treatment effects. Here, we show the full regression results for those estimates, and additional specifications that add respondent-level characteristics as controls. Table C.1 shows the regression results when we regress posteriors about Russian guilt on the ICC and USA treatments together. In other words, these regressions compare the two treatment groups with the control group. Table C.2 shows the same thing, only with support for the policy responses as the outcome measures. Table C.3 and Table C.4 do the same thing, only they exclude control group respondents. In other words, they compare outcomes between the ICC and USA treatment groups only.

Table C.1: Effect of Treatment on War Crimes Beliefs

	Dependent variable:	
	Russia Committed War Crimes	
	No Controls	With Controls
	(1)	(2)
ICC Treatment	−0.282	−0.810
	(0.788)	(0.843)
USA Treatment	−2.131***	−2.367***
	(0.790)	(0.846)
Age		0.031
		(0.032)
Female		7.669***
		(0.691)
Education		−1.435***
		(0.175)
Income		0.191
		(0.135)
Vote for Incumb.		0.801
		(0.723)
Constant	68.722***	74.241***
	(0.558)	(1.883)
Observations	6,508	5,415
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table C.2: Effect of Treatment on Policy Preferences

	Dependent variable:					
	Non-mil. Aid	Non-mil. Aid	Mil. Aid	Mil. Aid	Sanctions	Sanctions
	(1)	(2)	(3)	(4)	(5)	(6)
ICC Treatment	0.068*	0.057	0.019	0.052	0.063	0.013
	(0.039)	(0.040)	(0.039)	(0.042)	(0.044)	(0.043)
USA Treatment	−0.025	−0.024	−0.022	−0.067	−0.039	−0.042
	(0.039)	(0.040)	(0.039)	(0.043)	(0.044)	(0.043)
Age				0.010***	−0.009***	−0.001
				(0.002)	(0.002)	(0.002)
Female				0.016	0.347***	0.349***
				(0.035)	(0.036)	(0.035)
Education				0.039***	−0.024***	0.004
				(0.009)	(0.009)	(0.009)
Income				0.007	−0.002	0.011
				(0.007)	(0.007)	(0.007)
Vote for Incumb.				−0.108***	0.219***	0.006
				(0.036)	(0.037)	(0.037)
Constant	3.344***	2.882***	3.012***	2.745***	3.190***	2.824***
	(0.027)	(0.028)	(0.028)	(0.095)	(0.097)	(0.096)
Observations	6,516	6,516	6,516	5,415	5,415	5,415

Note: *p<0.1; **p<0.05; ***p<0.01

Table C.3: Effect of ICC Treatment on War Crimes Beliefs (Restricted to only ICC/USA conditions)

	Dependent variable:	
	Russia Committed War Crimes	
	No Controls	With Controls
	(1)	(2)
ICC Treatment	1.849**	1.548*
	(0.792)	(0.851)
Age		0.046
		(0.039)
Female		7.431***
		(0.853)
Education		−1.316***
		(0.216)
Income		0.084
		(0.165)
Vote for Incumb.		0.554
		(0.894)
Constant	66.591***	71.251***
	(0.561)	(2.272)
Observations	4,344	3,614
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table C.4: Effect of ICC Treatment on Policy Preferences (Restricted to only ICC/USA conditions)

	Dependent variable:					
	Non-mil. Aid	Non-mil. Aid	Mil. Aid	Mil. Aid	Sanctions	Sanctions
	(1)	(2)	(3)	(4)	(5)	(6)
ICC Treatment	0.093**	0.082**	0.040	0.119***	0.103**	0.055
	(0.038)	(0.040)	(0.039)	(0.042)	(0.044)	(0.043)
Age				0.011***	−0.008***	−0.0004
				(0.002)	(0.002)	(0.002)
Female				0.009	0.322***	0.341***
				(0.042)	(0.044)	(0.043)
Education				0.053***	−0.018	0.003
				(0.011)	(0.011)	(0.011)
Income				0.003	−0.006	0.009
				(0.008)	(0.008)	(0.008)
Vote for Incumb.				−0.094**	0.228***	−0.014
				(0.044)	(0.046)	(0.046)
Constant	3.319***	2.858***	2.990***	2.555***	3.116***	2.784***
	(0.027)	(0.028)	(0.028)	(0.112)	(0.116)	(0.116)
Observations	4,349	4,349	4,349	3,614	3,614	3,614

Note: *p<0.1; **p<0.05; ***p<0.01

C.2 Aggregate effect on posteriors, with country-specific intercepts

To show aggregate effects of treatment on posterior beleifs about Russian guilt, the main manuscript showed results with a single intercept. Figure C.1 shows results with a country-specific intercept. Results are very similar.

