The Domestic Sources of International Trust

Michael A. Goldfien[*] Michael F. Joseph[†] Roseanne W. McManus [‡]

April 17, 2025

Abstract

How do states overcome mistrust? Scholars argue that costly foreign policy signals can build trust. But when trust is low, such as during rivalries, states are unwilling to use these signals for fear of being cheated. We argue that domestic policies can also build trust by revealing information about a state's likelihood of cooperating internationally when there is a correlation between domestic and international preferences. We further argue that domestic policies have a distinct advantage: the value states accrue from them depends less on international reciprocation. As a result, domestic choices can reassure counterparts at moments when trust is so low that costly international signals appear prohibitively risky. We test the implications of our theory in a detailed case analysis of the Cold War's end and several shorter case studies, illuminating two situations the current literature struggles to explain: initial trust-building between enduring rivals and autocracies' reassurance of liberal states via illiberal domestic policies.

Keywords: Trust; Reassurance; Rivalry; Major Power-Client Relations; Domestic Politics and IR

[*] Assistant Professor, Department of National Security Affairs, U.S. Naval War College (michael.goldfien@usnwc.edu).

[†] Assistant Professor, Department of Political Science, UC San Diego (mfjoseph@ucsd.edu)

[‡] Associate Professor, Department of Political Science, Pennsylvania State University (rum842@psu.edu). The order of the authors is alphabetical and the contributions are equal.

Prominent research on the security dilemma has argued that states can overcome it using costly signals (e.g., Kydd 2000; Glaser 2010; Jervis 1978; Yoder 2019b; Haynes and Yoder 2020; Yoder and Haynes 2025). According to this literature, states that genuinely desire international cooperation reveal their security-seeking intentions by pursuing *foreign* policies that greedy states would not, such as reducing arms, joining international agreements, and pursuing retrenchment. But history is also replete with cases in which a government's *domestic* policy choices engendered trust with foreign counterparts. For example, in 1958, Gamel Abdel Nasser "unleashed a crackdown" on Egyptian Communists, which encouraged U.S. officials to "mend fences with the man whom they just months earlier compared to Hilter" (Radchenko 2024, 228). In 1961, a military coup and subsequent top-down economic reform in South Korea reinvigorated Washington's faith in its northeast Asian ally (Brazinsky 2009). In the late 1970s, market reforms implemented by Deng Xiaoping accelerated the rapprochement between Washington and Beijing begun earlier by Richard Nixon and Mao Zedong (Zagoria 1984). From 1985 to 1987, Mikhail Gorbachev released dissidents from internal exile, loosened emmigration restrictions for persecuted Soviet Jews, and introduced *glasnost* and *perestroika*, all of which increased Western trust in Soviet intentions (Bartel 2022; Wilson 2014). In the 2000's, limited political liberalization in Cuba whet the Obama administration's appetite to normalize relations several years later (LeoGrande 2015, 473-488).

Most recently, after rebels who were formerly aligned with ISIS and al-Qaeda toppled the long-standing Syrian regime, President Biden's remarks indicated that domestic policies would be a key litmus test for cooperation. Biden stated, "It's now incumbent upon all the opposition groups who seek a role in governing Syria to demonstrate their commitment to the rights of all Syrians, the rule of law, the protection of religious and ethnic minorities" (Biden 2024). Similarly, British Cabinet Minister Pat McFadden said that the United Kingdom would reconsider its legal designation of Syrian rebels as terrorists if they followed through on their leader's pledges "about the protection of minorities, about respecting people's rights" (Rhoden-Paul 2024).

These examples demonstrate that the standard costly signaling logic of trust-building could extend to domestic policies. But they also raise puzzles. First, the conventional wisdom in IR is that illiberal and authoritarian actors are mistrusted, especially by liberal states (Maoz and Russett 1993a; Russett and Oneal 2001; Tomz and Weeks 2013a). But in the examples listed above, there is evidence that *both* liberal and illiberal—even violent or coercive—actions reassured the United States. Second, the reassurance literature suggests that initial trust-building measures are undertaken with an eye toward influencing international opinion (Jervis 1978; Kydd 2005). Yet many ultimately reassuring domestic choices—such as Deng's economic reforms following the Cultural Revolution—appear driven primarily by domestic and not international concerns (Cable 2017). Finally, existing literature finds that trust-building processes between international rivals must start small, for example via modest arms control agreements, since even security-seekers fear being exploited by international rivals (Kydd 2005). Yet, as will be discussed in greater detail below, Gorbachev initiated a trust-building process with the West via domestic choices that, in the Soviet context, were nothing short of revolutionary.

To our knowledge, the possibility that domestic choices could serve as trust-building signals has previously been mentioned only in passing (Kydd 1997). Yet the puzzles above suggest that domestic choices build trust in ways that depart from our understanding of trust-building through foreign policy. Thus, a systematic analysis of the link between domestic policies and international trust is needed. We take up this task. We theorize the unique features of domestic policies that create unappreciated opportunities to engender trust between international rivals. To begin, we construct a simple two-period model of international trust problems (similar to Kydd 2005; Yoder and Haynes 2025). We then advance two conceptual claims about how domestic policy choices resemble and differ from international signals, which we introduce into the model.

Our first conceptual claim is that states hold overarching values that contribute to both their domestic and international policy preferences. Therefore, knowing which domestic policies a state prefers can shed light on whether that state will reciprocate or exploit international cooperation. Of course, the relationship between domestic and international motivations is imperfect. This limits the kinds of domestic choices that can communicate foreign policy motives to those where a security-seeking state is likely to prefer one domestic policy and a greedy state is likely to prefer another. Nonetheless, in cases where the correlation between domestic and international preferences is sufficiently strong, standard costly signaling arguments about international trust can also apply to domestic choices. For example, during the Cold War, the United States viewed socialist states as less trustworthy. A socialist state could, in principle, engender false trust with the US by instituting right-wing domestic policies. However, just as a greedy state receives a lower payoff from cutting arms in standard trust-building accounts, a socialist state receives a lower payoff from right-wing domestic policies. Therefore, as on the international level, it is costly for one type of state to imitate another on the domestic level. This allows domestic policies to function as a costly trust-building signal on the international stage when domestic and international preferences are sufficiently correlated.

Our second conceptual claim is that domestic policy choices depart from international policy choices in a critical way: they tend to be more independent. We use the term *independence* to describe the extent to which the payoffs from a state's policy choice depend upon an international counterpart's choice. In the classic security dilemma logic used to study costly international signaling, scholars assume that choices are highly dependent; that is, one state's payoff greatly depends on what the other does (Jervis 1978; Kydd 2005). We argue that domestic policy choices tend to be more (but not necessarily entirely) independent, meaning that the benefits and costs that a state derives from domestic choices depends less on what their counterpart does. For example, how much China benefits from cutting arms depends to a great extent on whether the U.S. responds in kind. By contrast, China's benefit from releasing or incarcerating political prisoners has much less to do with what the U.S. government does. The relative independence of domestic choices allows them to spark trust-building between states that seek cooperation but to whom international signals appear too risky. At the international level, even security-seeking states may hesitate to send the first international signal of goodwill for fear that rivals will exploit their cooperative gesture (Montgomery 2006; Rosato 2014). By contrast, the independence of domestic choices means that states can reveal their security-seeking intentions without exposing themselves to unacceptable risk.

These insights help to address empirical puzzles both highlighted by the anecdotes listed at the outset of the article and found in surprising cases of trust-building which the existing literature struggles to explain. First, our theory explains trust-building following illiberal domestic choices. While liberal policies can establish trust in some cases, our theory indicates that compatibility of preferences is the key determinant of cooperation in world affairs. Thus, there are some situations—e.g., harsh anti-Communist repression—in which illiberal domestic choices could reassure even liberal states. We demonstrate this logic below with several case anecdotes of trust-building between the U.S. and its client states during the Cold War, in which Washington viewed autocratic regimes' violent suppression of domestic Communist groups as evidence of alignment with Washington's international agenda.

Second, our theory indicates that domestic choices can play a uniquely important role in trust-building between enduring rivals. Because years of bitter rivalry has eroded all trust, rivals are especially unlikely to find international signals that are salient (i.e., costly) enough to reassure one another but not so salient as to create unacceptable risk. The independence of domestic choices provides a powerful solution to this problem, allowing security-seekers to undertake highly salient and reassuring domestic choices without fear of international exploitation. These domestic choices are not necessarily taken with the primary goal of signaling to international rivals, but they can nonetheless engender trust. We illustrate this logic with a detailed case study of the end of the Cold War. The case shows that Soviet domestic choices early in Gorbachev's tenure paved the way for later cooperation with the West on international security issues. Soviet choices amounted to "grand gestures" of the sort that are difficult to explain under existing rationalist accounts of trust building, and were as or more important in reassuring the West than anything Moscow did on the international stage.

Our argument has several implications. First, we offer a general theory of domestic politics and international trust. The extensive literature on trust and reassurance has paid scant attention to domestic politics. We rectify this and in doing so highlight a unique feature of domestic choices—their independence—that makes them especially useful in kick-starting a trust-building process between international rivals. Prior research has linked domestic politics to signals of resolve (Schultz

1999; Weeks 2008; Renshon, Yarhi-Milo, and Kertzer 2023; Fordham 1998; McManus 2017). We show that domestic politics may be equally important for reassurance.

Given that political elites spend the vast majority of their time on domestic politics (Lindsey and Hobbs 2015), our research exposes many avenues for trust-building that past scholarship may have overlooked. This also forges a stronger connection between international and comparative politics, by illuminating unappreciated international implications of domestic choices such as land and tax reform (e.g., Flores-Macías 2019), purges (e.g., Bokobza, Krishnarajan, Nyrup, Sakstrup, and Aaskoven 2022), and discrimination against minorities and immigrants (e.g., Peters 2015). We provide a common framework for understanding why, whether, and under what conditions, these or any other domestic choice can facilitate international trust or mistrust.

Finally, we advance research on the democratic peace (e.g., Maoz and Russett 1993a,b; Russett and Oneal 2001). Building on existing research, we offer a novel explanation for the apparent ability of democracies to identify one another as sharing cooperative international preferences. Yet our approach goes beyond the democratic peace by explaining trust-building between autocracies or between mixed-regime pairs. It sheds light on puzzling cases in which illiberal actions engender trust, and also allows for the possibility that domestic reforms that fall well short of regime change can reassure international rivals. This holds a critical policy implication: it is possible that the U.S. and China can reestablish trust, even if autocracy persists in China.

1 Trust Problems and Foreign Policy Trust Building

Trust problems are central to international relations. They play a role in power transitions (Yoder 2019a), conventional and nuclear arms races (Debs and Monteiro 2014; Bas and Coe 2016), arms control negotiations (Coe and Vaynman 2020), great power rivalry and rapprochement (Glaser 2010; Mattes and Weeks 2022), and general problems of cooperation (Crescenzi 2018). Trust problems are also an important focus of policymakers. For example, as American policymakers manage relations with China, building trust may be as important as signaling resolve

(Glaser 2015).

Scholars studying the security dilemma consider two ideal-types of states: security-seeking and greedy.[1] The fundamental difference between them is the value they get from cheating each other. Security-seekers get the highest value from reciprocal cooperation. In contrast, greedy states prefer to defect no matter what their partner will do.[2] A trust problem arises because states are uncertain about each other's intentions; and the value security seekers accrue from cooperation depends on what the other does.

We visualize a formalization of this problem in Table 1. As it shows, the trust problem focuses on two states that will soon confront a critical foreign policy choice. This could be the choice to comply with an arms control agreement or not; or to withdraw troops from an area or not. In such situations, each state can chose to cooperate or defect. These choices are shown in the 2×2 matrices presented in each panel.

Table 1: Representation of the trust problem

		Security-Seeking (y_B)		Greedy ($1-y_B$)	
		coop	defect	coop	defect
Security-Seeking (y_A)	c	H,H	0, H-a	H,H	0, H+a
	d	H-a, 0	1,1	H-a, 0	1,1
		coop	defect	coop	defect
Greedy ($1-y_A$)	c	H,H	0, H-a	H,H	0, H+a
	d	H+a, 0	1,1	H+a, 0	1,1

Note: The y_i represent the probability state i is a security seeker. That information is private. Each panel represents payoffs given player's type, and whether states cooperate or defect. Top-left realizes Stag Hunt preferences, bottom right realizes Prisoners' Dilemma preferences. We make **two assumptions** to capture the order of preferences in the Stag Hunt and Prisoners' Dilemma substantively motivated by, and depicted in, Jervis (1978, 171): $H > 1$; $1 > H - a$.

Each panel in Table 1 represents one of the four combinations of preferences that a pair of states

[1] We take these specific terms from Glaser (2010). Others use mean/nice, revisionist/status quo, etc. Jervis (1978) associates security-seeking preferences with the Stag Hunt game and greedy preferences with the Prisoner's Dilemma game.

[2] The sources of greedy or security-seeking motives are beyond the scope of this study. For a recent overview of the possible sources of greedy motives for territorial expansion, see Altman and Lee (2022).

can hold. Inside each panel are four possible sets of payoffs, based on different combinations of actions.[3] H represents the payoff from cooperating and is assumed to be greater than 1, meaning that mutual cooperation is always preferable to mutual defection. The parameter a, which is greater than 0 but less than H, captures both the reduction in utility that security-seekers experience from defecting while the other cooperates and the increase in utility that greedy types experience from defecting while the other cooperates. States always receive the worst possible payoff of 0 (i.e., the sucker's payoff) if they cooperate while their counterpart defects. Not only does this put their national security at risk, but unreciprocated international cooperation can damage a leader's domestic standing (Colaresi 2004).

Because security-seekers prefer to reciprocate rather than exploit cooperation, we might intuit that two security-seekers could always cooperate. Indeed, this would be the case if two security-seekers knew each other's type. In reality, however, states are usually uncertain about each other's intentions, and this can make cooperation more difficult. In the trust game, we assume that Player A knows her own type, but she only knows that B is a security-seeker with probability y_B. She believes that B is greedy with probability $1 - y_B$. Similarly, B believes that A is a security-seeker with probability y_A.

The most basic questions trust scholars ask are: Given that states are uncertain about each other's motives, when are security-seekers willing to cooperate with each other, and when does uncertainty cause them to defect? Scholars show that we can support mutual cooperation if both states are sufficiently confident that the other holds security-seeking preferences. However, if even one state is sufficiently uncertain about the other's motives, mutual defection is the unique solution in a one-period interaction (Kydd 2005; Yoder 2019b).

Uncertainty about a counterpart's preferences can cause cooperation between security-seekers to fail through two mechanisms. First, A may defect because A does not trust B enough (i.e., believes that B is likely greedy). Second, even if A is a security-seeker and trusts B, B may not

[3] Some scholars—e.g., Kydd (2005)—the payoffs inside these boxes using parameters that equate to the bargaining model of war (Fearon 1995). Their theoretical findings follow because security-seekers and greedy types vary in their preference ordering over different combinations of choices. We chose the smallest number of payoff parameters to re-create this basic problem.

trust A. In this situation, A also defects. The reason is that A expects B to defect due to B's distrust. Because B's strategy is a foregone conclusion, A also defects. This second mechanism illustrates the difficulties of two-sided trust problems. Both sides must not only trust the other, but also believe that the other trusts them.

The central insight of the trust-building literature is that distrustful states can build trust through costly signals. Many scholars model this as two sequential foreign policy choices (Kydd 2005; Yoder 2019b). The second period choice is often conceptualized as a broad or high-stakes choice between cooperation and competition. The initial choice is conceptualized as a signaling opportunity. Through their initial choice, security-seekers can potentially reveal their type by taking actions that have higher payoffs for them than for greedy types (Kydd 2005; Glaser 2010). Consider the example of arms reductions. Cutting arms is very costly for greedy states, since it undermines their ability to pursue expansionist foreign policies. It is less costly for security-seekers, who have more modest ambitions. Therefore, a state's willingness to cut arms can credibly signal its security-seeking motives. A similar logic applies to other costly trust-building signals that scholars have identified, including signing arms control treaties (Kydd 2005), visits and other symbolic gestures (Berenji 2020), building purely defensive weapons (Glaser and Kaufmann 1998), and retrenchment (Yoder 2019b).

It is important to note that the logic of using costly signals to communicate security-seeking intentions in trust-building settings is different from the logic of signaling resolve to fight in crisis bargaining (Kertzer 2016; Dafoe, Renshon, and Huth 2014). In crisis bargaining, states face a choice of whether to send a costly signal of resolve—for example, whether to build more arms. More highly resolved states are less sensitive to the costs of arming, and therefore arming can signal resolve. However, all types of states are assumed to find arming costly to some degree. Thus, both resolute and irresolute types would prefer to arm less, all else equal. In contrast, greedy and security-seeking states that are faced with the choice to cooperate or defect in a trust-building setting hold *opposite* preferences over outcomes: Security seekers prefer mutual cooperation over cheating, while greedy states prefer cheating over mutual cooperation. Therefore, consistent with past trust theories, we can conceptualize costly signaling in trust-building settings as the "opportunity cost" that is paid by security seekers for cheating and by greedy types for cooperating. Cooperative signals thus become credible indicators of security-seeking intentions because greedy types find it more costly to send them.