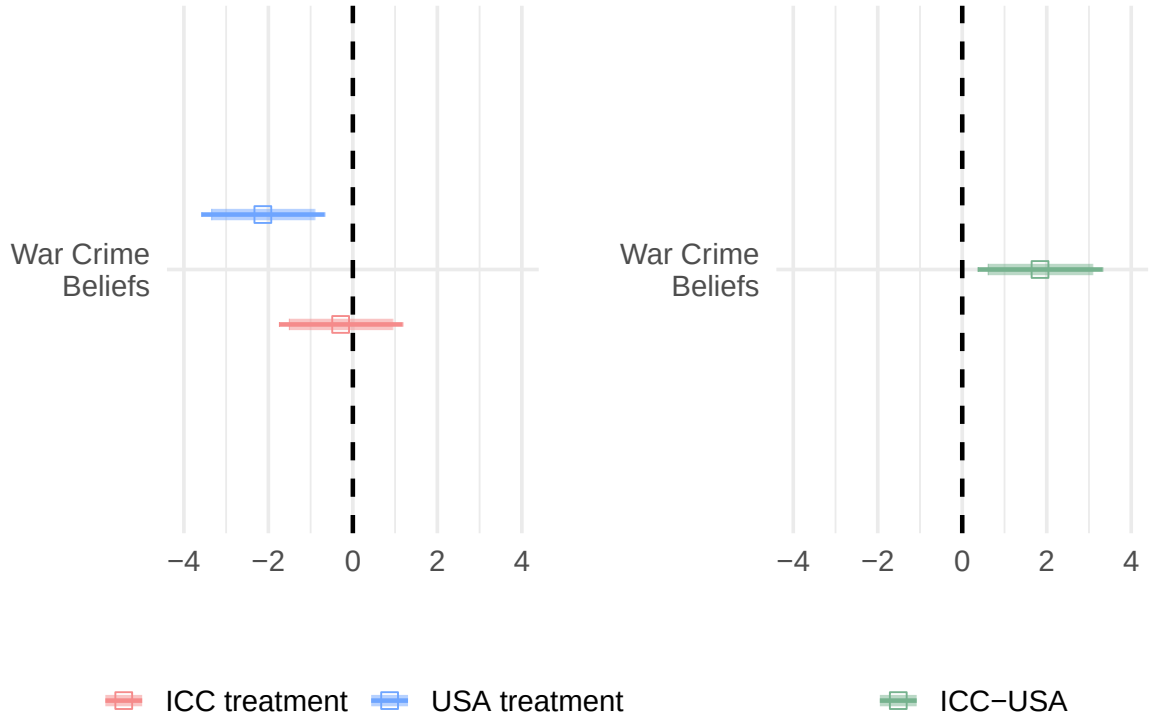
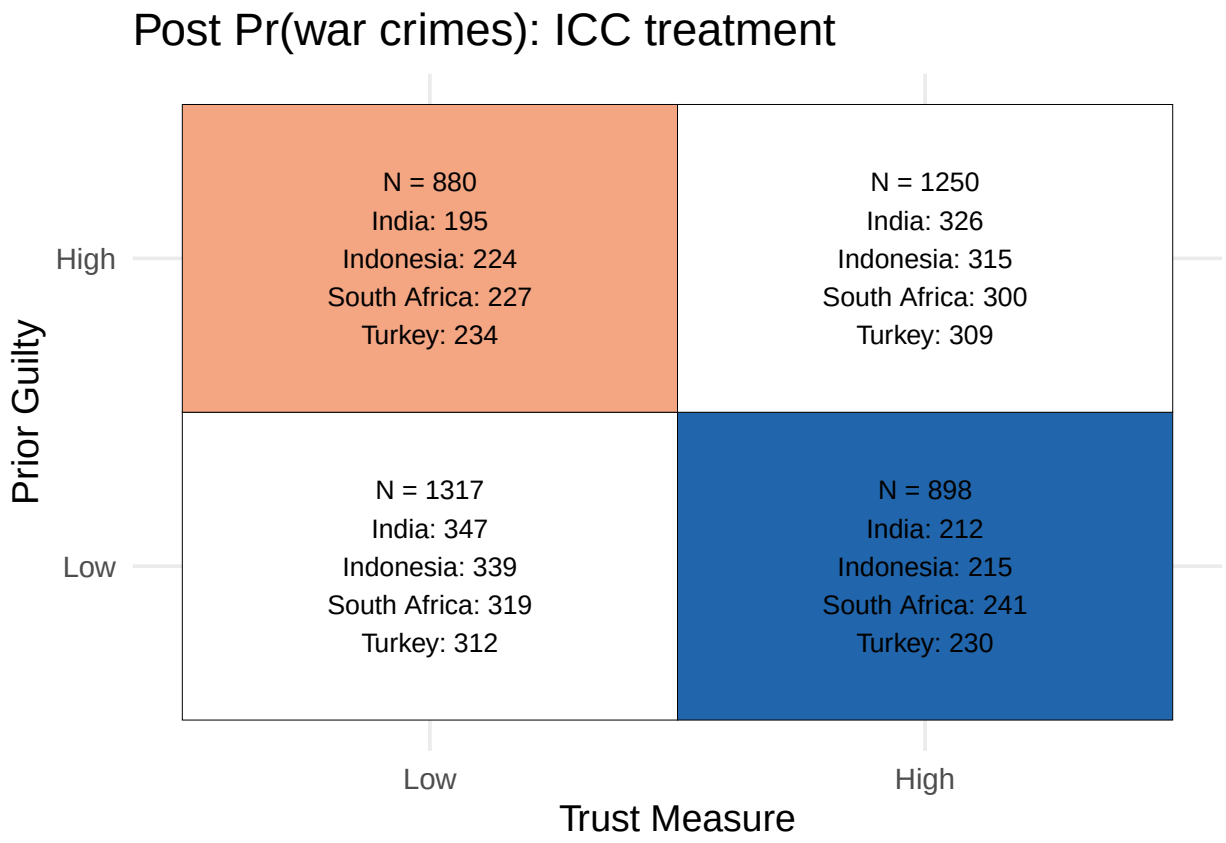
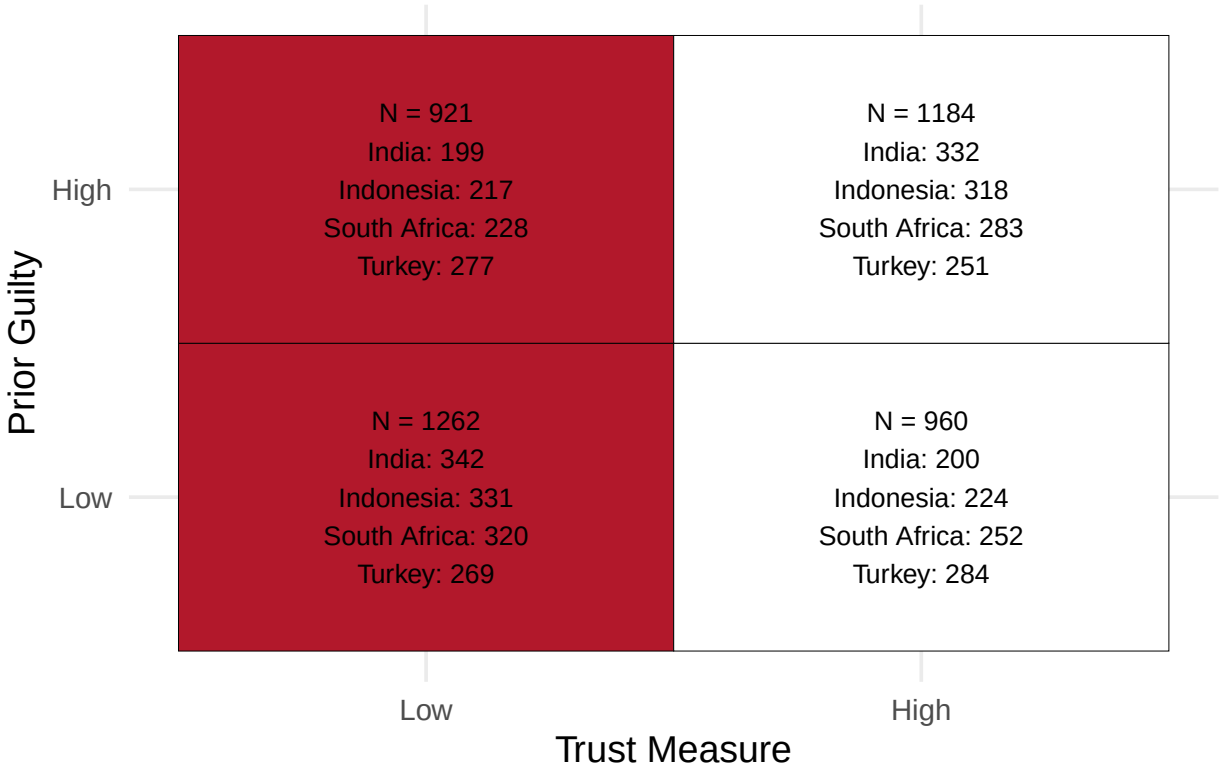


Figure C.1: Treatment effects on posteriors about Russian guilt, with country-specific intercepts.

C.3 Hypothesis 1: Sample sizes in the boxes

There are different sample sizes in the four quadrants. The figure below shows the number of observations in each cell, with the same coloring as the first figure. The top pane is for the ICC versus control analysis. The bottom pane is for the US versus control group analysis. We broke out the sample sizes by country in each cell, as well.




Post Pr(war crimes): USA treatment
N = 921
India: 199
Indonesia: 217
South Africa: 228
Turkey: 277
N = 1184
India: 332
Indonesia: 318
South Africa: 283
Turkey: 251
N = 1262
India: 342
Indonesia: 331
South Africa: 320
Turkey: 269
N = 960
India: 200
Indonesia: 224
South Africa: 252
Turkey: 284
High
Low
Prior Guilty
Low
High
Trust Measure

C.4 Hypothesis 1: Alternate specifications for the boxes

Evaluating Hypothesis 1 requires making decisions about how to compare the surface in Figure 1 with the empirical results. Recall, Figure 1 shows the predicted treatment effect – posterior beliefs about the state of the world minus priors – as priors and trust in the source vary. There is no simple linear estimation strategy to ask "do the treatment effects reflect the predictions in Figure 1?" In the main manuscript, we estimated treatment effects by regressing outcomes on a treatment indicator interacted with indicator variables for which of the four cells the respondent was in. The main manuscript shows results when cells are defined by country-specific medians for the priors beliefs and trust measures. There are different possible ways to define the cells. Respondents could be assigned to cells based on whether they were above or below a *global* median, for example. The regression itself could include country fixed effects and or control variables. These are each different ways to approximate the contoured surface in Figure 1.

Here, we describe results from a wide array of alternate specifications. In general, the specifications showed in the main manuscript are representative of those here. Point estimates vary and statistical significance does change for some parts of some specifications. But the overall patterns are similar: persuasion is strongest in the bottom right and backlash is most prominent in the top left. The ICC effects are more split between persuasion and backlash, which the U.S. treatment triggers more backlash overall.

C.4.1 H1: Boxes based on universal medians

The figures below replicate the boxes from the main manuscripts, but rather than classifying respondents in relation to a country-specific median, they use a global median.

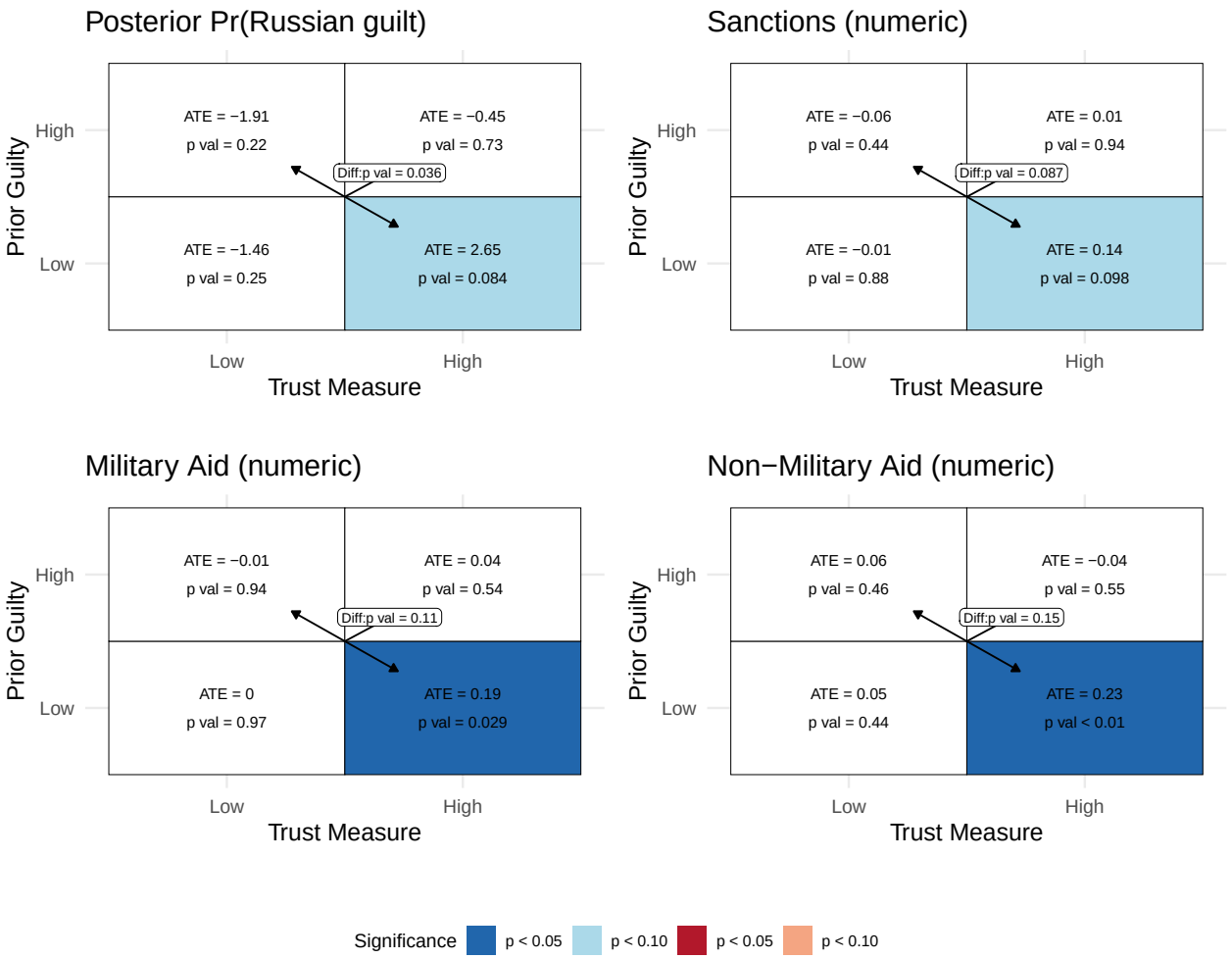


Figure C.2: Global Median - ICC Treatment

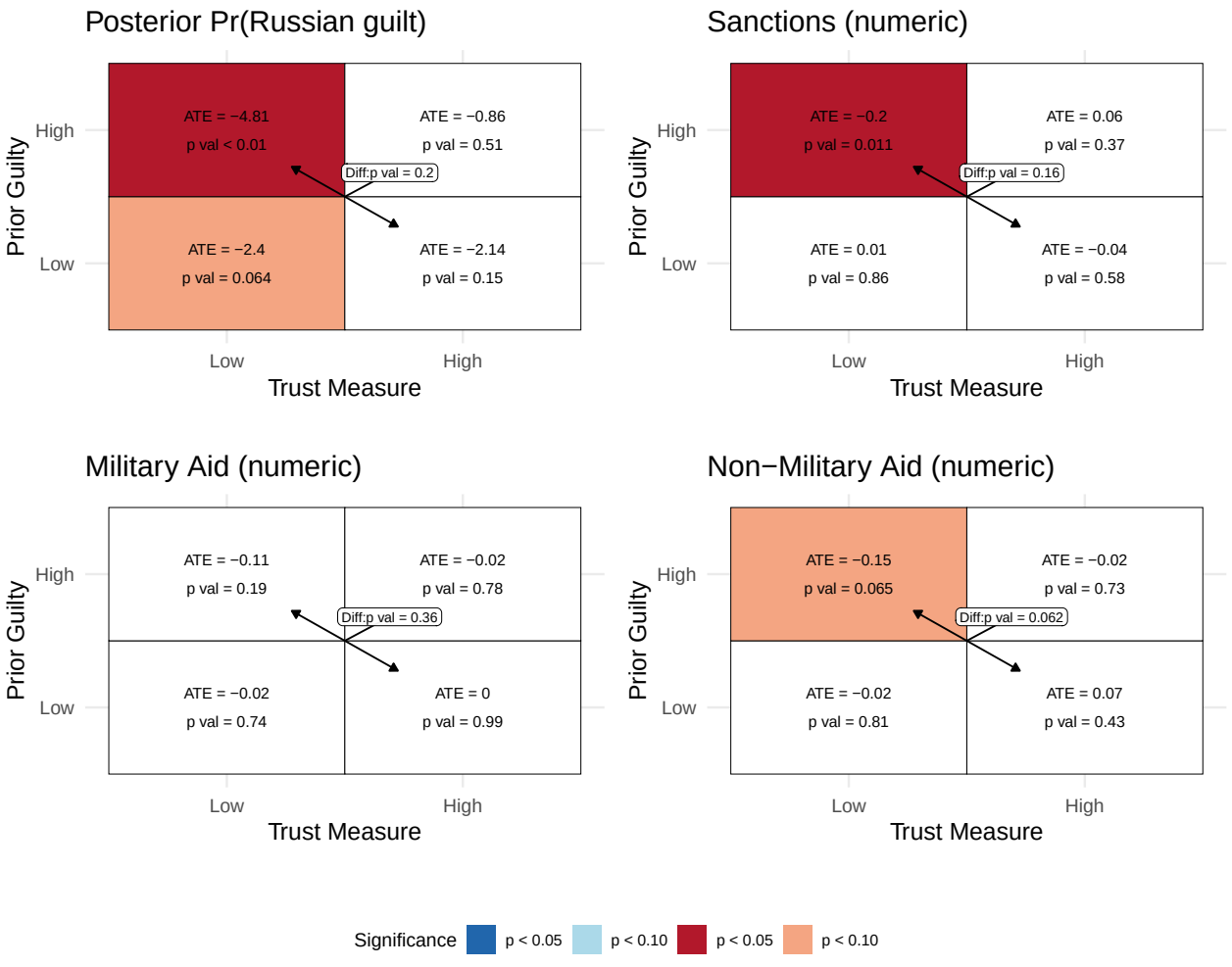
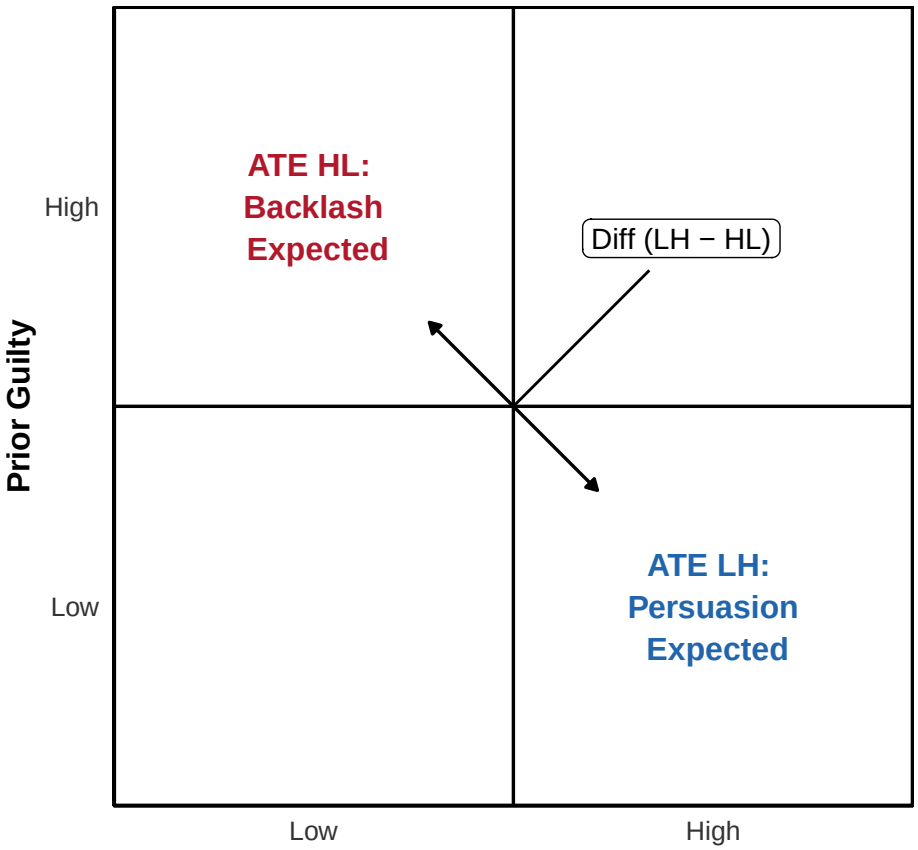