We treat extant research on trust-building through foreign policy signals as our baseline. In Appendix A.1, we set up a two-period model of trust that first privately draws each state's intentions (using probabilities y_A, y_B), then iterates the simultaneous-move model visualized in Figure 1 over two periods. Three insights from the philosophy of science motivate us to use a simultaneous moves model. First, doing so harnesses the power of incrementalism (Ashworth, Berry, and de Mesquita 2021, 58-59). By adding a single parameter (payoff independence, introduced in the next section) to the framework used by most scholars of trust-building (e.g, Jervis 1978; Kydd 2000, 2005; Shultz 2005; Yoder and Haynes 2025; Acharya and Ramsay 2013),[4] we assure the greatest hope of accumulating knowledge transparently across studies (see Ramsay 2017, for review). Second, this setup is substantively motivated to capture the strategic problem we hope to study: Achieving cooperation given initial *mutual* distrust. In the cases that interest us, such as enduring rivalries, arms control, and major power-client relations, both sides have the ability to exploit the other by adjusting their policy before the other realizes and can take countermeasures (e.g., Kydd 2005; Yoder and Haynes 2021; Powell 1996; Vaynman 2022; Glaser 2010; Waltz 1979; Braumoeller 2008). Thus, at every moment, both worry that the other may be cheating them, and this creates strategic pressure to cheat first. If we instead chose to sequence choices, it would artificially resolve the trust problem for whoever moves second.[5] Third, it is necessary to assume mutual distrust as a precondition in order to eventually explore how the independence of domestic choices can alleviate this problem.

We isolate a theoretical baseline in Appendix A.2 that reflects previous trust-building argu-

[4] While Powell (1996) studies continuous time, it too is a simultaneous moves model.

[5] We accept there are some substantive areas where exploitation can only be one-sided. For example, in proliferation cases, only the nuclear aspirant can secretly cheat because the great power already has nuclear weapons and has a known preference to prevent proliferation (Debs and Monteiro 2014). But in our substantive domain, any state can cheat their rival at any moment.

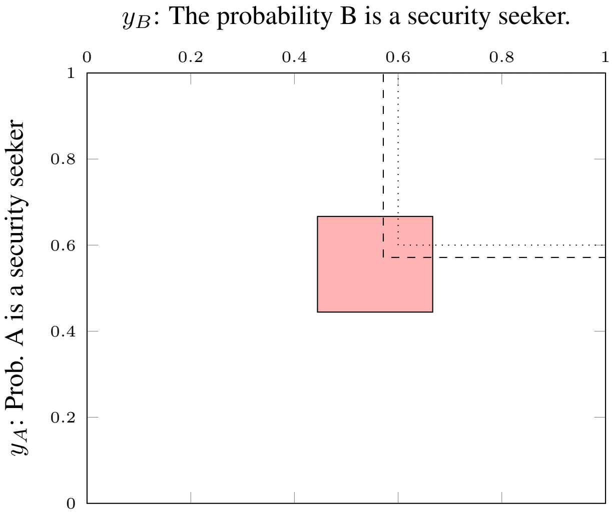
ments, where we initially assume that the direct costs and benefits of both periods are the same because both choices represent foreign policy choices and are thus similar in structure.[6] Thus, in later sections we can identify the effects of domestic signaling opportunities by manipulating features of the first period choice based on our substantive arguments about why domestic policies are different.

Given we start with the same baseline structure as others (e.g, Kydd 2005), our baseline results are consistent with past scholarship on trusting building through international costly signaling. Figure 1 plots the pure strategy equilibria of the baseline game as a function of each player's initial trust levels. Each box represents the conditions under which we can support a specific equilibrium. When initial trust is very high (top-right corner) two equilibria emerge (dotted and dashed boxes) where security-seekers trust their counterparts enough and cooperate in the second period. The equilibrium that interests us the most is the trust-building equilibrium (red). In it, security-seekers always cooperate in the first period, but greedy states do not. Then security-seekers cooperate in the second period only if their counterpart cooperated in the first. We call it a trust-building equilibrium because players learn about each other's motives from their first-period actions. When they both cooperate, they come to trust each other and this allows for cooperation in the second period.

We verify that even with the option to send a costly signal of reassurance in the first period, trust-building between security-seekers is not always possible. When at least one state initially believes the other is greedy with high probability, the states defect in both periods. This finding explains why trust-building is difficult for enduring rivals. Over time, rivals may genuinely come to desire peace as a result of exhaustion, shifting values, or leadership change. But they may be too fearful to take the first step towards trust-building because they are mistrustful of their rival and fear that cooperative gestures will be exploited. This motivates our question: when states start out deeply mistrustful of each other, how can they initiate trust-building?

[6] Some game theorists assume that the relative salience of each issue varies, because they argue that rivals can always calibrate the relative value of the two periods so that trust-building is possible (Kydd 2005). But these arguments face empirical, substantive, and technical criticisms (Lieber 2011; Montgomery 2006; Rosato 2014). In section 2.2, we will return to these arguments in detail.

Figure 1: Pure Strategy Perfect Bayesian Equilibria in the Standard Trust-Building Game



Description	Strategies	
	Greedy	Security-Seeking
Trust-building (Shaded red)	d, d	c, then d if other d; or c if other c.
No learning (dotted)	d,d	d,c
Suckers (dashed)	c,d	c,c
Semi-tragic (dotted)	d,d	c, d
Tragic (all parameters)	d,d	d,d

Assumes $H = 1.5, a = .75$. The table describes the five pure strategy equilibria of the trust-building game. c, d represents cooperate in the first period, defect in the second. The plot represents the parameter ranges. We solve for these Pure Strategy Equilibria in Appendix A.2. We examine mixed strategies, and justify our focus on pure strategies in Appendix A.2.1.

2 Domestic Policies and International Signaling

We argue that when two security-seekers distrust each other, they can kick-start the trust-building process through domestic policy choices. These domestic choices operate like the costly foreign policy signals studied in the existing literature in that security-seekers and greedy states derive the highest payoff from different policies.

When we say domestic choices, we do not primarily mean broad attributes of a state, such as regime type (Russett and Oneal 2001). Regime type could be viewed as a large-scale domestic policy choice, but it is not our primary focus because regimes rarely change. We also do not mean leader turnover (Wolford 2007), which is insufficient in itself to alleviate the security dilemma. Previously distrustful foreign counterparts will not immediately trust new leaders because they often come from the same pool of elites and because reputations adhere to both states and leaders (Renshon, Dafoe, and Huth 2018; Goldfien, Joseph, and McManus 2023). Of course, some new leaders do have sharply different preferences from their predecessors, but we expect this to become observable through their policy choices.

Instead, the domestic choices that we focus on involve government decision-makers who are confronted with a problem and can select among different policy options to resolve it.[7] One example might be a situation where the leadership is faced with nationwide protests and must decide between repression or negotiation with demonstrators.[8] As another example, a government facing an economic crisis could choose to address it by nationalizing industries or allowing its currency to float. One important feature of these choices is that they have a structure resembling the international choices that states face in the signaling stage of the trust-building game. Like typically-studied international choices, these domestic choices present states with at least two different policy options, and a state's overarching values determine its preferences over the options.[9]

[7] Some policy choices have costs and benefits at both the domestic and international levels. As we explain below, it is not the precise balance of domestic and international interests that matters for our theory, but rather the extent to which the payoffs from the choice are independent of what an international adversary does.

[8] For example, the Cuban government faced such protests in 2021 and cracked down.

[9] To be clear, leaders have more than two possible responses to any given domestic problem, but that is also true of international problems. Like the international trust-building literature, we emphasize two choices that might be seen as the options that would be favorable to the security-seeking and greedy types.

As on the international level, if a state chooses a domestic policy that does not align with its true preferences, it is less satisfied with the outcome. Therefore, opportunity costs give states an incentive to act true to their preferences at both the domestic and international levels. However, for domestic policies to communicate information about foreign policy motives, there must be a correlation between domestic and international preferences.

We argue that domestic policies can shed light on international motives because states hold overarching values that contribute to both their domestic and international policy preferences. Therefore, states with certain kinds of domestic policy preferences are likely to have certain kinds of foreign policy preferences. Our argument builds on work by Goldfien et al. (2023), who show that decision-makers have underlying dispositional attributes that influence their sensitivity to both the costs of fighting in a crisis and the costs associated with certain domestic choices. We extend this logic to account for correlation in preferences across the domestic and international policy spaces in trust-building scenarios. We argue that the preference to exploit international cooperation is likely to be correlated with certain domestic policy preferences. For example, governments that care little about legal restraints may be more likely to violate both domestic law and international arms control agreements.

One version of our claim overlaps with the liberal peace. Scholars have argued that domestic policies such as free elections and respect for ethnic minorities and human rights can predict security-seeking international intentions (Kydd 1997; Tomz and Weeks 2020; Maoz and Russett 1993a; Tomz and Weeks 2013b). In line with these previous claims, our first case study will illustrate how the U.S. interpreted Soviet liberalizing domestic reforms as evidence of greater international trustworthiness. But our claim goes further. We argue that what really matters is compatible international preferences, not liberal actions per se.[10] Thus, which policies promote trust depends on the context. In certain contexts, domestic actions that are illiberal, but are likely to be correlated with a preference to cooperate with a particular international partner, can induce trust. For

[10] For research highlighting the importance of compatible preferences for international cooperation, see Voeten (2021), Gartzke (1998), and Spaniel and Smith (2015). These arguments identify compatibility as important, but do not examine trust problems or domestic signaling opportunities.

example, the decision of right-wing Cold War dictators to brutally suppress domestic Communist groups was illiberal. Still, as we show later, U.S. officials inferred that the preferences revealed by these actions were correlated with a preference to cooperate with the U.S. internationally and increased their trust in these regimes.[11]

Since the relationship between domestic preferences and willingness to cooperate internationally is context dependent, we cannot identify a particular set of domestic choices that will always promote trust with every other regime. Democracies, Communist regimes, right-wing dictatorships, theocratic regimes, Western states, and non-Western states might all interpret different types of domestic choices as evidence of willingness to reciprocate their own cooperative gestures. They might even interpret the same actions differently. For example, analysts from the European Union are likely to interpret the Turkish government's push for a greater role of Islam in public life as evidence of lower willingness to cooperate, but analysts in Islamist states may have a different reaction. Yet even if they do not draw the same conclusions, we argue that all types of states can potentially learn something from each other's domestic policy choices.

To be clear, we do *not* argue that security-seeking and greedy states hold reliably different preferences over all or even most domestic policy choices, such as where to set the speed limit. Thus, most domestic choices have no bearing on world politics. Further, even when a domestic choice is correlated with international preferences, the degree of correlation is usually imperfect. The international choices that trust-building models have historically focused on, such as cutting arms, directly impact a state's ability to fight a war and thus plausibly have a strong correlation with greedy or security-seeking preferences.[12] Since domestic choices rarely have such a direct effect on warfighting capabilities, their correlation with greedy or security-seeking preferences can be more variable. For example, allowing freedom of expression is consistent with liberal values and thus likely correlated with willingness to cooperate with the United States. However, the

[11] Of course, domestic suppression and brutality may be correlated with greater international resolve (see, e.g., Goldfien et al. (2023)), but high resolve is not incompatible with cooperation among states with similar preferences. Moreover, brutal domestic actions typically do not position a state for easier international expansion in the way that hawkish international actions often do.

[12] Yet even at the international level, this correlation is likely to have some variation. This is an issue that existing trust-building models have failed to account for (Kydd 2005).

correlation is not perfect because even some U.S. allies suppress free speech in order to prevent hate speech, social unrest, or dissent.

How strong does the correlation need to be for domestic choices to signal international motives? Because this is partly a strategic problem, we explore the level of correlation necessary using our formal model.[13] We find that perfect correlation between domestic and international preferences is not necessary for learning from domestic choices to occur. Indeed, domestic choices can shed some light on international preferences with only a moderate correlation (defined precisely in Appendix A.5). However, the amount of information communicated is greater when the correlation is stronger and lesser when the correlation is weaker.

This raises the question: If domestic and international preferences are imperfectly correlated, why would domestic choices play an important role in international trust-building? Why are their effects not overshadowed by international signals of security-seeking motives? This is the question to which we now turn.

2.1 The Domestic Advantage: Payoff Independence

We argue that many domestic policy choices have an unexplored advantage that enhances their capacity to forge trust: their level of payoff independence. In brief, independence refers to the extent to which a state's payoffs (that is, the combination of benefits and costs accrued) from a choice depend upon what a foreign counterpart does.[14] The concept of independence becomes relevant whenever a state faces a domestic or international policy problem and can choose among different policy options to address it. A state's payoff from each option always depends on the state's own preferences. Yet sometimes the payoff from one or both options also depends, at least in part, on what option a foreign counterpart chooses.

We conceptualize the degree to which choices have payoffs that are independent from or dependent on a foreign counterpart's actions as a continuum. At one extreme, a state's choice is

[13] See Appendix A.5 for a formal discussion.

[14] The concept of independence is relevant to both international and domestic choices, although we argue below that domestic choices are often more independent.

maximally independent if the value it gets from selecting either policy option is the same no matter what its foreign counterpart does. For example, suppose State A was faced with nationwide protests (the policy problem) and considering whether to repress (option 1) or negotiate with (option 2) protesters. If State A's value from either repressing or negotiating does not depend on anything State B does, then State A's decision about how to respond to protests is fully payoff independent. This plausibly fits some empirical contexts, but not all. If instead State A's payoff from this decision depends somewhat on State B's response (e.g., whether State B lodges diplomatic criticism or even imposes sanctions), then the choice would only be partially independent. Put another way, if a state can calculate its own payoffs from each policy option without considering what any other state will do, then the choice it faces is fully independent. On the other hand, if the payoffs depend slightly, partially, or highly on what another state does, then the decision is slightly, partially, or highly dependent.

Existing research into the security dilemma assumes choices are highly dependent (Kydd 2005). For example, in the arms control variant, the immediate payoff A gets from the decision to reduce arms (cooperate) depends on whether B also reduces arms (cooperates) or instead continues to arm (defects). Similarly, the benefit A gets from the decision to arm (defect) depends on what B does. Indeed, we think this assumption is reasonable because the value states accrue from many international policy options that past trust scholars have studied, such as retrenchment or arms control, are heavily determined by what rivals do.

In contrast, domestic choices can have much greater variation in their level of independence. Examples of domestic choices that could plausibly be fully independent include reforming social welfare, reducing domestic regulations, or changing land use policy. The costs and benefits associated with these choices have little or nothing to do with the choices that foreign countries make. On the other hand, choices about immigration policy and press freedom could be moderately independent. Easing immigration standards could result in different levels of immigration depending on other states' policy choices. The costs and benefits of permitting press freedom could depend somewhat on whether foreign counterparts engage in influence campaigns (Levin 2021).[15] Overall, though, few domestic choices are as dependent as the foreign policy choices prominently examined in the trust literature. Table 2 summarizes how various domestic and international choices can vary in independence.

Table 2: Examples of Choices with Varying Degrees of Average Independence

Choice	Independence	Explanation
Low or high speed limit	High	Costs and benefits of either option do not depend on other countries' choices.
More or less social spending	High	Costs and benefits of either option do not depend on other countries' choices.
Permit press freedom or not	Moderate	Costs and benefits of both options are primarily domestic, but the costs of free media are higher if another country engages in influence operations.
Repress protesters or not	Moderate	Costs and benefits of both options are primarily domestic, but criticism or sanctions from another state could increase the costs of repression.
Instituting tariffs or allowing free trade	Low	The effect of either option on exporting firms depends greatly on whether other countries reciprocate.
Cutting or building arms	Low	Cutting arms puts a state in a very vulnerable position if a foreign rival does not reciprocate. Building arms leads to a military advantage only if the other side does not do the same.

We will use our formal model to show that more independent choices, which tend to be domestic, have an important advantage when it comes to trust-building. As noted earlier, the primary barrier to trust-building with international signals is the fear of being cheated. However, this fear assumes that one state's value for taking a trust-building action hinges on what the other state

[15] This last example highlights that the options States A and B are choosing between need not be identical in order for them to be dependent on each other. All that is necessary for A's choice to be dependent is that it depends on *some* choice by B, not necessarily the same choice.

does (i.e., full dependence). If the payoff from taking an action is independent of what the other does, this fear vanishes, and security-seekers face little risk when they make choices that reflect their genuine cooperative preferences.[16] At the same time, greedy states are more easily identified because they can no longer rely on the excuse, "I am not cooperating because I am afraid that you will cheat me." Independence thus gives security-seekers the necessary confidence to make an initial choice that will build trust and makes it easier for both sides to learn about each other's preferences.[17] Therefore, domestic choices that are both independent and correlated with security-seeking or greedy international preferences can play a crucial role in international trust-building.

It is important to emphasize that independence *does not* mean that policy choices are costless. It merely means that the benefits and costs are not contingent on a counterpart's behavior. In the absence of dependence on the counterpart's behavior, the payoffs of choices will depend even more heavily on a state's own preferences. This means that it would entail an even larger opportunity cost for a state to play against type. Therefore, the payoff structure promotes behavior that reveals more information about a state's type in the presence of independence.