Figure C.3: Global Median - USA Treatment

C.4.2 H1: ICC Treatment, Tables of different specifications

The tables below show how the treatment effects in each quadrant of the data vary, depending on decisions about how to define the quadrants or estimate ef fects within quadrants. Each table has the same format. The four rows correspond to the four outcome measures we consider, posterior beliefs about Russia and the three policy responses. Columns 2-4 describe the particular empirical decision made for those estimates: whether to use medians or means, whether to use demographic controls, whether to use country fixed effects. Columns 6 and 7 describe the estimated treatment effect for the top left and bottom right quadrants.

We labeled them "HL" and "LH", where "HL" means "the respondent is high in terms of their priors about Russia and low in their trust in the source". "LH" means "the respondent is low in their prior beliefs about Russian guilt and high in their trust of the source." Estimated treatment effects for HL should be negative (ie backlash) and estimates for LH should be positive (ie persuasion). Finally, Column 8 shows the statistical test for whether treatment effects are different in the HL versus LH quadrants. In other words, they show the same thing as the annotations on the diagonals of figures like Figure 7.



There are lots of specifications. The key takeaway is that the estimates are similar to those in the main manuscript. These estimation choices tend to all suggest persuasion where expected (positive coefficients), backlash where expected (negative coefficients), and a difference in the estimates for the two quadrants.

Table C.5: ATEs and Settings: ICC vs Control — Median split, no controls, no FE

					Estimates		
1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	6) ATE HL	7) ATE LH	8) Diff (LH–HL)
ICC vs Control	Median	No	No	War Crime Beliefs	-3.11 (p-val = 0.05)	3.26 (p-val = 0.04)	6.37 (p-val = 0.00)
ICC vs Control	Median	No	No	Support: Sanctions	-0.16 (p-val = 0.05)	0.11 (p-val = 0.18)	0.28 (p-val = 0.02)
ICC vs Control	Median	No	No	Support: Military Aid	-0.01 (p-val = 0.90)	0.24 (p-val = 0.01)	0.25 (p-val = 0.05)
ICC vs Control	Median	No	No	Support: Non-military Aid	-0.00 (p-val = 0.99)	0.25 (p-val = 0.00)	0.25 (p-val = 0.04)

*

Note:
Positive (negative) coefficients show persuasion (backlash). HL = High Prior & Low Trust; LH = Low Prior & High Trust Diff = LH − HL.

Table C.6: ATEs and Settings: ICC vs Control — Mean split, no controls, no FE

1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	Estimates		
					6) ATE HL	7) ATE LH	8) Diff (LH–HL)
ICC vs Control	Mean	No	No	War Crime Beliefs	-2.64 (p-val = 0.08)	2.31 (p-val = 0.16)	4.96 (p-val = 0.03)
ICC vs Control	Mean	No	No	Support: Sanctions	-0.16 (p-val = 0.05)	0.08 (p-val = 0.34)	0.25 (p-val = 0.04)
ICC vs Control	Mean	No	No	Support: Military Aid	-0.06 (p-val = 0.49)	0.17 (p-val = 0.06)	0.23 (p-val = 0.06)
ICC vs Control	Mean	No	No	Support: Non-military Aid	0.04 (p-val = 0.62)	0.20 (p-val = 0.02)	0.16 (p-val = 0.17)

*

Note:
Positive (negative) coefficients show persuasion (backlash). HL = High Prior & Low Trust; LH = Low Prior & High Trust Diff = LH – HL.

Table C.7: ATEs and Settings: ICC vs Control — Median split, with controls, no FE

1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	Estimates		
					6) ATE HL	7) ATE LH	8) Diff (LH–HL)
ICC vs Control	Median	Yes	No	War Crime Beliefs	-3.18 (p-val = 0.05)	2.85 (p-val = 0.07)	6.02 (p-val = 0.01)
ICC vs Control	Median	Yes	No	Support: Sanctions	-0.16 (p-val = 0.06)	0.12 (p-val = 0.15)	0.28 (p-val = 0.02)
ICC vs Control	Median	Yes	No	Support: Military Aid	-0.02 (p-val = 0.78)	0.22 (p-val = 0.01)	0.25 (p-val = 0.04)
ICC vs Control	Median	Yes	No	Support: Non-military Aid	-0.02 (p-val = 0.78)	0.25 (p-val = 0.00)	0.28 (p-val = 0.02)

*

Note:
Positive (negative) coefficients show persuasion (backlash). HL = High Prior & Low Trust; LH = Low Prior & High Trust Diff = LH – HL.

Table C.8: ATEs and Settings: ICC vs Control — Median split, no controls, with FE

1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	Estimates		
					6) ATE HL	7) ATE LH	8) Diff (LH–HL)
ICC vs Control	Median	No	Yes	War Crime Beliefs	-2.64 (p-val = 0.08)	3.46 (p-val = 0.02)	6.10 (p-val = 0.00)
ICC vs Control	Median	No	Yes	Support: Sanctions	-0.16 (p-val = 0.06)	0.12 (p-val = 0.15)	0.28 (p-val = 0.02)

Table C.8: ATEs and Settings: ICC vs Control — Median split, no controls, with FE *(continued)*

1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	6) ATE HL	7) ATE LH	8) Diff (LH–HL)
ICC vs Control	Median	No	Yes	Support: Military Aid	-0.01 (p-val = 0.87)	0.26 (p-val = 0.00)	0.27 (p-val = 0.03)
ICC vs Control	Median	No	Yes	Support: Non-military Aid	0.00 (p-val = 0.97)	0.25 (p-val = 0.00)	0.24 (p-val = 0.04)

*

Note:
Positive (negative) coefficients show persuasion (backlash). HL = High Prior & Low Trust; LH = Low Prior & High Trust Diff = LH – HL.

Table C.9: ATEs and Settings: ICC vs Control — Median split, with controls, with FE

					Estimates		
1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	6) ATE HL	7) ATE LH	8) Diff (LH–HL)
ICC vs Control	Median	Yes	Yes	War Crime Beliefs	-2.91 (p-val = 0.05)	3.00 (p-val = 0.05)	5.91 (p-val = 0.01)
ICC vs Control	Median	Yes	Yes	Support: Sanctions	-0.17 (p-val = 0.05)	0.13 (p-val = 0.12)	0.30 (p-val = 0.01)
ICC vs Control	Median	Yes	Yes	Support: Military Aid	-0.03 (p-val = 0.72)	0.24 (p-val = 0.00)	0.27 (p-val = 0.03)
ICC vs Control	Median	Yes	Yes	Support: Non-military Aid	-0.02 (p-val = 0.79)	0.26 (p-val = 0.00)	0.28 (p-val = 0.02)

*

Note:
Positive (negative) coefficients show persuasion (backlash). HL = High Prior & Low Trust; LH = Low Prior & High Trust Diff = LH – HL.

C.4.3 H1: USA Treatment, Tables of different specifications

The tables below show the same thing as the preceding sub-section, only for the USA treatment. Here, too, the key takeaway is that the results are similar to those in the main manuscript's specifications. The U.S. treatment triggers more backlash, overall, compared to the ICC treatment.

Table C.10: ATEs and Settings: usa vs Control — Median split, no controls, no FE

					Estimates		
1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	6) ATE HL	7) ATE LH	8) Diff (LH–HL)
USA vs Control	Median	No	No	War Crime Beliefs	-4.03 (p-val = 0.01)	-0.70 (p-val = 0.64)	3.32 (p-val = 0.12)
USA vs Control	Median	No	No	Support: Sanctions	-0.22 (p-val = 0.01)	-0.03 (p-val = 0.71)	0.19 (p-val = 0.10)

Table C.10: ATEs and Settings: usa vs Control — Median split, no controls, no FE
(continued)

1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	6) ATE HL	7) ATE LH	8) Diff (LH–HL)
USA vs Control	Median	No	No	Support: Military Aid	-0.12 (p-val = 0.15)	-0.04 (p-val = 0.63)	0.08 (p-val = 0.49)
USA vs Control	Median	No	No	Support: Non-military Aid	-0.12 (p-val = 0.15)	0.10 (p-val = 0.20)	0.23 (p-val = 0.05)

*

Note:
Positive (negative) coefficients show persuasion (backlash). HL = High Prior & Low Trust; LH = Low Prior & High Trust Diff = LH – HL.

Table C.11: ATEs and Settings: usa vs Control — Mean split, no controls, no FE

					Estimates		
1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	6) ATE HL	7) ATE LH	8) Diff (LH–HL)
USA vs Control	Mean	No	No	War Crime Beliefs	-3.82 (p-val = 0.01)	0.09 (p-val = 0.95)	3.91 (p-val = 0.07)
USA vs Control	Mean	No	No	Support: Sanctions	-0.21 (p-val = 0.01)	-0.08 (p-val = 0.33)	0.13 (p-val = 0.25)
USA vs Control	Mean	No	No	Support: Military Aid	-0.11 (p-val = 0.18)	-0.01 (p-val = 0.87)	0.09 (p-val = 0.43)
USA vs Control	Mean	No	No	Support: Non-military Aid	-0.16 (p-val = 0.05)	0.08 (p-val = 0.37)	0.24 (p-val = 0.05)

*

Note:
Positive (negative) coefficients show persuasion (backlash). HL = High Prior & Low Trust; LH = Low Prior & High Trust Diff = LH – HL.

Table C.12: ATEs and Settings: usa vs Control — Median split, with controls, no FE

					Estimates		
1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	6) ATE HL	7) ATE LH	8) Diff (LH–HL)
USA vs Control	Median	Yes	No	War Crime Beliefs	-3.61 (p-val = 0.02)	-1.59 (p-val = 0.29)	2.01 (p-val = 0.35)
USA vs Control	Median	Yes	No	Support: Sanctions	-0.21 (p-val = 0.01)	-0.02 (p-val = 0.79)	0.19 (p-val = 0.10)
USA vs Control	Median	Yes	No	Support: Military Aid	-0.14 (p-val = 0.11)	-0.04 (p-val = 0.60)	0.09 (p-val = 0.44)
USA vs Control	Median	Yes	No	Support: Non-military Aid	-0.12 (p-val = 0.16)	0.13 (p-val = 0.14)	0.25 (p-val = 0.04)

*

Table C.12: ATEs and Settings: usa vs Control — Median split, with controls, no FE *(continued)*

1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	6) ATE HL	7) ATE LH	8) Diff (LH–HL)

Note:
Positive (negative) coefficients show persuasion (backlash). HL = High Prior & Low Trust; LH = Low Prior & High Trust Diff = LH – HL.

Table C.13: ATEs and Settings: usa vs Control — Median split, no controls, with FE

					Estimates		
1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	6) ATE HL	7) ATE LH	8) Diff (LH–HL)
USA vs Control	Median	No	Yes	War Crime Beliefs	-2.81 (p-val = 0.05)	-1.13 (p-val = 0.42)	1.68 (p-val = 0.40)
USA vs Control	Median	No	Yes	Support: Sanctions	-0.20 (p-val = 0.01)	-0.03 (p-val = 0.70)	0.17 (p-val = 0.13)
USA vs Control	Median	No	Yes	Support: Military Aid	-0.10 (p-val = 0.21)	-0.03 (p-val = 0.75)	0.08 (p-val = 0.50)
USA vs Control	Median	No	Yes	Support: Non-military Aid	-0.12 (p-val = 0.16)	0.11 (p-val = 0.19)	0.23 (p-val = 0.05)

*

Note:
Positive (negative) coefficients show persuasion (backlash). HL = High Prior & Low Trust; LH = Low Prior & High Trust Diff = LH – HL.

Table C.14: ATEs and Settings: usa vs Control — Median split, with controls, with FE

					Estimates		
1) Comparison	2) Split	3) Ctrls?	4) FE?	5) DV	6) ATE HL	7) ATE LH	8) Diff (LH–HL)
USA vs Control	Median	Yes	Yes	War Crime Beliefs	-2.68 (p-val = 0.06)	-1.51 (p-val = 0.29)	1.17 (p-val = 0.56)
USA vs Control	Median	Yes	Yes	Support: Sanctions	-0.20 (p-val = 0.02)	-0.01 (p-val = 0.89)	0.19 (p-val = 0.11)
USA vs Control	Median	Yes	Yes	Support: Military Aid	-0.12 (p-val = 0.15)	-0.02 (p-val = 0.85)	0.10 (p-val = 0.37)
USA vs Control	Median	Yes	Yes	Support: Non-military Aid	-0.12 (p-val = 0.17)	0.13 (p-val = 0.13)	0.24 (p-val = 0.04)

*

Note:
Positive (negative) coefficients show persuasion (backlash). HL = High Prior & Low Trust; LH = Low Prior & High Trust Diff = LH – HL.

C.4.4 H1: Linear interaction terms

Since we used linear interaction term models in the Hypothesis 2 analysis, here are the results from those models where the beliefs about Russian guilt is the DV. The lines should be upward sloping and they are.