2.1.1 The Strategic Implications of Independence for Trust-Building

To explore the strategic implications of independence, we introduce the concept into the two-period trust game described earlier. We conceptualize the level of independence of each state's choices using the continuous parameters β_A, $\beta_B \in [0, 1]$. We apply these parameters *only* to the first period of the model. We omit the independence parameters from the second period because they are effectively equal to 0—that is, the choices are fully dependent. We do this in order to make the second period identical to the classic trust problem.[18]

We visualize A's first-period utility with independence in Table 3.[19] When β_A and β_B both

[16] The fact that these choices reflect true preferences arguably allows them to function as "indices," sources of information that are credible because they are believed to be "inextricably linked to the actor's capabilities and intentions" (Jervis 1989, 18).

[17] Although we argue that domestic choices are more likely to be independent, this argument could also apply to any international choices that have high independence and are correlated with future preferences to cooperate.

[18] This creates a hard test for the ability for first-period actions to build trust. If second-period actions could also be partially independent, the trust problem would be less severe.

[19] The effect on B's utility is symmetrical.

equal 0, the model converges to the classic model of trust-building through international actions presented earlier (see Table 1 and Figure 1). In this classic model, A's value for cooperating or defecting depends on a combination of A's type and B's choice. In contrast, when β_A equals 1, it means that B's choice in the first period has no impact on A's value from cooperation or defection in the first period.[20] Rather, A's first-period value from cooperation depends entirely on A's preferences. When β_A and β_B are between these values, it means both states' trust-building payoffs are partially dependent.

Table 3: Player A's first-period payoffs in the game with independence

A is Security-Seeking

	coop	defect
c	H	$\beta_A H$
d	H-a	$\beta_A(H-a)+1-\beta_A$

A is Greedy

	coop	defect
c	H	$\beta_A H$
d	H+a	$\beta_A(H+a)+1-\beta_A$

Note: These updated payoffs only apply to the *first period*. The second-period payoffs are still the same as in Table 1. B's payoffs are symmetrical.

We solve for all the Perfect Bayesian Equilibria (PBE) of our model in Appendix A.3 under the assumption that the level of independence can vary.[21] We plot the pure strategy PBE in Figure 2, assuming intermediate levels of independence ($\beta_A = .7, \beta_B = .6$).[22] Figure 2 also holds the H and a values at the same levels as in Figure 1, allowing for a direct comparison. We also use the same coloring scheme when the equilibria overlap in both models. Our goal is to establish that trust-building (the equilibrium shaded red) can occur through domestic choices when initial trust is low.

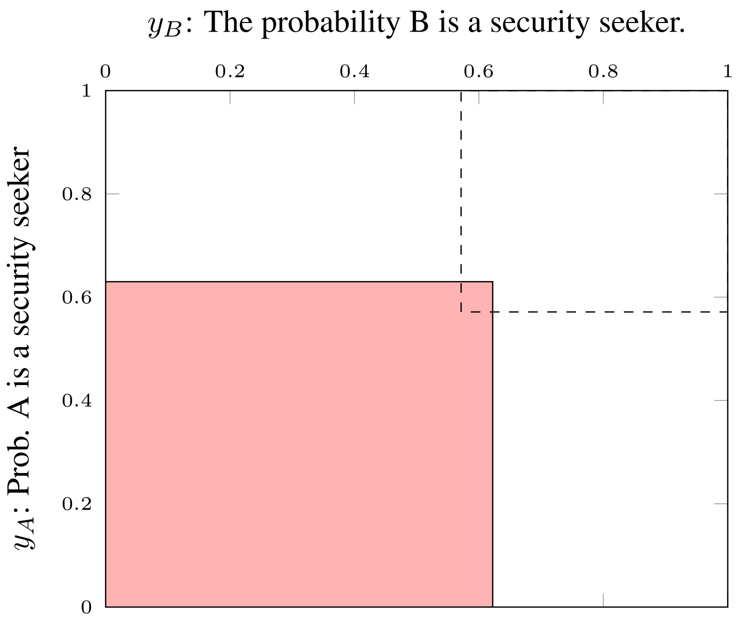
Our primary focus is on the conditions under which we can support the trust-building equilibrium in the model with domestic choices, relative to the standard model of international costly

[20] We continue to use the "cooperation" terminology for consistency, but with full independence one state is not really cooperating with the other. Rather, it is making an independent choice that is correlated with willingness to engage in future cooperation.

[21] The analysis partly relies on preliminaries reported in A.1.2.

[22] See Appendix A.3.5 for mixed strategy PBE, and Appendix A.2.1 for an explanation of why omitting them in the manuscript does not alter our conclusions.

Figure 2: Pure Strategy Perfect Bayesian Equilibria in the Domestic Signaling (Independence) Game



Description	Difference from International Signaling Model Caused By Independence (β)
Trust-building (Shaded red)	As independence increases, the equilibrium expands into low levels of trust. Once $\beta > \frac{1}{1+a}$, trust-building is possible even at the lowest initial trust levels.
No learning	When independence is sufficiently high ($\beta > \frac{1}{1+a}$), we can no longer support this equilibrium.
Tragic	When independence is sufficiently high ($\beta > \frac{1}{1+a}$), we can no longer support this equilibrium.
Suckers (dashed)	No change
Semi-Tragic	When independence is sufficiently high ($\beta > \frac{1-y(1+a)}{(1+a)(1-y)}$), we can always support this equilibrium.
Asymmetric Trust (not shown)	Independence allows for a lopsided equilibrium. In period 1, A defects and security-seeking types of B agree to be cheated, facilitating period 2 cooperation. It requires B to have high trust in A, while A has low trust in B, and moderate independence. There is a symmetric equilibrium in which A agrees to be cheated.

Assumes $\beta_A = .7, \beta_B = .6, H = 1.5, a = .75$. This implies $\beta_A \sim \beta_B > \frac{1}{1+a}$. Parameters and figure shading are chosen for direct comparison to the international signaling results in Figure 1. Recall, Figure 1 specifies each equilibrium strategy profile.

signaling. Proposition A.6 formally characterizes the trust-building equilibrium in the model with payoff independence. We report it in the Appendix. Here we describe the empirical implications by contrasting when trust-building occurs here, relative to the baseline condition. Thus, the results described below follow from a comparative static analysis of the trust-building equilibrium, which is generally presented in section A.3.2.

Recall that in the model where states were forced to signal using dependent international actions, the trust-building equilibrium did not survive under low initial trust. By contrast, our model of domestic signaling verifies the following.

Result 1a: When both players' choices are sufficiently independent (i.e., $\beta_A, \beta_B > \frac{1}{1+a}$), a trust-building equilibrium always exists for states that start out with the highest possible level of confidence that the other is greedy (i.e., $y \to 0$).

States are willing to build trust with domestic policies even at the lowest possible levels of initial trust because independence resolves both of the security-seeker's strategic problems. The direct strategic problem was that one state did not trust the other. This problem is eliminated because the value of one's choices no longer depends on what the counterpart does. With sufficient independence, a security-seeking state would still want to pursue the choice associated with cooperation even if its counterpart defected. The indirect problem arose because one state knew that their counterpart viewed them as untrustworthy. Therefore, they knew that their counterpart would defect as a matter of defense even if they chose to cooperate. But when their value from cooperation is independent, the state does not worry about this either. Since the trust-building equilibrium is perfectly separating, the states' choices are entirely informative and initial beliefs are overshadowed by the first-period choices.

This result exposes an *independence threshold* ($\beta_A, \beta_B > \frac{1}{1+a}$) for the policy choices that states face. Of course, it can be challenging to map real-life policy choices onto this threshold. For example, lifting immigration restrictions is mainly a domestic political choice, but a counterpart's policies could influence how much a state profits (or loses) from immigration reform. One might wonder: Does it matter if we cannot confidently code which choices meet the independence threshold for trust-building?

Result 1b: Even when the independence threshold characterized in 1a is not met, as the level of independence increases, a trust-building equilibrium can be supported at decreasing levels of initial trust.

Figure 3 plots the parameter ranges where we can support the trust-building equilibrium as a function of a state's initial belief that the other is greedy, and the level of independence of the first-period choice. When a choice has maximum dependence ($\beta_A = 0$), we can only support trust-building at intermediate levels of trust. Then, as independence increases, the range within which we can support trust-building expands downward until it is possible to support trust-building at the lowest possible levels of initial trust. Meanwhile, the maximum level of trust at which the greedy type will not imitate a security-seeker in the first period hardly changes. The reason is that independence only influences first-period payoffs, but the greedy type's value from cooperation in the first period is primarily derived from advantages it will later accrue by cheating in the second period.

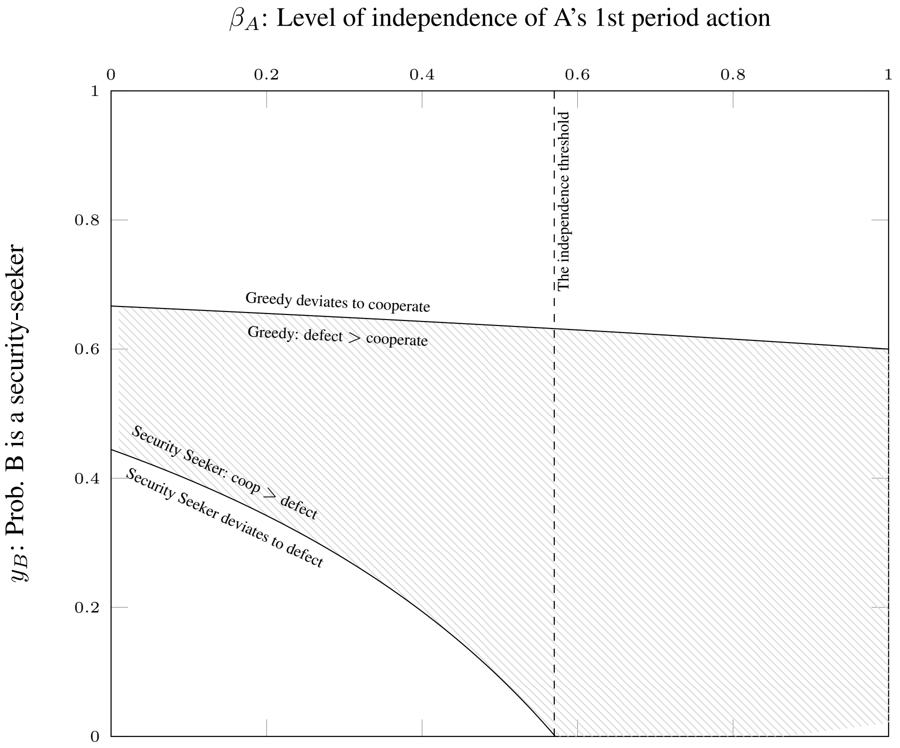
2.2 Salience of the Trust-Building Exercise

So far, we have assumed that the first-period domestic choice and second-period international choice are equally important. In reality, domestic policy choices vary in salience relative to international cooperation.[23] Variation in the salience of different international policies is an important theoretical focus of existing trust-building theories. In classic models, the salience of the first-period international choice must be neither too low nor too high relative to the second-period choice. Kydd argues that for any given level of initial trust, there is a first-period choice with exactly the right salience to signal reassurance without exposing the sender to too much risk (Kydd 2005). Therefore, Kydd concludes that security-seekers, in principle, can always find a right-sized signal to build trust when starting from any initial level of trust.

In practice, however, it may be difficult to find international trust-building gestures that are so finely calibrated. Skeptics call this the "Goldilocks" problem (Lieber 2011; Rosato 2014; Mont-

[23] In the literature, what we refer to here as "salience" is sometimes referred to as "costliness" (Kydd 2005). As noted above, there is an opportunity cost to playing against type in the game. The more salient or important the first-period choice, the greater that cost becomes.

Figure 3: When Is Trust-Building Possible Given Variation in β?



In the shaded parameter ranges, we can support trust-building for that level of independence β_A and a level of initial trust y_B. The upper (lower) boundary is driven by the greedy (security-seeker) incentive to deviate from trust building. The plot is identical for β_B, y_A. Assumes $H = 1.5, a = .75$.

gomery 2006). Calibration problems can be exacerbated by problems with interpretation. For example, when another state is constantly adjusting its force posture, it may be hard to notice if it dismantles a few missiles and even harder to interpret the salience of such a gesture. Moreover, psychological biases may cause the sender and receiver to disagree on the salience of a gesture (Kim 2022). These interpretation problems are especially likely to impede trust-building when the range of acceptable signals is small, a condition which is likely to prevail in enduring rivalries. Kydd's model shows that the calibration window is smaller when initial trust is low. Yoder and Haynes (2021), for their part, find in a laboratory experiment based on Kydd's model that trust-building is particularly unlikely to succeed in low-trust conditions. This suggests that trust-building through international signals will be especially difficult when we most desperately want it to work: to end enduring rivalries.

Is our argument about trust-building through domestic choices similarly sensitive to these calibration problems? To address this question, we introduce a salience parameter, θ, into the model. When $\theta = 2, 1/2, 1$, it means that the second-period foreign policy choice is twice as, half as, and equally important as the first-period domestic policy choice, respectively.[24] We test the robustness of the trust-building equilibrium to the inclusion of salience in Appendix A.4. Here we describe the primary empirical implication of that result:

Result 2a If the independence threshold described in Result 1a ($\beta > \frac{1}{1+a}$) holds, then there is no Goldilocks problem.

1. There is no upper bound on the salience of domestic choices for which we can support the trust-building equilibrium. The security-seeker prefers first-period cooperation for any level of salience, even at very low levels of initial trust.

2. There is still a lower bound on the salience of domestic choices for which we can support the trust-building equilibrium. The greedy type can profitably deviate to first-period cooperation if first-period salience is low.

This finding has two important substantive implications that help us contrast international costly signals and domestic choices. First, trust-building via domestic choices entails an easier

[24] Of course, each domestic policy choice will have a particular level of salience, payoff independence, and relationship to foreign policy motives. The value of the model is that we can study each dimension independently.

calibration task than international trust-building. In international politics, with high dependence, states must build trust on issues of moderate stakes, neither too high nor too low. But with trust-building at the domestic level, given sufficient independence, no first-period choice is too salient for security-seekers to make the choice that signals their type. Therefore, rather than the baby-steps toward establishing trust that we would expect to see with international signaling, states can potentially build trust much more rapidly through major domestic choices that have high stakes.

Second, the result has implications for how domestic and international actions can work together. Initially, when the inability to perfectly calibrate signals makes international signaling challenging, states can instead build trust through domestic choices. Since domestic choices are imperfectly correlated with international preferences, they may not completely resolve the trust problem. But they may increase trust enough to make international signals of reassurance, such as arms control agreements, more feasible. Thus, states may start the trust-building process with independent domestic choices and then shift to more dependent international signals to build trust further.

We might wonder: If domestic choices work so well for international trust-building, does this mean that countries deliberately manipulate their domestic policies to build trust? While our model does not rule this out, it suggests that trust building through domestic choices works best *when it is not intentional*. The salience parameter in our model (θ) captures how much a state cares about its initial domestic choice relative to the second-round international interaction. When a state cares relatively more about the international round, it can seek to manipulate its domestic policy to gain international cooperation. If a state follows this strategy, it would increase international trust, but only to a limited extent. The reason is that observers understand a greater focus on second-period payoffs increases the incentive for first-period deception. Therefore, observers cannot rule out that a greedy state is manipulating its domestic policy to look like a security-seeker so that it can exploit future cooperation.

In contrast, when the current domestic choice is much more salient to a state than the future international interaction, the state will have little or no incentive to take into account the opinion of foreign observers when setting its domestic policy. Therefore, the state will make the choice based on its true preferences, which (assuming sufficient preference correlation) reveals whether it is security-seeking or greedy. Ironically, a state thus conveys the most information through its domestic choices when it does not consider their international implications.

Naturally, major domestic choices (e.g., overhauling human rights policy) will have a greater salience than smaller domestic choices (e.g., freeing a single dissident). This implies that we may occasionally see countries make small domestic policy gestures to influence international opinion, but these gestures will have a modest impact. Major domestic policy choices are less likely to reflect concern with foreign opinion, and counterintuitively will change foreign opinion to a greater extent.

Result 2b Independent domestic policy choices that are intrinsically important for the state taking them will have greater, and potentially unintended, consequences on international trust than choices that are lower in salience.

3 Implications for Ending Enduring Rivalry: End of the Cold War

Our theory of trust-building is broadly applicable, but for purposes of testing it, we focus on two types of trust-building that existing research struggles to explain: ending enduring rivalries and illiberal trust-building. We first consider enduring rivalries. Existing research lacks a satisfying explanation for how two bitter rivals whose trust has completely eroded can find the right-sized gestures to launch a trust-building process. In this section, we illustrate empirically how independent domestic choices help enduring rivals overcome this problem by detailing the end of one particularly important enduring rivalry: the Cold War. In the next section, we shall examine illiberal trust-building.

Our focus on the Cold War is driven by two considerations. First, it is one of the most important historical examples of trust-building. The end of the Cold War is notable not only for the dramatic consequences it had for the entire international system, but also for how quickly it occurred. In 1979, trust between the West and Soviet Union plummeted due to Moscow's invasion of Afghanistan. In the early 1980s, both the U.S. and U.K. elected anti-Communist leaders, Ronald Reagan and Margaret Thatcher. Reagan called the Soviet Union an "evil empire" whose Communist ideology would end up on the "ash heap of history" (Garthoff 2000, 9-11). Yet the 1980's turned out to be the last years of the Cold War. In 1988, Reagan stood with his counterpart, Mikhail Gorbachev, in Red Square and declared that he no longer saw the Soviet Union as an evil empire, that such beliefs were of "another time, another era" (Miles 2020, 62). By 1991, the Cold War reached a peaceful conclusion.