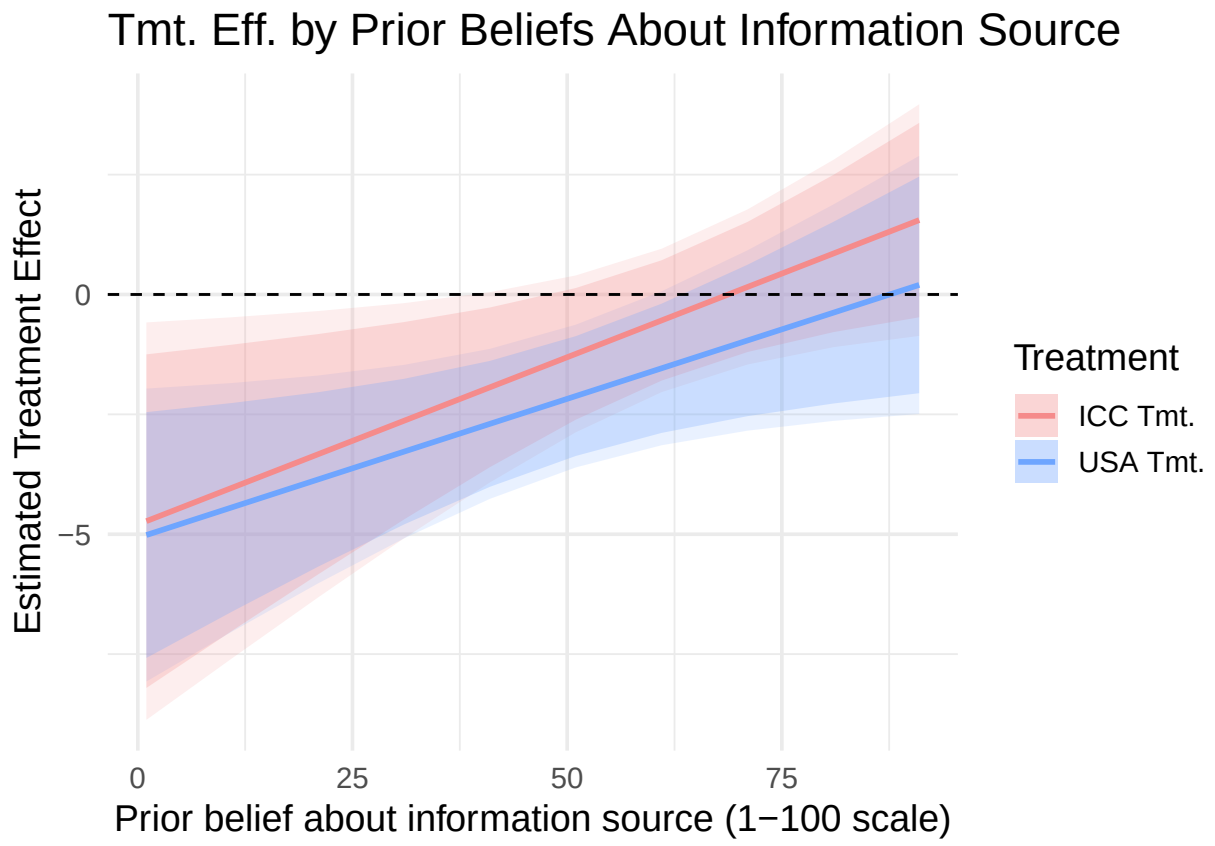


Figure C.4: Effect of treatment on posteriors about Russian guilt, as beliefs about the source vary.

C.5 Hypothesis 2: Treatment effects on ICC legitimacy

We also tested whether the ICC treatment influenced perceptions of ICC legitimacy. Figure C.5 shows these estimates graphically. The ICC treatment raised mean legitimacy scores by 0.12 points on a five-point scale (SE = 0.037, p = 0.0013), amounting to roughly a 3.6 percent increase.[67] While the effect size is modest, it is highly significant and consistent with our findings on source trust. When the ICC makes a statement about violations of international law, this meaningfully bolsters public perceptions of its legitimacy.

[67] We did not ask a question about U.S. legitimacy.

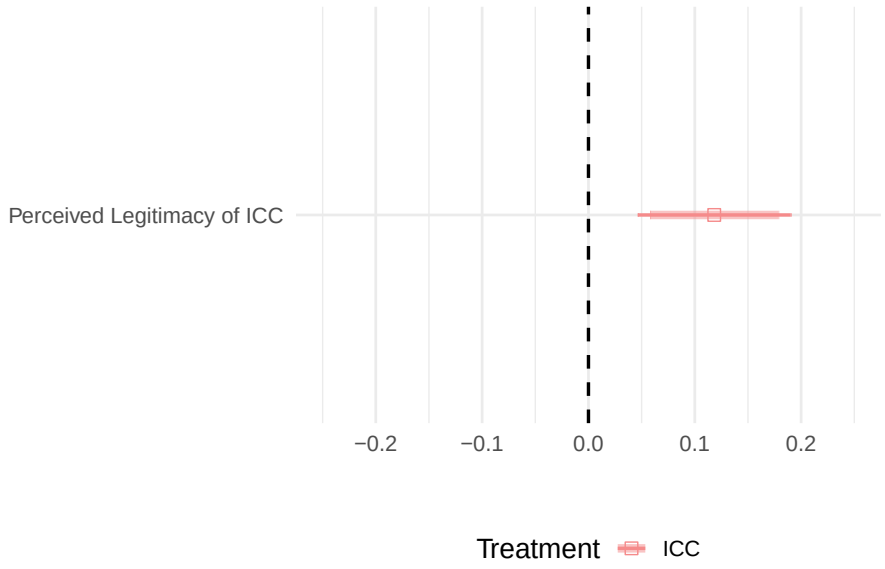


Figure C.5: Effect of treatment on perceptions of ICC legitimacy

D Appendix for cooperative internationalism moderation

Did cooperative internationalism moderate treatment effects? Our theoretical model suggests that the effects of CI (cooperative internationalism) are ambiguous. One the one hand, respondents that scored higher on the CI measures should show larger treatment effects for the ICC. Presumably, they should have higher pre-treatment beliefs about the trustworthiness of the ICC's information, which should magnify the ICC treatment effect. This is analogous to the argument most commonly found in existing research. On the other hand, they also are likely to already have higher prior beliefs about Russian guilt, which has a non-monotonic effect on the magnitude of predicted treatment effects. It could mute treatment effects for respondents that already strongly believe in Russian guilt. In our sample, both of these correlations were apparent. Higher CI respondents had higher beliefs about the trustworthiness of the ICC and higher prior beliefs in Russian guilt.

Our surveys included standard, pre-treatment cooperative internationalism items. We asked about the respondents' agreement (on a 5 point scale) with the statements: (1) "It is essential for my country to work with other countries to solve problems such as overpopulation, hunger, and pollution" (2) "It is important for countries to work together to tackle global challenges," (3) "Countries should work together through international organizations," (4) "Protecting the global environment is very important," and (5) "Helping to improve the standard of living in other countries is very important."[68]

Note that this same ambiguity applies to moderation based on partisanship.[69] A respondent's party identification could affect their perception of sources. In South Africa, the African National Congress (ANC) is generally less aligned with the United States than the Democratic Alliance (DA). On the one hand, this could mean that ANC members would be less moved by information from the United

[68] We randomized the order of these items. We did not ask about militant internationalism, since our focus was on international law.

[69] For examples where partisanship moderates the effects of an informational treatment, see Chaudoin (2023) and Brutger (2021).

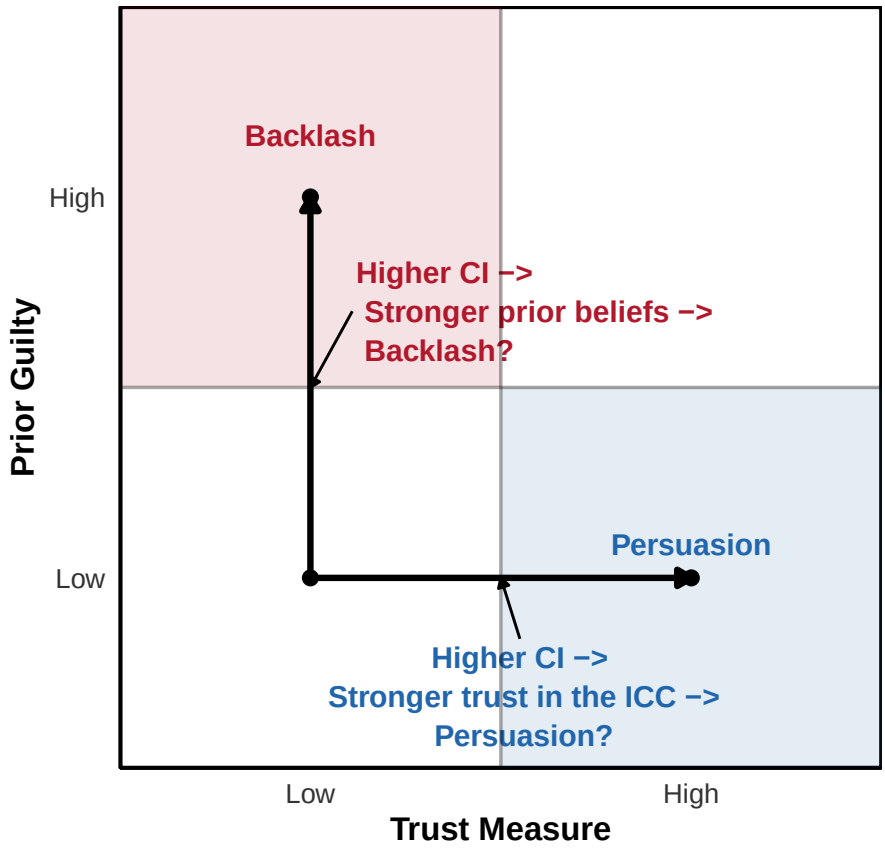


Figure D.1: Moderating Effects of Cooperative Internationalism

States. On the other hand, their members may not already have a deeply held belief that Russia is guilty of war crimes, which means their opinions are more movable. A DA member might be more trusting of the United States, which magnifies treatment effects. But they may already think Russia is guilty, muting treatment effects. Increasing trust in a source of information unambiguously increases the treatment effect of information from that source. But moving prior beliefs has a non-monotonic effect on treatment effects. Treatment effects are biggest for people with moderate prior beliefs. Since many moderating variables, like party identification, are correlated with both, their net effect is hard to predict, theoretically.

Figure D.2 shows the estimated treatment effects, broken down by whether respondents were above or below the average score on the CI items. The left pane shows effects on posterior beliefs about Russian war crimes. The right pane shows effects on the policy responses.

Cooperative internationalism has inconsistent moderating effects. Looking first at only the ICC treatment effects, for three of the four outcome measures, higher CI respondents had weaker ICC treatment effects. This is contrary to expectations that are based only on a theory that links CI with perceptions of an IO's credibility. On the other hand, this would be consistent with a theory that links CI to a ceiling effect, where high CI respondents already believe Russia is guilty, so they can't raise this posterior probability much higher.

Looking next at a comparison of ICC and USA treatment effects, the results are also inconsistent in their support or disconfirmation of arguments about CI. On the one hand, the ICC treatment effect was generally larger than the USA treatment effect for high CI respondents. However, the ICC effect was larger than the ICC effect among low CI respondents for two out of four outcomes (beliefs about Russian guilt and non-military aid). For those outcome measures, the difference between ICC and USA treatment effects for high CI respondents was comparable to the difference for low CI respondents.[70]

Our point here is not that CI contains no useful information or that it has no effect on attitudes towards foreign policies. On the contrary, it is well-correlated with important parameters, like prior beliefs about the world or about the trustworthiness of sources. CI is a good predictor of foreign policy attitudes. However, it is theoretically ambiguous as a moderator of treatment effects. CI is a bundle of things related to priors, and therefore its net impact on predicted treatment effects is theoretically ambiguous. CI also likely contains other things that moderate treatment effects in ways that go beyond Bayesian updating. In our application, this ambiguity was born out, even though CI was correlated with prior attitudes as expected.

Above, we stated that correlations between cooperative internationalism and prior beliefs about Russian guilt / perceptions of the ICC were as we would expect. We show that here. Higher CI respondents were more likely to believe that Russia was guilty *ex ante* and they had higher pre-treatment perceptions of the ICC.

[70] In the appendix, we also estimate 2x2 boxes using CI and prior beliefs as the two moderating variables. There are not clear patterns.

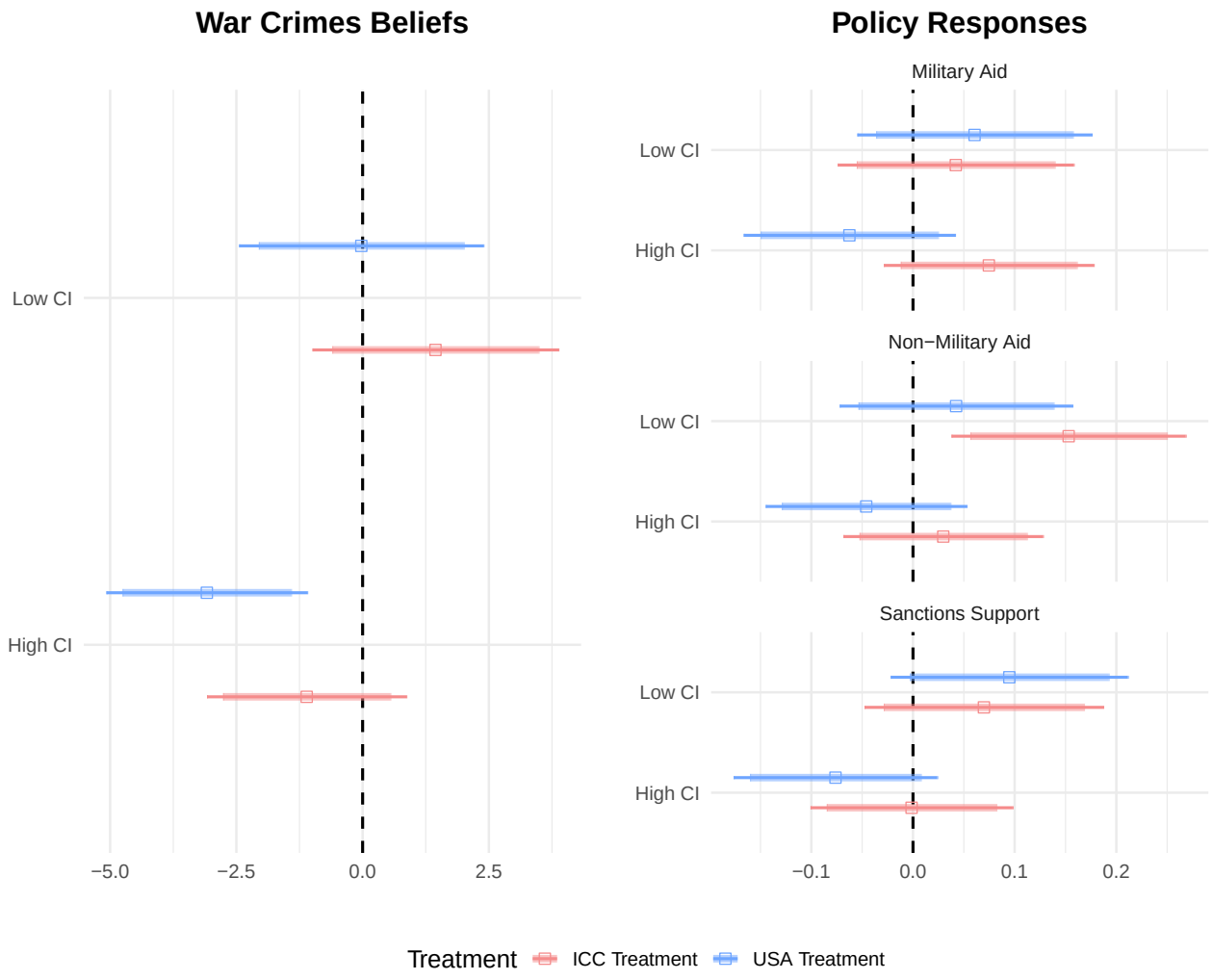
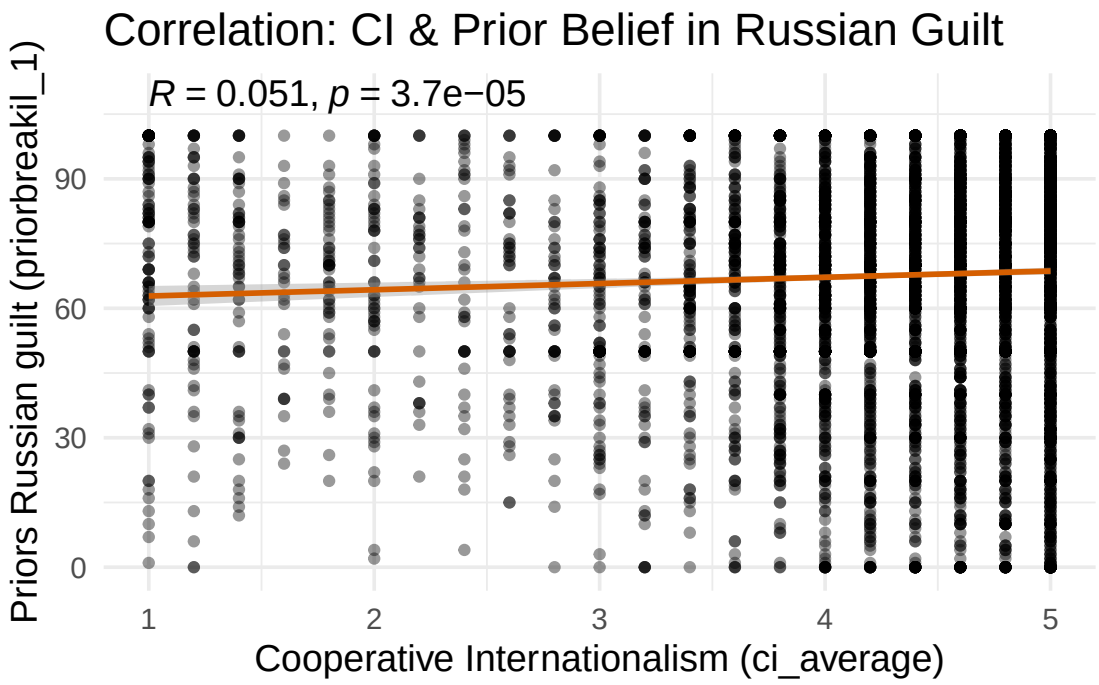
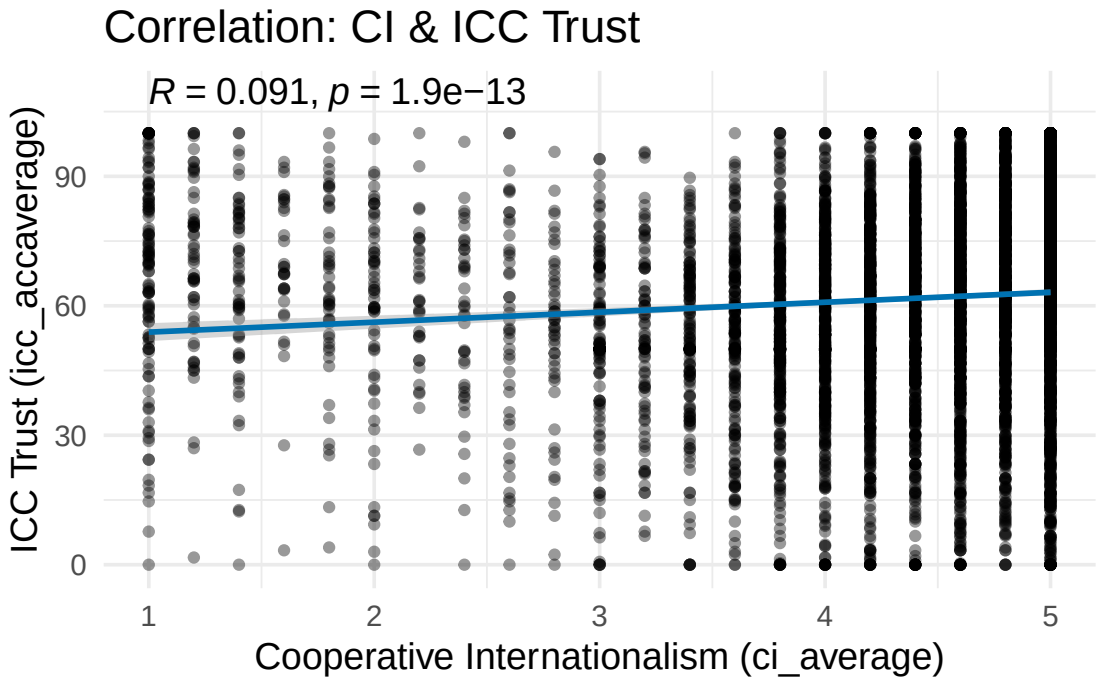