A second reason for our focus on the Cold War is that previous research has used it to illustrate how international gestures build trust. Most prominently, Kydd identifies the Intermediate-range Nuclear Forces (INF) Treaty, signed in late 1987 and ratified in 1988, as the key turning point in East-West relations. The INF Treaty represented the Soviet's "first important costly signal" to the United States and its allies that Moscow had security-seeking intentions (Kydd 2005, 227). In this account, the process of reassurance and rapprochement was initiated by the INF Treaty and reinforced primarily by additional Soviet foreign policy choices. While we acknowledge the importance of the INF Treaty, our review of the history suggests that Soviet domestic policies, such as improved treatment of dissidents, *glasnost*, and *perestroika*, kick-started the trust-building process even earlier. Therefore, analyzing this case enables us to show how domestic and international policies work together, with domestic policies playing a potentially necessary preliminary role.

Our approach to the Cold War assumes that both sides genuinely desired cooperation by the late Cold War, as signals of security-seeking intentions can only establish cooperation among states that genuinely desire it. This assumption is consistent with recent scholarship from Fritz Bartel, who argues that Soviet motives shifted following economic shocks in the 1970's (Bartel 2022), and also with more classic research on the effects of both "New Thinking" and structural constraints (Risse-Kappen 1994; Brooks and Wohlforth 2000). While establishing precisely how cooperative motives arose is beyond the scope of our study, we ask how the U.S. and Soviet Union were able to *recognize* each other's cooperative motives after decades of distrust.

Like Kydd, we focus our analysis on U.S. trust of the Soviet Union. Of course, the trust-building in our model, as in Kydd's, is two-sided. Yet, as we showed, even if one side holds low initial trust, international trust-building is not an equilibrium. By the late Cold War, the United States was the more distrusting side. Whereas Soviet leader Yuri Andropov had seriously worried about a surprise U.S. nuclear attack, discussion of such fears abated when Gorbachev took office. Moreover, Gorbachev, motivated by Soviet financial constraints, was more eager to push for arms control than Reagan, showing little concern that the U.S. would cheat on such agreements (Brown 1997; Wilson 2014; Chernyaev 1993).Therefore, our analysis focuses on how Soviet domestic policy alleviated U.S. concerns that the Soviet Union could not be trusted.

Following best practices in case (i.e, within-unit) evaluation of models (Gerring 2004),we verify that the Cold War case fits parameter ranges where domestic signaling is plausible, but international signaling is difficult (Bates 1998; Lorentzen, Fravel, and Paine 2017). Although the U.S. and Soviet Union had discussed arms control proposals from the early 1980's onward, the benefits of arms control were highly dependent (low β), creating an impediment to progress under low trust. In contrast, the Soviets' domestic policy choices had moderate or high independence (moderate or high β). The payoffs from more lenient treatment of dissidents, greater free speech, and reduced central planning of the economy were not, for the most part, dependent on Western policies. For example, the main benefits that Soviet leaders expected to receive from *glasnost*, a policy of greater openness and freedom of speech, were reduced corruption and increased efficiency. These benefits did not depend greatly on Western responses. We also classify Soviet domestic choices as having high salience (θ). Policies such as *glasnost* and *perestroika*, which introduced some market forces into the Soviet economy, created a fundamental shift in Soviet political culture and governance, with direct implications for the Soviet Union's governing ideology and the Kremlin's control over the population. Finally, we argue that domestic and international preferences were sufficiently correlated to enable learning. The Cold War had a strong ideological component, pitting a liberal, capitalist West against an illiberal, Communist Eastern Bloc. The nature of Soviet domestic reforms reflected a shift in intrinsic values, indicating greater compatibility in international preferences between East and West. Overall, given the relatively high independence and salience of Soviet domestic choices, and the correlation between domestic and international preferences, our theory predicts that these choices had all of the characteristics necessary to launch trust-building.

3.1 Initial Trust-Building through Domestic Choices

East-West trust in the early 1980s could hardly have been lower. President Reagan and officials throughout his administration believed that the Soviet Union was bent on global domination. To Reagan, the Soviets had exploited the U.S. during the 1970s, opportunistically jumping out ahead in the arms race and engaging in military adventurism while misguided Western leaders sought détente. As Reagan prepared to take office, he confided in a friend that “I don’t really trust the Soviets and I don’t really believe that they will really join us in a legitimate limitation of arms agreement” (Wilson 2014, 17). In his first press conference as commander-in-chief, Reagan expressed his belief that the Soviets “reserve unto themselves the right to commit any crime, to lie, to cheat, in order to attain” their goal a global Socialist revolution (Gwertzman 1981).

In 1985, Mikhail Gorbachev became General Secretary of the Communist Party of the Soviet Union. Gorbachev and his advisors desired to end the costly competition between the Soviet Union and the US (Brands 2014). However, the combination of low trust and low independence inherent in defense policy made it difficult to initiate trust-building through costly international actions such as arms control. Reagan administration officials were wary of arms control with the Soviets, believing that “the Soviet Union had violated every arms agreement it had ever signed” (Wilson 2014, 66). As a result, arms control proposals in the early and mid-1980s were often more about public relations than genuine attempts to build trust (Colbourn 2022). Although the Soviet government under Gorbachev was more eager for arms control and less fearful that the U.S. would cheat (Chernyaev 1993), making unilateral arms cuts would have raised security concerns and potentially damaged Soviet prestige. Therefore, arms control could not proceed without trust on both sides.

Fortunately, domestic choices offered another means for initiating reassurance. Consistent with our theory, a shift in Soviet policy on dissidents and minorities became an early signal that the Kremlin might be a more trustworthy partner for the West. One particularly notable case was that of Andrei Sakharaov, the Soviet nuclear physicist and Nobel Peace Prize-winning dissident. Sakharov's release from internal exile in 1986 was a major event and increased Western trust that Gorbachev and his government were different from their predecesors. According to Shultz, Sakharov's freedom, "affecting a man of towering intellect and moral authority, made an impact on some of Gorbachev's most severe skeptics" (Shultz 2010, 1095). More systematic policy changes re-enforced Western beliefs. Jack Matlock, the senior Soviet expert on the Reagan NSC and later U.S. ambassador to Russia, highlighted their impact on Shultz's beliefs:

> [T]he evolution in Shevardnadze's attitude toward human rights in the Soviet Union made *probably the most important contribution* to Shultz's feeling that the two had compatible goals. Shevardnadze had always tolerated a discussion of human rights with more courtesy than Andrei Gromyko could summon, but by 1987 he began to do more than simply arranging an exit visa once in a while. He actually began to try to change the system (Matlock 2004, 265).

From 1986 to 1988 the number of exit visas issued to Soviet Jews ("refuseniks") increased from 1,000 to 80,000, a shift which even the CIA—generally more skeptical of Soviet intentions than others in the U.S. government—called "remarkable" (Brands 2014, 137). Illustrating the moderate-to-high independence of the issue, the Soviets exhibited little concern that these moves could be exploited by the West. Shevardnadze even invited Shultz to provide Moscow with a list of potential émigrés for the Soviets to consider (Shultz 2010, 986).

Gorbachev and his allies in the Kremlin eventually pursued broader reforms, *glasnost* and *perestroika*, aimed at increasing political and economic freedom. *Glasnost* focused on transparency and openness. For example, in 1986 Soviet leaders green-lighted the release of the film *Repentence*, which critically represented Stalin, knowing full well that it was a "bombshell" that would "change our social system" (Taubman 2017, 248). The film's significance registered abroad. The

New York Times called its release a "cultural and political milestone" (Barringer 1986). By 1987, *glasnost* was "spreading like wildfire on the steppe," and had led to something closer to freedom of speech and freedom of the press (Taubman 2017, 314). Under *perestroika*, new laws legalized private enterprise (1986), allowed state enterprises to determine output on the basis of demand (1987), and permitted co-operatives (1988). As with emigration policy, the Soviets did not seem particularly concerned that *perestroika* and *glasnost* left the USSR vulnerable to exploitation by the West. On the contrary, Shultz and Gorbachev enjoyed open discussions about economic policy (Wilson 2014, 132-33). However, these larger reforms were designed primarily to benefit the Soviet domestic economy rather than specifically to reassure the West (Bartel 2022).

Nonetheless, *perestroika* and *glasnost* further increased Western trust of the Soviet Union. As early as April 1986, Soviet reforms had sparked "prominent" deliberations in the U.S. government about the possibility that the Soviets were truly changing (Savranskaya, Blanton, and Zubok 2010, 116).[25] By 1987, Soviet reform had convinced senior U.S. policymakers that international cooperation was possible. During a meeting between Gorbachev and Thatcher in late 1987, journalists observed that Thatcher "placed almost as much emphasis on the Soviet leader's internal reforms as on the superpower talks, considering *perestroika* and *glasnost* as evidence of a determination which also promised progress in East-West negotiations" on arms control and other security issues (Naughtie 1987). Reagan concurred, subsequently saying that Gorbachev's book, *Perestroika*, which outlined the Soviet leader's vision for reform, made him hopeful about relations between Washington and Moscow (Matlock 2004, 294).

3.2 International-Level Trust-Building Begins

Soviet domestic reforms not only built trust, but facilitated meaningful international cooperation, including the INF Treaty, the first U.S.-Soviet agreement to significantly *reduce*, rather than limit, nuclear arms. The conclusion of the INF Treaty reflected, to be sure, important concessions

[25] Not everyone agreed that the Soviet Union was changing for good, but the very existence of debates is evidence of changing U.S. attitudes, given initial certainty in the Reagan Administration that the Soviets had malign intentions.

by Gorbachev (e.g., including the SS-23 in the deal, de-linking missile defense). However, its signing was aided by the goodwill that Gorbachev had generated with *glasnost* and *perestroika*. During his April 1987 visit to Moscow, Shultz and Gorbachev discussed Soviet economic reform in-depth, "establishing greater trust" and helping to make "the prospect of the elimination of INF a reality" (Wilson 2014, 133). Aiming to whet Reagan's appetite for a summit in Washington at which an INF Treaty could be signed, Shultz reported from Moscow that "the Soviet Union is changing" (Leffler 2007, 399).

When Gorbachev eventually did come to Washington to sign the INF Treaty in December 1987, he received a hero's welcome. Gorbachev interpreted the success of the Washington summit as evidence that his reformist domestic program changed perceptions of the Soviet Union abroad. Briefing the Politburo afterwards, Gorbachev observed:

> In Washington we saw for the first time with our own eyes what a great interest exists for everything that is happening here, for our *perestroika*. And the goodwill, even enthusiasm to some degree, with which prim Washington received us, was an indicator of the changes that have started taking place in the West. These changes evidence the beginning of the crumbling 'image of the enemy,' beginning of the destruction of the 'Soviet military threat' myth (Savranskaya et al. 2010, 361).

Soviet domestic reform also built trust with the U.S. Senate, which would go on to ratify the INF Treaty. As the influential chairman of the Senate Armed Services Committee, Sam Nunn, observed, "the advent of Gorbachev, *glasnost*, and *perestroika* has undeniably improved the overall climate for the conduct of superpower relations" (Nunn 1988).

In 1988, the Kremlin's domestic reforms provided even stronger evidence that the Soviet Union had fundamentally changed. On the eve of Reagan's May visit to Moscow, the Kremlin released a set of "theses" for the upcoming 19th Party Conference which indicated that Gorbachev wanted to further liberalize the Soviet system. Matlock, by then the U.S. ambassador to Russia, was "electrified" when he read the document (Savranskaya et al. 2010, 110). The following day, the ambassador briefed Reagan, telling him that "the Soviet Union will never be the same" (Matlock

2004, 296). Just a few days later, in Red Square, Reagan declared that he no longer saw the Soviet Union as an "evil empire."

The 19th Party Conference in the summer of 1988 indeed proved a major milestone in the Kremlin's reform program. Though falling far short of democracy, the conference resulted in a political liberalization unprecedented in the Soviet context. Gorbachev and his allies secured popular elections at lower levels of government and greater judicial independence and rule of law. The conference "dealt only briefly with international and security affairs" (Garthoff 2000, 361). The lack of attention to foreign policy is notable both because it speaks to the relative independence of domestic reform, and because a party conference focused on domestic policy had such a big impact on Western perceptions of the Soviet Union. The conference contributed further to Washington's appetite for cooperation with Moscow; high-level diplomacy with the Soviets thereafter "expanded rapidly" (Matlock 2004, 306). In September, Gorbachev worked to entrench his reforms by removing several conservative hardliners from high-ranking posts. Speaking at a roundtable on the end of the Cold War, the senior CIA Soviet analyst during the mid- to late-1980s identified these events as a critical juncture for Western perceptions of the Soviet Union:

> When I talked about 1988, it was after the 19th Party Conference, and then in the period after that, in September, when the major restructuring took place. . . before that happened there was still room for those who wanted to disparage the implications of the events in the USSR to make their arguments. Whether you believed it or not, they had room to argue that 'that's all right, it will eventually drift back to a Brezhnev-style system.' But I think that after the end of 1988, no matter what your slant, you could not very well argue that some major lines had not been crossed. . . You could not very well argue that it was just talk and political rhetoric (Savranskaya et al. 2010, 116-17).

All told, domestic reforms undertaken by the Kremlin from 1986 to 1988 played a crucial role in creating an atmosphere in which international cooperation was easier to sustain. In addition to the INF Treaty, the U.S. and Soviet Union concluded a number of smaller but meaningful agreements on issues as varied as monitoring nuclear tests, peaceful nuclear energy, fishing rights, space exploration, cultural and educational exchanges, and maritime navigation (Garthoff 2000, 353). Although U.S.-Soviet rapprochement slowed somewhat during the transition between the Reagan and Bush presidencies, Bush declared in fall 1989 that "[t]he world will be a better place if *perestroika* succeeds" and laid out more than a dozen proposals to increase U.S.-Soviet cooperation, including lifting trade restrictions and supporting Soviet efforts to join the GATT as an observer (Engel 2017, 297). Soviet policymakers saw this as "the end of economic warfare" between the two states (Engel 2017, 298). US-Soviet cooperation continued to bear fruit even as the Soviet Union disintegrated and revolution swept through Eastern Europe. The U.S. and Soviet Union concluded the Treaty on Conventional Forces in Europe and the START Treaty, and the Kremlin further demonstrated its benign intentions through domestic choices by showing restraint toward separatists in the Baltics.

Overall, given the independence and high salience of Soviet domestic choices, our theory predicts that trust-building should have been possible despite low initial trust, and that Soviet domestic choices should have had a substantial impact on Western perceptions. This is what we observe. As shown above, domestic political reform was among the earliest indicators of changing Soviet intentions. Moreover, domestic policies, including Soviet treatment of dissidents, *glasnost*, and *perestroika*, were seemingly as or more important than foreign policy in convincing key Western officials that the two sides had compatible goals after decades of tension and conflict. By increasing trust, Soviet domestic reforms thus contributed to an unprecedented level of East-West cooperation.

4 Illiberal Trust-Building

While Soviet liberalizing reforms reassured U.S. officials of the Soviet desire to cooperate with the United States—itself a liberal state—liberalizing reforms do not always have this effect. For example, much earlier liberalizing reforms in the Soviet Union under Khrushchev arguably worsened relations with China and contributed to the Sino-Soviet split (Haynes and Yoder 2020). In fact, illiberal domestic policies can have a more reassuring effect in some circumstances. This is a phenomenon that the democratic peace theory and research by Kydd are unable to explain. In contrast, our theory allows for the possibility that which domestic policies are perceived to be correlated with a desire to cooperate internationally can vary with the context and depend upon which state is evaluating the policies. Therefore, our theory can encompass both liberal and illiberal trust-building. We illustrate illiberal trust-building with several case anecdotes. Most are drawn from the Cold War, but we also consider the British reaction to Hitler's antisemitic policies in the run-up to World War II.

4.1 US Relations with Autocratic Governments in the Cold War

The Cold War illustrates how domestic policies that are viewed as reassuring by one state can be viewed as alarming by another. For example, when potential client states adopted policies such as land reform and nationalizing industry, it increased Soviet trust but reduced U.S. trust. We focus, however, on a more surprising aspect of the Cold War: Although the United States was a liberal state itself, it was sometimes reassured by the illiberal domestic policies of others, as long as they were anti-Communist.

Throughout the Cold War, the U.S. feared that political instability in developing states would create a breeding ground for Communism and undermine American security by bringing these states into Moscow's and Beijing's orbit. Thus, the emergence of new governments in the developing world presented a vexing problem for Washington. The U.S. wanted to support fledgling regimes that would join U.S. efforts to contain Communist influence internationally. However, some governments might exploit U.S. support initially, only to align with the Communist camp over time. How could the U.S. determine which governments it could trust and support, and which were lost causes?

We argue that the U.S. came to see the domestic politics of new governments in Latin America, Asia, and elsewhere as a key indicator of their reliability as partners. Despite Washington's oft-stated goal of fostering the spread of democracy, Washington found itself reassured by decidedly non-democratic and illiberal policies, including the brutal repression of leftist opposition groups and top-down, right wing economic development planning.

Below, we briefly highlight U.S. policy in South Korea, Indonesia, and Latin America during the first two decades of the Cold War. We show that illiberal policy choices by these states reassured Washington because they indicated a shared preference for limiting the expansion of Communism internationally. These states lacked feasible options to send international signals of reassurance without U.S. cooperation, so their domestic policies were key to establishing initial trust.

South Korea. When a group of South Korean military officers overthrew the country's democratically elected government in May 1961, Washington's initial reaction was concern. While the U.S. had some anxiety about the stability of the young government of erstwhile Prime Minister Chang Myon, the coup plotters were not immediately seen as preferable (Brazinsky 2009). In particular, their intentions were a question mark, and the coup leader, Park Chung-hee, had prior Communist ties, leading one intelligence estimate to comment that "we cannot rule out the possibility that [Park] is a long-term Communist agent, or that he might redefect" (FRUS 2024, 1961–1963, Volume XXII, Northeast Asia, No. 224).

The coup came at a precarious moment, with U.S. intelligence analysts concerned that a "lack of national purpose" in South Korea coupled with "enticements" from Pyongyang, "could lead to some movement in the south toward an accommodation with the north" (FRUS 2024, 1961–1963, Volume XXII, Northeast Asia, No. 243). Beyond its potential to fall under the influence of North Korea, U.S. officials viewed the fate of South Korea in terms of Domino Theory; as President Kennedy himself argued, if "Korea were not free, Japan would not be free and that would mean the whole Pacific area would go too, so Korea has a vital interest for us" (FRUS 2024, 1961–1963, Volume XXII, Northeast Asia, No. 247).

Given its concerns, Washington was greatly encouraged by the military government's anti-Communist actions and swift, if authoritarian, implementation of modernizing policy reforms across South Korea's economy and society. In a telegram from the U.S. embassy in Seoul in

October 1961, the reassurance felt by American officials is palpable. American Ambassador Samuel Berger wrote that the military regime "has taken hold with energy, earnestness, determination and imagination, albeit with certain authoritarian and military characteristics which have hampered its public image" (FRUS 2024, 1961–1963, Volume XXII, Northeast Asia, No. 244). Further, "[v]igilence against communist subversion and quality and volume anti-communist propaganda have greatly improved." Concluding, Berger expressed that "From United States point of view. . . this government still offers much hope" (FRUS 2024, 1961–1963, Volume XXII, Northeast Asia, No. 244).

Efforts to improve economic performance—even by non-democratic means—and a demonstrated anti-Communist orientation facilitated U.S. cooperation and support. The positive impression that the military government had made early in its tenure led the initially skeptical Kennedy administration to host Park at the White House in 1961, which "played a significant role in stabilizing and legitimizing the South Korean regime" (Brazinsky 2009, 120-21). Ambassador Berger, reflecting on the state of play in Korea at the end of 1961, wrote that despite "shortcomings"—i.e., its authoritarian style—the U.S. could expect that new aid "will be more effectively used than by any previous government" and that "the record of reform and improvement which has taken place is heartening and suggests we can go to Congress this spring in good conscience that this has been a constructive year and that our continuing massive support is well justified" (FRUS 2024, 1961–1963, Volume XXII, Northeast Asia, No. 244). Overall, though Park Chung-hee's military "revolution" was a step backward for democratization, it nonetheless proved reassuring to Washington because it signaled compatible interests with the U.S.

Indonesia. Following Indonesia's independence in 1949, American officials questioned the reliability of President Sukarno, a hero of the Indonesian nationalist movement. By the early 1960s, American officials were increasingly worried not only that Sukarno would lead Indonesia into the Communist camp, but that American aid was enabling him. Lyndon Johnson feared that "the closer we get to Sukarno the more difficult he becomes" and questioned whether support for Indonesia was in the "national interest" (Schmitz 2006, 41). In particular, American officials suspected that Sukarno had become captured by Indonesian leftists. Reacting to a speech by Sukarno in 1964, the CIA concluded that the Indonesian president was pursuing a "course—both international and domestic—which is close to the immediate objectives of the Indonesian Communist Party" (FRUS 2024, 1964–1968, Volume XXVI, Indonesia; Malaysia-Singapore; Philippines, No. 62).

The emergence of General Suharto of the Indonesian Army as a key political player in 1965 greatly shifted U.S. perceptions.[26] Suharto proved to be an anti-Communist, and reversed the leftward drift of the Indonesian government. Specifically, Suharto moved to brutally eradicate suspected members of the Communist Party of Indonesia (PKI), resulting in what the State Department characterized as the "decimation of the PKI as an organized political force."(FRUS 2024, 1964–1968, Volume XXVI, Indonesia; Malaysia-Singapore; Philippines, No. 202) With hundreds of thousands killed, the CIA estimated that "the anti-PKI massacres in Indonesia rank as one of the worse mass murders of the 20th century, along with the Soviet purges of the 1930's, the Nazi mass murders during the Second World War, and the Maoist bloodbath of the early 1950's" (Schmitz 2006, 48). Indeed, the evidence suggests that the U.S. knew many of the "victims were entirely innocent" (Bevins 2017).

Yet while anti-Communist repression in Indonesia was judged by U.S. officials to rival the worst excesses of America's main 20th century adversaries, Washington was largely reassured by Suharto's ruthlessness because it marginalized the Indonesian left and, therefore, indicated that Indonesia would remain reliably outside Beijing's sphere of influence (Bevins 2017). American officials communicated to the Indonesian army that the "Embassy and the USG" were "generally sympathetic with and admiring of what the army [was] doing" (Schmitz 2006, 49). The NSC reflected that Jakarta had been "well on the way to becoming another expansionist Communist state" but conditions had "sharply reversed" (Schmitz 2006, 51). Having previously feared that aid to Indonesia was contrary to U.S. national interests, President Johnson in March 1966 authorized emergency food aid to Jakarta to help put the two countries "on the road back to cooperative

[26] Suharto, who rose to prominence following an attempted coup against Sukarno, would later sideline Sukarno entirely.

relations" (Schmitz 2006, 49).

Latin America. Since the presidency of James Monroe, American policymakers had viewed a stable and friendly Western Hemisphere as a vital interest. Large trade flows connected the U.S. with Latin America, and the region was an important source of raw materials. Though the U.S. would have preferred the emergence of stable liberal democracies in the aftermath of the Second World War, Washington repeatedly confronted the reality of illiberal governments in Latin America. In this context, the crucial task was determining whether a given autocratic regime could be counted on to support U.S. interests. According to the historian David Schmitz:

> In an analysis that became central to justifications for supporting right-wing dictators throughout the Cold War, [a 1949 State Department] report emphasized that in determining policy toward these nations 'it is important to determine if a dictatorial regime is of the traditional Latin American military or authoritarian type, or if it is of Communist, Nazist, or other police state type.' This distinction was crucial. The former was acceptable, but the latter was grounds for 'legitimate concern' (Schmitz 1999, 172).

As such, domestic repression in Latin America could be reassuring to U.S. policymakers if it targeted the "correct" political opposition, such as Communists. None other than George Kennan explained the logic in a meeting with U.S. ambassadors to Latin American countries. Kennan opined that the U.S. "should not hesitate before police repression by the local government. This is not shameful since the Communists are essentially traitors." It was better, Kennan argued, to have an anti-Communist authoritarian government than even "a liberal government if it is indulgent and relaxed and penetrated by Communists" (Schmitz 1999, 149-51). Since stable, pro-Western democracies were unlikely to pop up, American policymakers came to see right-wing authoritarian rule as more or less "desirable" (Schmitz 1999, 149-51).

In this context, harsh repression at home could be used by Latin American dictators to signal their reliability as partners to the U.S., thereby earning American support. Dictators such Anastasio Somoza, who was credited by the U.S. ambassador in Nicaragua for keeping "his foot firmly on the spark of communism here," could be welcomed to the White House with "their anticommunist credentials" seen as a signal of alignment with U.S. interests (Schmitz 1999, 155-56).

4.2 Nazi Germany's Reassurance of the United Kingdom

Another example of illiberal domestic policies building trust is Hitler's anti-Jewish policies prior to the 1938 Sudetenland crisis. In the crisis, Hitler—who had already remilitarized the Rhineland and annexed Austria—demanded control over the Sudetenland, a German-populated region of Czechoslovakia. Hitler threatened to attack if his demand was not met, but promised that this would be the end of his territorial ambitions. In fact, Hitler told British Prime Minister Chamberlain that he "wanted to be friends with England" (Churchill 1948, 307-08).

Whether to trust that Hitler really would cease seeking more territory and cooperate with Britain after receiving control of the Sudetenland was a key question for the British government. A policy document from Alexander Cadogan of the British Foreign Office explains why. As Telford Taylor summarizes, Cadogan argued that the Sudetenland was not worth fighting over in itself and therefore "the case for supporting Czech territorial integrity must depend, not on the Sudetenland, but on what Hitler might do *after* annexing it" (Taylor 1979, 625, emphasis original).[27]

When the Sudetenland crisis reached its peak in September 1938, the infamous Kristallnacht pogrom was still two months in the future and mass killings in concentration camps had not yet begun. However, Hitler's government had already issued dozens of policies excluding Jews from participation in the economy, education, and public life and singling them out for persecution. The British government was fully aware of these policies from diplomatic reports. An early report from Ambassador Rumbold emphasized that "the National Socialist programme is intensely anti-Jewish," and predicted, "it is certainly Hitler's intention to degrade and, if possible, expel the Jewish community from Germany ultimately" (Ascher 2012, 25). In 1935, MI6 officer Edward Foley wrote a report on "The Situation of Jews in Germany," which concluded that "the ultimate

[27] Of course, Britain had an incentive to delay fighting while it built up its own capabilities, and we acknowledge this might have also influenced British decision-making. However, beliefs about Hitler's aims played a role in the British internal debate, and we focus on the reason for these beliefs.

aim [of the Nazi Party] remains the disappearance of the Jews from Germany or, failing that, their relegation to a position of powerlessness and inferiority in Germany" (Ascher 2012, 40). Moreover, there was widespread condemnation of anti-Jewish actions such as the boycott of Jewish shops in 1933 and anti-Jewish violence during the annexation of Austria among the British public (Kushner 1989, 148).

Although Kydd argues that policies discriminating against domestic minorities should be a source of international distrust, Hitler's anti-Jewish policies actually increased British trust that Hitler would abide by his promise not to expand Germany's territory beyond the Sudetenland (Kydd 1997). While some members of the British elite shared Hitler's antisemitic beliefs and approved of his policies (Kushner 1989), the reasoning of key decision-makers was more subtle. They believed that Hitler's persecution of the Jews was more likely to be associated with a racially-motivated desire for German unification than a desire for global expansion. The former could allow future cooperation between Britain and Germany, whereas the latter meant war.

Prime Minister Chamberlain told his cabinet, "The crucial question was whether Herr Hitler was speaking the truth when he said that he regarded the Sudeten question as a racial question which must be settled, and that the object of his policy was racial unity and not the domination of Europe. Much depends on the answer to this question" (Parker 1993, 169). Since expanding Germany's territorial control beyond the Sudetenland would bring people other than Germans into the Third Reich and undermine Hitler's apparent goal of a racially pure Germany, Chamberlain expected Hitler's racism to be negatively correlated with a desire for international expansion. Chamberlain told his cabinet, "the seizure of the whole of Czechoslovakia would not be in accordance with Herr Hitler's policy, which was to include all Germans in the Reich but not include other nationalities" (Weinberg 2005, 549). Similarly, Chamberlain told his inner circle that he was "satisfied that Herr Hitler was speaking the truth" when Hitler said he viewed the Sudetenland dispute "as a racial question" (Faber 2008, 345). While Chamberlain was the clearest proponent of this line of reasoning, some other officials seemed to think similarly. For example, Foreign Secretary Halifax wrote that he did not think it "necessary to assume that Hitler's racial ambitions are necessarily likely to expand into international power lust" (Weinberg 2005, 549).

Of course, Hitler was actually expansionist. The British were correct to infer from Hitler's domestic policies that he was a genuine racist. After all, Hitler's policies lowered Germany's productivity, took resources to implement at a time when Germany was in severe debt, and invited international backlash. Since these policies were costly, Hitler's choice of them was a credible signal of racist beliefs. Yet the British were wrong about how racism correlated with international expansionism. They assumed, somewhat logically, that a racist would not want to expand Germany's borders to incorporate non-German people. However, the British failed to consider the Nazi concept of *Lebensraum*, the racist idea that German superiority entitled the German people to more "living space." This shows that the correlation between domestic and international policy preferences is sometimes uncertain in the real world.

To be clear, we do not claim that racist domestic policies will always or even commonly promote international trust. Soviet persecution of Jews undermined U.S.-Soviet relations during the Cold War and, as noted above, Gorbachev's relaxation of emigration restrictions for Jews facilitated trust-building. The Anglo-German case is relatively unique because of how Hitler framed his international demands and domestic policies. However, we use it to show that the types of domestic policies that can promote expectations of international cooperation are wide-ranging and can be illiberal.

5 Conclusion

We argued that domestic policy choices which are sufficiently correlated with foreign policy motivations can operate as costly signals that engender international trust. Moreover, costly domestic choices hold a key advantage over international choices that make them vital for trust-building: payoff independence. Independence means that the value a state accrues from domestic choices depends mainly on that state's true motivations, and less on how their counterpart responds. We focus particularly on situations where initial trust is so low that security-seekers are unwilling to use international actions to signal their motivations because their fear of exploitation is too great. In these conditions of extreme mistrust, even moderately independent domestic actions can facilitate trust-building that would otherwise be impossible. Thus, the domestic reforms that often occur before rivals achieve a rapprochement may not be a coincidence. Rather, they may be an important, potentially necessary, step that enduring rivals take to increase trust such that they are willing to engage in the international trust-building activities that others have studied (Kydd 2005; Yoder and Haynes 2021).

This project makes a number of important contributions. First, it contributes to literature on domestic politics and signaling in international relations (Goldfien et al. 2023; McManus 2017; Weeks 2008; Schultz 1999; Renshon et al. 2023; Fordham 1998). Second, it offers a new way of understanding the democratic peace and suggests the possibility of trust-building among a wider variety of regimes, based on domestic policy changes that fall short of regime change. Third, it expands our understanding of the set of policy choices that are relevant to international relations. Most importantly, it offers a novel solution to the problem that the trust-building literature has wrestled with for decades: how countries with high levels of distrust can engage in initial trust-building activities without exposing themselves to too much risk (Glaser 2010; Jervis 1978; Kydd 2005).

Our empirical analysis provides new insight into reassurance at the end of the Cold War, offering a clearer explanation for the linkage between reforms within the Soviet Union and rapprochement abroad. It also illustrates how even illiberal actions, such as suppressing left-wing groups and oppressing minorities, can increase trust in some circumstances, a puzzle that is not explained by the democratic peace approach. Finally, we provide insights into modern-day rivalries, which are particularly relevant during a time of increased tension with Russia and China. While these countries may currently have incompatible goals with the West, if their preferences eventually shift toward a desire for greater cooperation, our theory suggests a feasible way for Western policymakers to recognize this.

Our findings also speak to important policy issues, including the evolving competition between the United States and China. Many have argued that the U.S. and China have entered a period of rivalry or even a Cold War (Sanger 2021; Bekkevold 2022; Daly 2022). This is concerning because major power rivalries are often long and very costly (Thompson 2001). Intense competition between rivals lowers trust, making it hard to find a path back to peace even if both sides tire of competing. However, our research identifies how the U.S. and China may eventually be able to kick-start trust-building through unilateral domestic policy changes.

We also clarify how the United States can identify which states are trustworthy international partners. As rising populism brings new governments to power, the U.S. can no longer be certain about the foreign policy intentions of key international partners. Existing advice to put trust in states that ascribe to liberal values has limited utility in a world where the U.S. itself is moving toward protectionism, and many core U.S. partners are autocratic or experiencing democratic backsliding. If the U.S. distrusts all states that fall short of liberal ideals, it will potentially miss important opportunities for international cooperation. Our theory suggests that the U.S. can take a more nuanced look at the domestic policies of potential international partners and identify which individual policies are likely to be correlated with a desire to cooperate with or cheat the U.S. Our theory suggests that the U.S. can take a more nuanced look at the domestic policies of potential international partners and identify which individual policies are likely to be correlated with a desire to cooperate with or cheat the U.S.

References

Acharya, A. and K. W. Ramsay (2013). The calculus of the security dilemma. *Quarterly Journal of Political Science 8*(2), 183–203.

Altman, D. and M. M. Lee (2022). Why territorial disputes escalate: The causes of conquest attempts since 1945. *International Studies Quarterly 66*(4), sqac076.

Ascher, A. (2012). *Was Hitler a Riddle? Western Democracies and National Socialism*. Stanford: Stanford University Press.

Ashworth, S., C. Berry, and E. B. de Mesquita (2021). *Theory and Credibility*. Princeton.

Barringer, F. (1986). 'repentance,' a soviet film milestone, strongly denounces official evil. *New York Times*.

Bartel, F. (2022). *The Triumph of Broken Promises: The End of the Cold War and the Rise of Neoliberalism*. Harvard University Press.

Bas, M. A. and A. J. Coe (2016). A dynamic theory of nuclear proliferation and preventive war. *International Organization 70*(4), 655–685.

Bates, R. H. (1998). *Analytic narratives*. Princeton University Press.

Bekkevold, J. I. (2022). 5 ways the u.s.-china cold war will be different from the last one. *Foreign Policy*.

Berenji, S. (2020). Sadat and the road to jerusalem: Bold gestures and risk acceptance in the search for peace. *International Security 45*(1), 127–163.