Figure D.2: Treatment effects, broken down by cooperative internationalism





We also re-estimated the box plots from the main manuscript, using the CI measure instead of our pre-treatment measures of source trustworthiness. We do not find the same patterns.

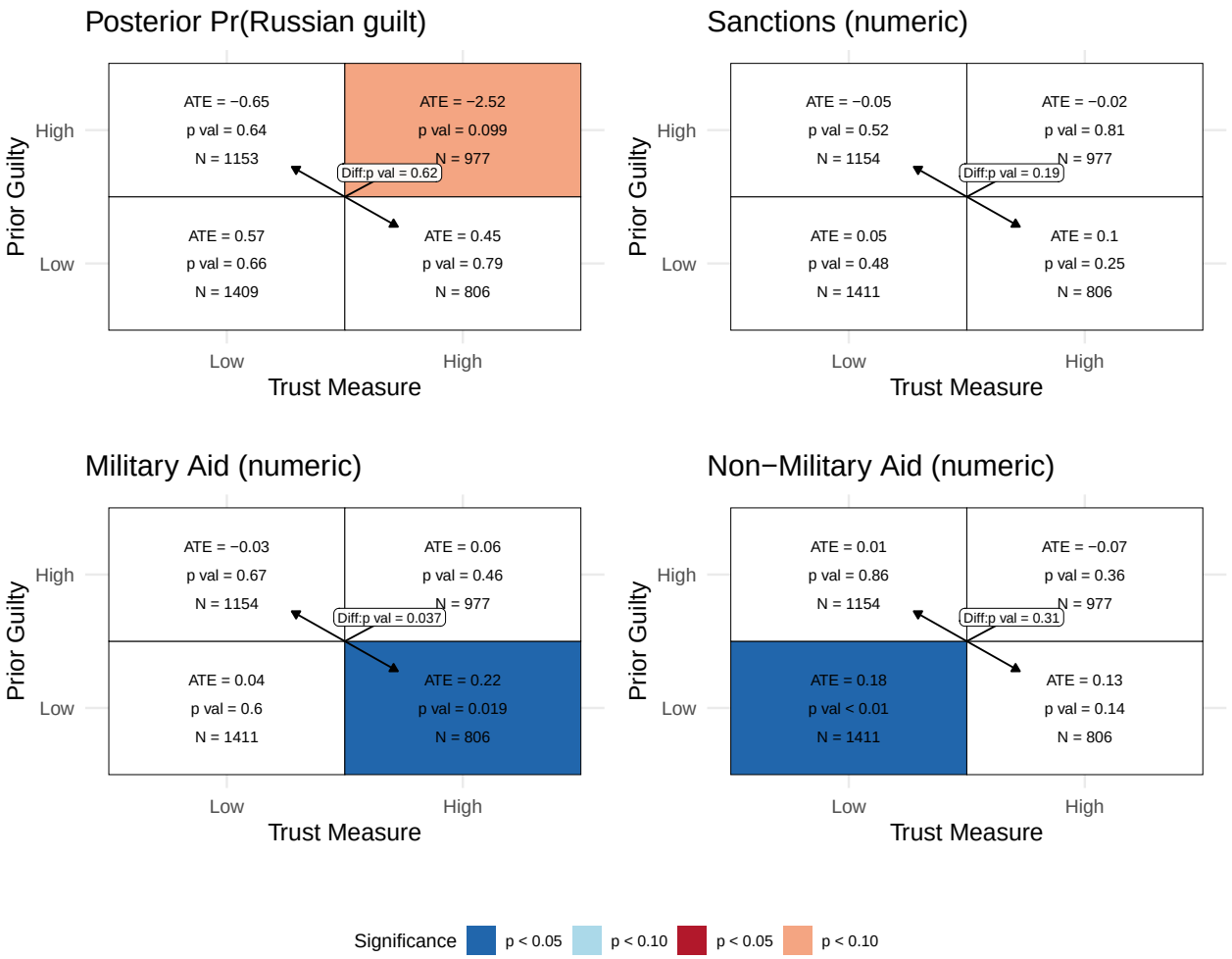


Figure D.3: Effect of ICC treatment, cooperative internationalism boxes.