Bevins, V. (2017). What the united states did in indonesia. *The Atlantic 21*.

Biden, J. (2024). President biden delivers remarks on the latest developments in syria.

Bokobza, L., S. Krishnarajan, J. Nyrup, C. Sakstrup, and L. Aaskoven (2022). The morning after: Cabinet instability and the purging of ministers after failed coup attempts in autocracies. *Journal of Politics 84*(3), 1437–1452.

Brands, H. (2014). What good is grand strategy? In *What Good Is Grand Strategy?* Cornell University Press.

Braumoeller, B. F. (2008, 2). Systemic politics and the origins of great power conflict. *American Political Science Review 102*, 77–93.

Brazinsky, G. A. (2009). *Nation building in South Korea: Koreans, Americans, and the making of a democracy*. Univ of North Carolina Press.

Brooks, S. G. and W. C. Wohlforth (2000). Power, globalization, and the end of the cold war: Reevaluating a landmark case for ideas. *International Security 25*(3), 5–53.

Brown, A. (1997). *The Gorbachev Factor*. OUP Oxford.

Cable, V. (2017). Deng: Architect of the chinese superpower. *LSE Ideas SR023*, 1–7.

Chernyaev, A. S. (1993). *My Six Years with Gorbachev*. University Park, PA: Pennsylvania State University Press.

Churchill, W. S. (1948). *The Gathering Storm*. Boston: Houghton Mifflin Company.

Coe, A. J. and J. Vaynman (2020). Why arms control is so rare. *American Political Science Review 114*, 342—-355.

Colaresi, M. (2004). When doves cry: International rivalry, unreciprocated cooperation, and leadership turnover. *American Journal of Political Science 48*(3), 555–570.

Colbourn, S. (2022). *Euromissiles: The Nuclear Weapons That Nearly Destroyed NATO*. Cornell University Press.

Crescenzi, M. J. (2018). *Of Friends and Foes: Reputation and Learning in International Politics*. Oxford University Press.

Dafoe, A., J. Renshon, and P. Huth (2014, 5). Reputation and status as motives for war. *Annual Review of Political Science 17*, 371–393.

Daly, R. (2022). China and the united states: It's a cold war, but don't panic. *Bulletin of the Atomic Scienctists*.

Debs, A. and N. P. Monteiro (2014). Known unknowns: Power shifts, uncertainty, and war. *International Organization 68*(1), 1–31.

Engel, J. A. (2017). *When the World Seemed New: George HW Bush and the End of the Cold War*. Houghton Mifflin Harcourt.

Faber, D. (2008). *Munich, 1938: Appeasement and World War II*. New York: Simon and Schuster.

Fearon, J. D. (1995). Rationalist explanations for war. *International organization 49*(3), 379–414.

Flores-Macías, G. A. (2019). *The Political Economy of Taxation in Latin America*. Cambridge University Press.

Fordham, B. (1998). The politics of threat perception and the use of force: A political economy model of u.s. uses of force, 1949-1994. *International Studies Quarterly 42*, 567–590.

FRUS (2024). *Foreign Relations of the United States*. Washington, DC: Department of State.

Garthoff, R. L. (2000). *The great transition*. Brookings Institution Press.

Gartzke, E. (1998). Kant we all just get along? opportunity, willingness, and the origins of the democratic peace. *American Journal of Political Science 42*, 1–27.

Gerring, J. (2004, 5). What is a case study and what is it good for? *American Political Science Review 98*, 341–354.

Glaser, C. L. (2010). Rational theory of international politics. In *Rational Theory of International Politics*. Princeton University Press.

Glaser, C. L. (2015). A us-china grand bargain? the hard choice between military competition and accommodation. *International Security 39*(4), 49–90.

Glaser, C. L. and C. Kaufmann (1998). What is the offense-defense balance and can we measure it? *International Security 22*, 44–82.

Goldfien, M. A., M. F. Joseph, and R. W. McManus (2023). The domestic sources of international reputation. *American Political Science Review 117*(2), 609–628.

Gwertzman, B. (1981). President sharply assails kremlin; haig warning on poland disclosed. *The New York Times*.

Haynes, K. and B. K. Yoder (2020). Offsetting uncertainty: Reassurance with two-sided incomplete information. *American Journal of Political Science 64*, 38–51.

Jervis, R. (1978). Cooperation under the security dilemma. *World politics 30*(2), 167–214.

Jervis, R. (1989). *The logic of images in international relations*. Columbia University Press.

Kertzer, J. D. (2016). *Resolve in international politics*. Princeton University Press.

Kim, S. J. (2022). Doom and gloom, from structure to human minds: What makes a north korean nuclear deal difficult? *Political Psychology 43*(4), 715–730.

Kushner, T. (1989). Beyond the pale? british reactions to nazi anti-semitism, 1933–39. *Immigrants Minorities 8*, 143—-160.

Kydd, A. (1997). Sheep in sheep's clothing: Why security seekers do not fight each other. *Security studies 7*(1), 114–155.

Kydd, A. (2000). Trust, reassurance, and cooperation. *International Organization 54*(2), 325–357.

Kydd, A. H. (2005). *Trust and mistrust in international relations*. Princeton University Press.

Leffler, M. P. (2007). *For the soul of mankind: the United States, the Soviet Union, and the Cold War*. Macmillan.

LeoGrande, W. M. (2015). Normalizing us—cuba relations: escaping the shackles of the past. *International Affairs 91*(3), 473–488.

Levin, D. H. (2021). *Meddling in the ballot box : the causes and effects of partisan electoral interventions*.

Lieber, K. A. (2011). Mission impossible: Measuring the offense-defense balance with military net assessment. *Security Studies 20*, 456–459.

Lindsey, D. and W. Hobbs (2015). Presidential effort and international outcomes: Evidence for an executive bottleneck. *The Journal of Politics 77*, 1089–1102.

Lorentzen, P., M. T. Fravel, and J. Paine (2017). Qualitative investigation of theoretical models: the value of process tracing. *Journal of Theoretical Politics 29*, 467–491.

Maoz, Z. and B. Russett (1993a). Normative and structural causes of democratic peace. *American Political Science Review 87*, 624–638.

Maoz, Z. and B. Russett (1993b). Normative and structural causes of democratic peace, 1946–1986. *American Political Science Review 87*(3), 624–638.

Matlock, J. (2004). *Reagan and Gorbachev: How the cold war ended*. Random House.

Mattes, M. and J. L. Weeks (2022). Reacting to the olive branch: Hawks, doves, and public support for cooperation. *International Organization 76*(4), 957–976.

McManus, R. W. (2017). *Statements of Resolve: Achieving Coercive Credibility in International Conflict*. Cambridge University Press.

Miles, S. (2020). Engaging the evil empire. In *Engaging the Evil Empire*. Cornell University Press.

Montgomery, E. B. (2006). Breaking out of the security dilemma: Realism, reassurance, and the problem of uncertainty. *International Security 31*, 151–185.

Naughtie, J. (1987). Thatcher, gorbachev hopeful on arms. *The Guardian*.

Nunn, S. (1988). Arms control in the last year of the reagan administration. *Arms Control Today;(United States) 18*(2).

Parker, R. A. C. (1993). *Chamberlain and Appeasement: British Policy and the Coming of the Second World War*. New York: St. Martin's Press.

Peters, M. E. (2015). Open trade, closed borders: Immigration in the era of globalization. *World Politics 67*(1), 114–154.

Powell, R. (1996). Uncertainty, shifting power, and appeasement. *American Political Science Review 90*(4), 749–764.

Radchenko, S. (2024). *To Run the World: The Kremlin's Cold War Bid for Global Power*. Cambridge University Press.

Ramsay, K. W. (2017). Information, uncertainty, and war. *Annual Review of Political Science 20*, 505–527.

Renshon, J., A. Dafoe, and P. Huth (2018). Leader influence and reputation formation in world politics. *American Journal of Political Science 62*(2), 325–339.

Renshon, J., K. Yarhi-Milo, and J. D. Kertzer (2023). Democratic reputations in crises and war. *Journal of Politics 85*(1), 1–18.

Rhoden-Paul, A. (2024). Uk could remove syrian rebel group from terror list.

Risse-Kappen, T. (1994). Ideas do not float freely: transnational coalitions, domestic structures, and the end of the cold war. *International organization 48*(2), 185–214.

Rosato, S. (2014). The inscrutable intentions of great powers. *International Security 39*, 48–88.

Russett, B. and J. R. Oneal (2001). *Triangulating Peace: Democracy, Interdependence, and International Organizations*. W Norton and Company.

Sanger, D. E. (2021). Washington hears echoes of the ’50s and worries: Is this a cold war with china? *New York Times*.

Savranskaya, S., T. S. Blanton, and V. M. Zubok (2010). *Masterpieces of history: the peaceful end of the Cold War in Eastern Europe, 1989*. Central European University Press.

Schmitz, D. F. (1999). *Thank God they’re on our side: the United States and right-wing dictatorships, 1921-1965*. Univ of North Carolina Press.

Schmitz, D. F. (2006). *The United States and right-wing dictatorships, 1965-1989*. Cambridge University Press.

Schultz, K. A. (1999). Do democratic institutions constrain or inform? contrasting two institutional perspectives on democracy and war. *International Organization 53*(2), 233—-266.

Shultz, G. P. (2010). *Turmoil and Triumph: Diplomacy, Power, and the Victory of the American Deal*. Simon and Schuster.

Shultz, K. A. (2005). The politics of risking peace: Do hawks or doves deliver the olive branch? *International Organization 59*, 1—-38.

Spaniel, W. and B. C. Smith (2015). Sanctions, uncertainty, and leader tenure. *International Studies Quarterly 59*, 735—-749.

Taubman, W. (2017). *Gorbachev: His life and times*. Simon and Schuster.

Taylor, T. (1979). *Munich: The Price of Peace*. Garden City, NY: Doubleday.

Thompson, W. R. (2001). Identifying rivals and rivalries in world politics. *International Studies Quarterly 45*, 557–586.

Tomz, M. R. and J. L. Weeks (2013a). Public opinion and the democratic peace. *American Political Science Review 107*, 849–865.

Tomz, M. R. and J. L. Weeks (2013b). Public opinion and the democratic peace. *American political science review 107*(4), 849–865.

Tomz, M. R. and J. L. Weeks (2020). Human rights and public support for war. *The Journal of politics 82*(1), 182–194.

Vaynman, J. (2022). Better monitoring and better spying: The effects of emerging technology on cooperation.

Voeten, E. (2021). *Ideology and International Institutions*. Princeton University Press.

Waltz, K. N. (1979). *Theory of international politics*. McGraw-Hill.

Weeks, J. L. (2008). Autocratic audience costs: Regime type and signaling resolve. *International Organization 62*(1), 35—-64.

Weinberg, G. L. (2005). *Hitler's Foreign Policy 1933-1939: The Road to World War II*. New York: Enigma Books.

Wilson, J. (2014). *The triumph of improvisation: Gorbachev's adaptability, Reagan's engagement, and the end of the Cold War*. Cornell University Press.

Wolford, S. (2007). The turnover trap: New leaders, reputation, and international conflict. *American Journal of Political Science 51*(4), 772–788.

Yoder, B. K. (2019a). Hedging for better bets: Power shifts, credible signals, and preventive conflict. *Journal of Conflict Resolution 64*, 923–949.

Yoder, B. K. (2019b). Retrenchment as a screening mechanism: Power shifts, strategic withdrawal, and credible signals. *American Journal of Political Science 63*, 130–145.

Yoder, B. K. and K. Haynes (2021). Signaling under the security dilemma: An experimental analysis. *Journal of Conflict Resolution 65*, 672–700.

Yoder, B. K. and K. Haynes (2025). Endogenous preferences, credible signaling, and the security dilemma: Bridging the rationalist–constructivist divide. *American Journal of Political Science 69*, 268–283.

Zagoria, D. S. (1984). China's quiet revolution. *Foreign Affairs 62*(4), 879–904.

Appendix: The Domestic Sources of International Trust

A Formal Analysis

We now report the formal model that supports our theory. First, we set up a complete model of trust that includes all the parameters that interest us (salience and independence). We use this general setup to develop a common definition of trust-building equilibria. We also use it to characterize a set of common on-path strategies that can appear in the second period. This broad overview lays the foundation for the equilibrium analysis.

Second, we restrict our attention to the core model commonly studied, and solve for all the equilibria of the model. Third, we introduce independence. Fourth, we introduce salience. Finally, we consider a robustness model that considers how similar domestic and international issues are.

A.1 Set up.

We study a two-period trust game between two players, A, B. Where appropriate, we notate period-relevant variables as $t \in \{1, 2\}$ and player-relevant variables as $i \in \{A, B\}$. Because the model is symmetric, we sometimes we refer to player j meaning not player i.

Players come in two types: greedy i_g and security-seeking i_s. Define $y_i = pr(i_s)$ meaning that player i is a security-seeker with probability y_i, and $1 - y_i = pr(i_g)$.

The sequence of moves is as follows.

- Nature draws player types i.i.d
- A trust scenario arises in Period 1, wherein players simultaneously chose between cooperation and defection.
- Players observe the result of the trust scenario in Period 1.
- A second trust scenario arises in Period 2, wherein players simultaneously chose between cooperation and defection.
- Payoffs are realized.

A strategy for A is $s^A(a_t)$ where $a_t \in \{c, d\}$. A strategy for B is $s^B(b_t), b_t \in \{c, d\}$. A's belief at the beginning of each period about the likelihood that B is a security-seeker is σ_t^A and B's belief is σ_t^B. Trivially, $\sigma_1^i = y_j$.

Each player's payoff depends on their type, their choice, and the choice that the other side makes. Second-period payoffs are θ times the values represented in Table 1. We make two substantively motivated assumptions

$$\mathcal{A}_1 \qquad H > 1$$

This assures that players prefer mutual cooperation to mutual defection.

$$\mathcal{A}_2 \qquad H - a < 1$$

This assures the preference over outcomes assumed in the literature. That is, if a player is a security seeker, they would still prefer mutual defection to cheating. Strictly, $\mathcal{A}_2$ is not necessary for any of our main results. It does, however, simplify the conditions for equilibria that are not the focus of our analysis. We will note when it is necessary to invoke.

The first-period payoffs for A are represented in Table 3. Note A's choice is the rows, and B's choice is the columns. B's payoffs are defined similarly, replacing $\beta_A \in [0,1]$ with $\beta_B \in [0,1]$ (e.g., if B cooperates and A defects, B gets $\beta_B H$). For example, assume B is a security-seeker, then the following details B's expected utilities from the declared strategies $EU_1^{Bs}|s^B(c,d), s^A(d,d) = \beta_B H + \theta$. Here, B is suckered in the first period. However, B does not get 0. Instead, B keeps $\beta_B H$ because β_B is how independent B's action is. In the second period, B expects the double defection payoff (1), weighted by the second-period salience θ.

Note that this set-up is sufficiently flexible to capture the baseline model that represents standard models of international signaling ($\beta_i = 0, \theta = 1$); our novel introduction of payoff dependencies ($\beta_i \in [0,1], \theta = 1$); and a robustness check to our theory that explores relative salience ($\beta_i \in [0,1], \theta > 0$).

We solve for a Perfect Bayesian Equilibrium (PBE).

Given the structure of the game, off-path beliefs can emerge if and only if we conjecture strategy profiles where all types pool in the first period. We restrict off-path beliefs as follows.

Definition: Feasible off-path beliefs In any equilibrium strategy profile with an on-path pooling first period action $i_1 = c$, restrict off-path beliefs $\sigma_2^j|i_1 = d \leq y_i$. In any equilibrium strategy profile with an on-path pooling first period action $i_1 = d$, restrict off-path beliefs $\sigma_2^j|i_1 = c \geq y_i$.

This restriction states that if all types of player i cooperate on path in the first period, then if i deviates to defect, j cannot increase her confidence that i is a security seeker. Similarly, if all types of player i defect on path in the first period, then if i deviates to cooperate, j cannot increase her confidence that i is greedy.

A.1.1 Definition: trust-building equilibria

We define a trust-building equilibria as a PBE with the following features:

- First-period discriminators: The greedy and security-seeking types face different incentives, and therefore make different choices in the first period. In equilibrium: $pr(a_1 = c|A_s) > pr(a_1 = c|A_g)$.
- Cooperation generates trust: Because greedy and security-seeking types make different choices in the first period, their rivals learn, and this allows first-period cooperation to engender trust. In equilibrium: $\sigma_{1A} < \sigma_{2A}|b_1 = c$.
- Trust breeds cooperation: The trust built on the first-period choice allows states to cooperate in the second period. Therefore, the inclusion of an initial period facilitates trust, that allows for cooperation in the second period. In equilibrium: $pr(a_2 = c|A_s, a_1 = b_1 = c) > pr(a_2 = c|A_s, b_1 = d, a_1 = .)$.

A.1.2 Preliminary analysis that serves all equilibrium analysis

We start by solving for all strategies that can appear on the path in the second period.

First we must define a mixing probability. Let $\omega_B^* = pr(b_2 = c) = \frac{1}{\sigma_2^A(1+a)}$, and $\omega_A^* = pr(a_2 = c) = \frac{1}{\sigma_2^B(1+a)}$.

Lemma A.1 *For some set of parameters, we can support three and only three strategy profiles on path in the second period of a PBE.*

1. *For all parameters we can support* $a_2 = b_2 = d$.
2. *If* $\sigma_2^i > \frac{1}{1+a}$ *we can support* $a_2 = c|A_s, a_2 = d|A_g$, $b_2 = c|B_s, b_2 = d|B_g$.
3. *If* $\sigma_2^i > \frac{1}{a+1}$ *we can support* $a_2 = \omega_A^*|A_s, a_2 = d|A_g$, $b_2 = \omega_B^*|B_s, b_2 = d|B_g$.

Before we consider a specific profile, note that greedy types defect in every on-path strategy profile. We now show this is strictly true. Conjecture $b_2 = a_2 = c$ in equilibrium. A_g can profitably deviate to defect if $H + a > H$. Conjecture $a_2 = c, b_2 = d$. A_g can profitably deviate to defect if $1 > 0$.

We now describe three strategy profiles we can support on the path, focusing on the security-seeker's preferences. We derive Bullet 1 as follows. A_s prefers to remain on the path (defect) rather than deviate to cooperate if $1 > 0$. Always true, as desired.

We derive bullet 2 as follows. Consider A_s prefers to remain on the path, rather than deviate to defection, so long as $\sigma_2^A H > \sigma_2^A(H - a) + 1 - \sigma_2^A$, this solves for the equilibrium condition with regard to σ_2^A. σ_2^B is solved the same way.

To verify these are the only supportable pure strategies, we must exhaust the other strategy profiles for security seekers. Consider the pure strategy profile: $a_2 = c, b_2 = d$. Note A can always profitably deviate to $a_2 = d$. There are no other pure strategy profiles to consider.

Turning to the mixed strategy profile. A is indifferent between cooperation and defection if $\sigma_2^A \omega_B H = \sigma_2^A(\omega(H - a) + 1 - \omega) + 1 - \sigma_2^A$. This solves for $\omega_B = \frac{1}{\sigma_2^A(1+a)}$. This can be solved within 0, 1 so long as $\sigma_2^A > 1/(a + 1)$. A's equilibrium mixing probability is solved similarly.

To verify this is the only supportable mixed strategy, we must exhaust the other strategy profiles for security seekers. Consider the pure strategy profile: $pr(a_2 = c) = \omega'_A \neq \omega_A^*$, $pr(b_2 = c) \in (0, 1)$. Note, B is only indifferent at $\omega'_A = \omega_A^*$, and thus B must hold a profitable deviation to $pr(b_2 = c) \in \{0, 1\}$. Now consider $pr(a_2 = c) = \omega'_A \neq \omega_A^*$, $b_2 = c$. Here, if A is a security seeker, A always holds a profitable deviation to $a_2 = c$. There are no other mixed strategies to consider.

Remark These three on-path strategy profiles yield the following expected utility for A at the onset of the second period.

1. $EU_2^A = 1$
2. $EU_2^A|A_s = H\sigma_A$, $EU_2^A|A_g = \sigma_A(H + a - 1) + 1$
3. $U_2^A|A_s = \frac{H}{1+a}$, $U_2^A|A_g = \frac{H+a-1}{1+a} + 1$

These utilities are useful for characterizing equilibria because they define the possible expected second-period utilities. For example, in an equilibrium where security-seekers cooperate in the first period and greedy types defect, we are certain that first-period defection yields an expected utility of 1 in the second period. We are also certain that first-period cooperation can generate only 1 of 3 potential expected utilities.

Finally, we emphasize a point that we will later use to justify our focus on pure strategy equilibria in the manuscript.

Remark Holding beliefs constant in ranges we can support the respective equilibria ($\sigma_2^i > \frac{1}{a+1}$), both greedy types and security-seekers hold a higher expected utility from the pure strategy equilibrium described in bullet 2, than the mixed strategy equilibrium described in bullet 3.

Note, all expected utilities are increasing in σ_2^i. Subbing in the boundary condition $\sigma_2^i = \frac{1}{a+1}$, A_s, A_g's utilities are identical in bullet's 2 and 3.

A.2 Baseline Model of International Signaling (Section 1)

We start with the assumption that $\beta_i = 0, \theta = 1$. This is the set of assumptions discussed in our review of existing trust-building theory in Section 1.

We start with the pure strategy equilibria that we plot in Figure 1 that are not trust-building equilibria.

Proposition A.2 *We can support the following strategy profiles as PBE for any feasible off-path beliefs.*

1. ***Tragic:*** *For all parameters* $s^A(d, d), s^B(d, d)$.

2. ***Suckers:*** *If* $y_i > \frac{a}{H+a-1} \sim \frac{1}{1+a}$ *holds we can support* $s^A(a_1 = c, a_2 = c|(A_s\&a_1 = b_1 = 1), a_2 = d|$ ***Otherwise***$), s^B(b_1 = c, b_2 = c|(B_s\&a_1 = b_1 = 1), b_2 = d|$ ***Otherwise***$)$.

3. ***No Learning:*** *Given* $\mathcal{A}_2$, *if* $y_i > \frac{1}{1+a}$ *holds we can support* $s^A(a_1 = d, a_2 = c|A_s, a_2 = d|A_g), s^B(b_1 = d, b_2 = c|B_s, b_2 = d|B_g)$.

4. ***Semi-Tragic:*** *If* $y_i > \frac{1}{1+a}$ *holds we can support* $s^A(a_1 = c|A_s, a_1 = d|A_g, a_2 = d), s^B(b_1 = c|B_s, b_1 = d|B_g, b_2 = d)$.

The tragic equilibrium is obvious. We now analyze the **suckers** equilibrium. We start with A_g's strategy. All players cooperate in the first period. So in period 1, A_g prefers cooperation to defection if: $y_B(H + H + a) + (1 - y_B)(H + 1) > y_B(H + a + 1) + (1 - y_B)(H + a + 1)$. This solves for $y_i > \frac{a}{H+a-1}$, as desired. Turning to A_s's strategy. In the second period, we assert A_s prefers cooperation to defection given on path play. Given all players cooperate in period 1, $\sigma_2^A = y_B$. Thus, A prefers cooperation to defection if, $y_B H > y_B(H - a) + 1 - y_B$. This rearranges to $y_i > \frac{1}{1+a}$. Working backwards, in the first period, A_s prefers cooperation over deviating to defection if: $y2H + (1 - y)H > y(H - a + 1) + (1 - y)2$. This re-arranges to $y > \frac{2-H}{1+a}$. Taking $H = 1$ as the minimum bound, this is equivalent to $y_i > \frac{1}{1+a}$. Note we can sustain the equilibrium for any off-path beliefs. The reason is that the only-path action is $i_1 = d$ and we conjecture all players revert to $i_2 = d$ given either side deviates. We've shown we can support $i_2 = d$ for any beliefs, as desired. B's strategy is symmetric. This completes the proof.

We now analyze the **No Learning** equilibrium. In it, $a_1 = b_1 = d$. Thus, A's on-path belief is $\sigma_{2A} = y_B$. Lemma A.1.2 solves the second-period strategy. This gives is the condition $y_i > \frac{1}{1+a}$. Turning to the first period, notice the conjectured on-path second-period strategies leave all types with their maximum second-period expected value. Thus, we'll focus on the case where off-path beliefs match on-path beliefs. Given this case, no type can profit by deviating so long as $1 > 0$. B's strategy is symmetric. This completes the proof.

We note a second no learning equilibrium held together by a different off-path punishment: $s^A(a_1 = d, a_2 = c|A_s \;\&\; a_1 = b_1 = d, a_2 = d|\text{ otherwise}), s^B(b_1 = d, b_2 = c|B_s \;\&\; a_1 = b_1 = d, b_2 = d|\text{ otherwise})$. Focusing on A_s, there is no profitable deviation if $1 + y_B(H + a) > H - a + 1 \equiv y_B > \frac{H-a}{H+a}$. Note, $\frac{1}{1+a} > \frac{H-a}{H+a}$ if $2 > H - a$ true by assumption $\mathcal{A}_2$. If we violate $\mathcal{A}_2$, we simply require an additional condition that places $y > \frac{H-a}{H+a}$, but do not alter any of our substantive conclusions.

We now analyze the **Semi-Tragic** equilibrium. Lemma A.1.1 shows that we can support $a_2 = b_2 = d$ for all beliefs and parameters. Turning to the first period, A_g prefers defection if $y_B(H + a + 1) + (1 - y_B)2 > y_B(H + 1) + 1 - y_B \implies y_B(a - 1) > -1$, always true. A_s prefers cooperation if $y_B(H+1) + 1 - y_B > y_B(H - a + 1) + 2(1 - y_B) \implies y_B > \frac{1}{a+1}$, as desired. B's strategy is symmetric.

We now solve for the pure strategy, trust-building equilibrium.

Proposition A.3 *There is one pure strategy **trust-building equilibrium**. It arises if and only if*

$$\frac{1}{H} \geq y_i \geq \frac{1}{H + a} \tag{1}$$

holds. In it, greedy A plays $s^A(d, d)$. security-seeker A plays $s^A(a_1 = c, a_2 = c|(b_1 = c, a_1 = c), a_2 = d$ otherwise$)$. B's strategy is symmetric.

Since it is a pure strategy equilibrium, $\sigma_{2A}|b_1 = c = 1$.

The greedy A plays on the path if:

$$y_B(H + a + 1) + (1 - y_B)2 \geq y_B(2H + a) + 1 - y_B \equiv y_B \leq \frac{1}{H}$$

The security A plays on path if:

$$y_B(H + a + H) + 1 - y_B \geq y_B(H - a + 1) + (1 - y_B)2 \equiv y_B \geq \frac{1}{H + a}$$

These re-arrange as the upper and lower boundaries of our equilibrium conditions. This completes the proof.

Proposition A.4 *Given feasible off-path beliefs no other pure strategy equilibria exist.*

We've shown that we can only support two pure strategy profiles in the second period, and that greedy types always defect in the second period. As a result, there are only two cases to consider. First, there are a class of strategy profiles that include the following unconditional on path actions:

$s^A(a_1 = a_2 = c|A_s), s^B(b_1 = b_2 = c|B_s)$. We cannot support any equilibrium that includes this in the strategy profile. Suppose we could, it is easy to see that $s^A(a_1 = a_2 = d|A_g), s^B(b_1 = b_2 = d|B_g)$. This implies that both states must form posterior beliefs $\sigma_2^i = 0|i_1 = d$. This implies security seekers can profitably deviate from $i_2 = c \to d$.

Second, while we have ruled out asymmetric strategies in the second period, it is theoretically possible that we can support asymmetric strategies in the first, so long as second period strategies are conditional on first period. There are only two profiles to rule out, that vary in their off-path punishments. Profile 1 is: $s^A(a_1 = c|A_S, a_1 = d|A_g, b_2 = c|A_s \& a_1 = c \& b_1 = d, a_2 = d \text{ otherwise})$, $s^B(b_1 = d, b_2 = c|B_s \& a_1 = c \& b_1 = d, b_2 = d \text{ otherwise})$. Here if B deviates from $b_1 = d \to c$, then players revert to $a_2 = b_2 = d$, which we can always support. In the first period, A_s cannot profitably deviate from $a_1 = c \to d$ if: $y_B H > 1 + 1$, solves for $y_B > \frac{2}{H}$. A_g cannot profitably deviate from $a_1 = d \to c$ if: $1 + 1 > y_B(H + a) + 1 - y_B$, solves for $\frac{1}{H+a-1} > y_B$ as desired. These can only be jointly satisfied if $2 > H + 2a$, cannot be true if $\mathcal{A}_2$ holds.

Profile 2 is: $s^A(a_1 = d, a_2 = c|A_s \& b_1 = c, a_2 = d \text{ otherwise})$, and $s^B(b_1 = c|B_S, b_1 = d|B_g, b_2 = c|B_s \& b_1 = c \& a_1 = ., b_2 = d \text{ otherwise})$. Here $a_1 = .$, emphasizes that b_2 holds even given A's off-path deviation. The ICC on B_s's first period choice is: $0 + y_A(H) + (1 - y_A)0 > 2 \equiv y_A > 2/H$. Only solvable if $H > 2$. The ICC on B_g's first period choice is $0 + y_A(H + a) + (1 - y_A) < 2 \equiv y_A < \frac{1}{H+a-1}$. Putting these together, it must be that $2/H < \frac{1}{H+a-1} \& H > 2 \equiv H < 2 - 2a \& H > 2 \equiv 2a < 0$, which cannot be satisfied.

A.2.1 A comment on mixed strategy equilibria

We focus on pure strategy PBE in the manuscript. To be clear, there are many mixed strategy equilibrium and even mixed strategy trust-building equilibria. We omit them from the manuscript because (a) they do not alter our basic conclusions; (b) are cumbersome to solve for and cannot be easily grouped owing to many different off-path strategies that can emerge; and (c) are strictly less efficient than pure strategy equilibria. In section A.3.5 we will fully specify all the mixed strategy equilibria for the complete model (note this model is a special case where $\beta = 0$). Here we provide a preliminary analysis to demonstrate why they are both inefficient and also cannot alter our results. Recall, our main claim is that trust-building equilibria do not arise when y_i are too low. Thus, it would be misleading to omit mixed strategy trust-building equilibria if we could support them at lower levels of y_i. We shall solve for these in the complete model (i.e, once we introduce $\beta_i \in [0, 1]$). But here we demonstrate that they will not influence our core conclusion.

We cannot support them for lower levels of y_i. The reason is that the lower bound on y_i to support the contingent equilibrium is defined by the security-seeker's preference to engage in trust-building. When $y_B < \frac{1}{H+a}$, A_s's expected value from cooperation is too low to support trust, given her expectation that B is greedy. We've already shown that the mixed strategies we can support in the second period lower the security-seeker's expected utility. This means that the minimum y_i that will support trust-building must be high. To illustrate the point, we solve for the mixed strategy equilibrium that supports trust with the lowest level of y_i.

Proposition A.5 *If*

$$\frac{1+a}{2(H+a-1) - a(1+a)} > y_i > \frac{1+a}{2H + a(1+a)} \tag{2}$$

Then the following strategies are a mixed strategy, trust-building PBE. Greedy A plays $s^A(d, d)$. Security-seeker A plays $s^A(a_1 = c, a_2 = \omega^*_A | b_1 = c, a_1 = c; a_2 = d$ otherwise). B's strategy is symmetric.

Because the first period separates, $\sigma_{2A} = 1 | b_1 = c$, $\sigma_{2A} = 0 | b_1 = d$. The security-seeker remains on the path if, $y_B(H + \frac{2H}{1+a}) + 1 - y_B > y_B(H - a + 1) + 2(1 - y_B)$. This solves for the lower bound on y_i, as stated in the equilibrium.

The greedy type remains on the path if, $y_B(H + a + 1) + 2(1 - y_B) > y_B(H + \frac{2(H+a-1)}{1+a}$. This solves for the upper bound on y_i, as stated in the equilibrium.

Contrasting the lower bounds of inequalities 2 and 1, we can support 1 at lower levels of y_i if $\frac{1+a}{2H+a(1+a)} > \frac{1}{H+a}$, always true.

A.3 Our theory: Independence of domestic choices (Section 2.1, and Results 1a and 1b)

We now study our theoretical intervention by only changing the model above to allow $\beta \in [0, 1]$.

We proceed as follows. First, we solve for the trust-building pure strategy PBE. Since our formally stated results focus on this equilibrium, we detail the comparative statics of this equilibrium through a series of remarks, and clarify how the results map onto Results 1a and 1b. Second, we solve for all other pure strategy PBE and rule out those that we cannot support. Finally, we solve for all mixed strategy equilibria and rule out those we cannot support.

A.3.1 The pure strategy, symmetric trust-building equilibrium

Proposition A.6 *If*

$$y_i > \frac{1 - \beta_j(a + 1)}{H + a - B_A(a + 1)} \tag{3}$$

and

$$\frac{1 + \beta_j(a - 1)}{H + (a - 1)\beta_j} > y_i \tag{4}$$

hold, then there is a pure strategy trust-building equilibrium with the same strategy profile as written in Proposition A.3.

We start with A_s's strategy. Because this is a complete separating equilibrium, $\sigma^A_2 | b_1 = d = 0, \sigma^A_2 | b_1 = c = 1$. By Lemma A.1, we can support second-period cooperation. Turning to the first period, A_s prefers to cooperate, rather than defect if:

$$y_B(H + H) + (1 - y_B)(\beta_A H + 1) > y_B(H - a + 1) + (1 - y_B)(\beta_A(H - a) + 1 - \beta_A + 1)$$

This assumes that if A defects, A gets the second-period value 1 from $b_2 = a_2 = d$. This re-arranges to equilibrium condition 3, as desired.

Turning to A_g's strategy. We've already shown we can support second-period defection for any set of beliefs. Turning to the first period, A_g prefers to defect, rather than cooperate if:

$$y_B(H+a+1)+(1-y_B)(\beta_A(H+a)+1-\beta_A+1)<y_B(H+(H+a))+(1-y_B)(\beta_A H+)$$

$$y_B(H+(a-1)\beta_A)<1+\beta_A(a-1)$$

This rearranges to inequality 4 as desired. There are no off-path beliefs. The strategies are symmetric. This completes the proof.

A.3.2 Establishing Result 1a,1b

Result 1a: When both players' choices are sufficiently independent (i.e., $\beta_A, \beta_B > \frac{1}{1+a}$), a trust-building equilibrium always exists for states that start out with the highest possible level of confidence that the other is greedy (i.e., $y \to 0$).

Result 1b: Even when the independence threshold characterized in 1a is not met, as the level of independence increases, a trust-building equilibrium can be supported at decreasing levels of initial trust.

Results 1a and b and their implications are effectively a series of comparative static claims on equilibrium A.6 as a function of β, y.

Thus, we focus on the ICC for greedy and security types. Our main claims focuses on the incentives of security-seekers. The reason we care most about security-seekers is that their incentives impose a lower bound on y (condition 3). The classic model (absent β) is structured such that the security-seeker desires mutual cooperation when payoffs are dependent. Thus, if initial trust is too low, the security-seeker does not cooperate. Thus all of our claims about independence and trust relate to easing A_s's tension that prevents cooperation when y is too low.

Remark The security-seeker cannot profitably deviate from on-path cooperation in the trust-building equilibrium at any level of initial trust (even $y \to 0$), and any value of H so long as

$$\beta_i > \frac{1}{a+1}.$$

Outside this range, the level of independence that will sustain the security-seekers' incentive compatibility constraint is

$$\beta_i > \frac{1-y_j(H+a)}{(a+1)(1-y_j)}$$

wherein the right hand-side is strictly increasing in y_j.[28]

Both results come from re-arranging the security-seeker's ICC described in 3. The first claim establishes the boundary where initial trust does not affect A_s's incentives for trust-building. Consider the limit of 3, if $\beta = 1$, it solves for: $\frac{-a}{(H-1)}$. This is strictly negative for any permissible value of H, a. Re-arranging this condition, we see that the condition on y remains negative (and thus we

[28] recall, subscript j means not i.

can support trust-building for any level of initial trust), if $\beta_i > \frac{1}{a+1}$, as desired. The second claim comes from simply re-arranging 3 as a function of β. It is useful because it illustrates that A_s's incentives for cooperation are increasing in independence.

Turning to the incentives of greedy states. The classic model is structured such that if initial trust is too high, that greedy will try to cheat. Their incentive to deviate to cooperation is amplified when they believe the other side can be cheated (i.e., they are playing against a security-seeker). Thus, their ICC is governed by an upper bound on y. This is condition 4. To be clear, this is less important for our theory. After all, our main claim is that there is no lower bound on initial trust. But greedy types only determine the upper bound. Thus, our main goal is to establish that the greedy types are willing to comply under the same conditions that security-seekers are.

Remark The greedy type cannot profitably deviate from first-period defection in the trust-building equilibrium so long as $\beta_j > \frac{y_i(H-1)}{(a-1)(1-y_i)}$.

This simply comes from re-arranging 4. Notice two things. First, the condition is easier to satisfy at higher levels of β. Second, when $\beta = 1$, the LHS becomes $\frac{a}{(H+a-1)}$, which is strictly greater than 0 for any permissible value of H, a, as desired.

Putting both type's of incentives together, notice that

Remark For any set of parameters H, a, (a) all types' incentives to deviate from on path actions in the trust-building equilibria are decreasing in β; (b) at $\beta = 1$, we can support a trust-building equilibrium for every level of initial trust that satisfies $y_i < \frac{a}{(H+a-1)}$.

A.3.3 Asymmetric trust build equilibria

When $\beta_i = 0$, we ruled out the possibility of asymmetric pure strategy equilibria entirely. Here we demonstrate that they are viable with payoff independence.

Proposition A.7 *If*

$$\frac{\beta_A(a-1)}{H+a-1} > y_B > \frac{2-\beta_A(1+a)}{H}$$

$$y_A > \frac{\beta_B(a-1)+1}{\beta_B(a-1)+2-H+a}$$

then, there is an asymmetric, pure strategy, trust building equilibrium: $s^A(a_1 = c|A_S, a_1 = d|A_g, b_2 = c|A_s\&a_1 = c\&b_1 = d, a_2 = d$ *otherwise*$)$, $s^B(b_1 = d, b_2 = c|B_s\&a_1 = c\&b_1 = d, b_2 = d$ *otherwise*$)$. *There is an equivalent trust-equilibrium swapping, A and B.*

A's first period choice is fully separating, and thus we can sustain B's second period strategy. B's first period choice is pooling. Thus, we can sustain A's second period choice if $y_B > \frac{1}{a+1}$.

In the first period, A_s cannot profitably deviate from $a_1 = c \to d$ if: $\beta_A H + y_B H > \beta_A(H - a) + 1 - \beta_A + 1$, solves for $y_B > \frac{2-\beta_A(1+a)}{H}$, as desired. What is more, $\frac{2-\beta_A(1+a)}{H} > \frac{1}{a+1}$ given $\mathcal{A}_2$. A_g cannot profitably deviate from $a_1 = d \to c$ if: $\beta_A(H+a)+1 > \beta_A H + y_B(H+a) + 1 - y_B$, solves for $\frac{\beta_A(a-1)}{H+a-1} > y_B$ as desired.

In the first period, B_s cannot profitably deviate from $b_1 = d \to c$ if: $y_A(H - a + H) + (1 - y_A)(\beta_B(H+a)+1-\beta_B+1) > y_A(H+1)+(1-y_A)(\beta_B H+1)$, solves for $y_A > \frac{\beta_B(a-1)+1}{\beta_B(a-1)+2-H+a}$, as desired. B_g cannot profitably deviate from $b_1 = d \to c$ if $y_A(H+a+H+a)+(1-y_A)(\beta_B(H+a)+1-\beta_B+1) > y_A(H+1)+(1-y_A)(\beta_B H+1)$, which is always true.

Before moving on, we note two facts.

Remark (a) Asymmetric trust building requires some dependence $\beta_A > 0$. (b) it is easier to satisfy all three constraints as either dependency increases ($\beta_A \to 1, \beta_B \to 1$).

A.3.4 Other pure strategy equilibria

In what follows, we characterize all other pure strategy equilibria. To begin, we focus on the symmetric equilibria. Since the symmetric equilibria follow naturally from the equilibrium listed in proposition A.2, we only solve for the conditions where the addition of β makes a difference.

Proposition A.8 *There is a* ***tragic equilibrium*** *if and only if $\beta_i < \frac{1}{1+a}$. It has the same strategy profile as in proposition A.2.1.*

We focus on A_s's strategy. We've shown we can support mutual defection in the second period for any set of beliefs and parameters. We label A's expected value of $a_2 = b_2 = d$ as EU_2. We focus on first-period incentives. A_s prefers defection to cooperation if $\beta_A H + EU_2 < \beta_A(H - a) + 1 - \beta_A + EU_2$. This solves for the equilibrium condition as desired. This result is notable because it departs from the baseline model, and conventional wisdom that defect, defect is always an equilibrium.

Proposition A.9 *There is a* ***suckers equilibrium****. Its conditions and strategy profile are the same as in proposition A.2.2.*

See proof of A.2.2. Since all players cooperate in the first period, β does not factor into computing the conditions where deviating is profitable.

Proposition A.10 *There is a* ***No learning*** *PBE if $y_i > \frac{1}{1+a}$ and $\beta_i < \frac{1}{1+a}$. It has the same strategy profile as in proposition A.2.3.*

The only difference in the proof from A.2.3 is in A_s's incentive to play defect in the first period. A_s prefers defect to cooperate iff: $\beta_A(H - a) + 1 - \beta_A + y_B H > \beta_A H + y_B H$. This solves for $\beta_A < \frac{1}{1+a}$, as desired. This completes the proof.

As in the baseline, there is a second **No learning** PBE held together by a different off-path action. Specifically,

Proposition A.11 *If $y_i > \frac{\beta_j(1+a)}{H}$ and $\beta_i < \frac{1}{1+a}$. There is a second* ***No learning*** *PBE with strategy profile. $s^A(a_1 = d, a_2 = c|A_s \;\&\; a_1 = b_1 = d, a_2 = d|$ otherwise$)$, $s^B(b_1 = d, b_2 = c|B_s \;\&\; a_1 = b_1 = d, b_2 = d|$ otherwise$)$.*

Here the difference is that if either player deviates from $i_1 = d \to c$, then $i_2 = d$. The first period ICC for A_s is: $\beta_A(H - a) + 1 - \beta_A + y_B H > \beta_A H + 1$ This solves for the new condition.

Comparing these equilibria, $\frac{\beta_j(1+a)}{H} > \frac{1}{1+a} \equiv \frac{\beta_j(1+a)}{H} > \frac{1}{a+1}$. Subbing in the second condition that $\beta_j < \frac{1}{1+a} \equiv 1 > H - a$. Thus, it must be that the first No learning equilibrium occurs under broader conditions than the second if $\mathbf{A}_2$ holds.

Proposition A.12 *There is a **Semi-tragic** PBE if and only if $\beta_i > \frac{1-y(1+a)}{(1+a)(1-y)}$. It has the same strategy profile as in proposition A.2.4.*

The only difference in the proof from A.2.4 is in A_s's incentive to play cooperate in the first period. A_s prefers cooperate to defect iff: $y_B H + (1-y_B)\beta_A H + 1 > y_B(H-a) + (1-y_B)(\beta_A(H-a) + 1 - \beta_A) + 1$. This solves for $\beta_A > \frac{1-y_B(1+a)}{(1+a)(1-y_B)}$ as stated. B's incentives are symmetric. This completes the proof. For interest, the condition can be re-written as, $y > \frac{1-\beta-\beta a}{(1-\beta)(1+a)}$.

Finally, we turn to the asymmetric equilibria. We state one side. There is an equivelent equilibrium swapping the As and Bs.

Proposition A.13 *If $\beta_A > \frac{1}{H}$, and $y_A < \frac{1-\beta_B(1+a)}{(1-\beta_B)(1+a)}$ there is an **asymmetric, semi-tragic** PBE. In it, $s^A(a_1 = c|A_s, a_1 = d|A_g, a_2 = d), s^B(b_1 = d, b_2 = d)$.*

Trivially, A_g cannot profit from deviating. A_s cannot deviate from $a_1 = c \rightarrow d|b_1 = d$ if, $\beta_A H > 1 \equiv \beta_A > \frac{1}{H}$, as desired. Trivially, B_g cannot profit from deviating. B_s cannot profit from deviating from $b_1 = d \rightarrow c$ if: $y_A(H-a) + (1-y_A)(\beta_B(H-a)1 - \beta_B) > y_A H + (1-y_A)\beta_B H \equiv y_A < \frac{1-\beta_B(1+a)}{(1-\beta_B)(1+a)}$.

Ruling out other pure strategy equilibria Finally, we rule out three other classes of pure strategy equilibria. First, we cannot sustain any pure strategy equilibria that include against-type first period actions. That is $a_1 = c|A_g, a_1 = d|A_s$. It is trivial that we cannot support these if second period strategies are not contingent on first period actions. The reason is that if second period actions are not contingent, then we need only consider first period incentives. In terms of contingent second period strategies, the binding constraint is: $s^A(a_1 = c|A_g, a_1 = d|A_s, a_2 = c|(a_1 = b_1 = d, A_s), a_2 = d$ otherwise$)$. Note that we need not consider other second period contingent strategies because given the first period strategy, $\sigma_i^2 = 1|j_1 = d, \sigma_i^2 = 0|j_1 = c$. Given this strategy profile, A_g's ICC is: $y_B(H\beta_A+1) + (1-y_B)(H+1) > y_B(\beta_A(H+a) + 1 - \beta_A + H + a) + (1-y_B)(H+a+1)$. This reduces to, $0 > y_B(\beta_A a - \beta_A + H) + a$, which cannot be satisfied. It follows that for any contingent second period strategy profile we could support, A_g always has a profitable first period deviation from $a_1 = c \rightarrow d$.

Second, we cannot support any pure strategy PBE that include contingent second period strategies $s^A(a_2 = c|b_1 = d, A_s), a_2 = d|b_1 = c, A_s)$. This follows instantly from what was just shown.

Finally, we rule out other pure strategy asymmetric equilibria. Given what we just ruled out, the only remaining asymmetric equilibria to exhaust are those that include $a_2 = b_2 = d$. Note, we cannot support any equilibria that includes $i_1 = c|j_2 = i_2 = dA_g$. Suppose we could, i_g can always profit from the deviation $i_1 = c \rightarrow d$. Thus, we need only consider $i_1 = i_2 = d|A_g$. It follows that the only remaining asymmetric strategy profile to rule out is: $s^A(a_1 = c|A_s, a_1 = d|A_g, a_2 = d), s^B(b_1 = d, b_2 = d)$. We've solved for this profile.

A.3.5 Symmetric Mixed strategy equilibria

As reasoned above, none of our main conclusions are impacted by mixed strategy equilibria. Here we solve them for completeness and emphasize how the results are bounded by both initial trust y and dependenceβ. We will also emphasize which mixed strategy equilibria meet our definition of trust building. We'll show that no mixed strategy trust building equilibria arise given lower values of y_i than the pure strategy equilibria above.

Mixing in second period only There are two equilibria with only second period mixed strategies (i.e, pure strategies in first period).

As a reminder, we have solved for the unique, on path mixing strategy:

$$\omega_i^* = pr(a_2 = c) = \frac{1}{\sigma_2^j(1+a)}$$

Which produced second period expected utilities:

$$U_2^A|A_s = \frac{H}{1+a} \qquad U_2^A|A_g = \frac{H+a-1}{1+a} + 1$$

and always carried an equilibrium condition $\sigma_2^i > \frac{1}{a+1}$.

First,

Lemma A.14 *If*

$$y_i > \frac{1}{a+1}$$
$$\beta_A < \frac{H}{(1+a)^2}$$

then the following strategy profile is an equilibrium $s^A(a_1 = d, a_2 = \omega_A^*|(a_1 = b_1 = d\&A_s), a_2 = d$ *otherwise*$).s^B(b_1 = d, b_2 = \omega_B^*|(a_1 = b_2 = d\&B_s), b_2 = d$ *otherwise*$)$ *given any off-path beliefs.*

Given first period pooling, $\sigma_2^i = y_j$. This gives us the first condition. In the first period, if either player deviates from defection, the game reverts to mutual defection in the next period. Thus, A_g's ICC is: $\beta_A(H+a) + 1 - \beta_A + \frac{H+a-1}{1+a} + 1 > \beta_A H + 1$, which is always satisfied, as desired. Further, A_s's ICC is: $\beta_A(H-a) + 1 - \beta_A + \frac{H}{1+a} > \beta_A H + 1$, which solves for $\beta_A < \frac{H}{(1+a)^2}$, as desired. Note as we saw in the pure strategy equilibria, we cannot support mutual defection in the first period when $\beta_A \to 1$. This places an upper bound on dependence for this equilibrium. Further, this condition is strictly satisfied when $\beta_A = 0$.

Second,

Lemma A.15 *If*

$$y_i > \frac{1}{a+1}$$
$$H > a^2 + 1$$

then the following strategy profile is an equilibrium $s^A(a_1 = c, a_2 = \omega_A^*|(a_1 = b_1 = c\&A_s), a_2 = d$ *otherwise*$).s^B(b_1 = c, b_2 = \omega_B^*|(a_1 = b_2 = c\&B_s), b_2 = d$ *otherwise*$)$ *given any off-path beliefs.*

Given first period pooling, $\sigma_2^i = y_j$. This gives first condition 1. A_s ICC is $H + \frac{H}{a+1} > H - a + 1$, always satisfied. A_g's ICC is $H + \frac{H+a-1}{1+a} + 1 > H + a + 1$, which gives us $H > a^2 + 1$.

Finally, we **cannot support** any equilibrium where each type plays a separating pure strategy in the first period, and mixes in the second. If they separate in period 1, then $\sigma_2^i \in \{0, 1\}$, which assures we cannot support mixing in the second period. We've shown $i_2 = d|i_g$ strictly dominates. This implies we cannot support any equilibria where i_g mixes in the second period.

Mixing only in the first period. There are three equilibria with only first period mixed strategies (i.e, pure, possibly conditioned, strategies in second period).

Define a first period mixing probability:

$$\omega_i^x = \frac{1-\beta_j(1+a)}{y_i(1+a)(1-\beta_j)}$$

We'll prove that this leaves i_s indifferent in two equilibria. Note that for $\omega_i^x \in [0,1]$, it must be that $y_i > \frac{1-\beta_j(1+a)}{(1+a)(1-\beta_j)}$, with the special case $y_i > \frac{1}{(1+a)}$ given full dependencies. It also requires $\frac{1}{1+a} > \beta_j$.

The first equilibrium is:

Lemma A.16 *If $\omega_i^x \in [0,1]$, then the following strategy profile is an equilibrium $s^A(pr(a_1 = c) = \omega_A^x | A_s, a_1 = d|A_g, a_2 = d).s^B(pr(b_1 = c) = \omega_B^x | B_s, b_1 = d|A_g, b_2 = d)$.*

In period 2, players mutually defect regardless of type. This is proven to hold. First, we derive the mixing probability ω_i^x as what leaves i_s indifferent between cooperation and defection. Focusing on A_s, $\omega_B y_B H + (1-\omega_B y_B)\beta_A H + 1 = \omega_B y_B(H-a) + (1-\omega_B y_B)(\beta_A(H-a) + 1 - \beta_A) + 1$, which solves for ω_A^x. Trivially, if we can find a $\omega_i^x \in [0,1]$ than, i_s can be held indifferent.

The second is,

Lemma A.17 *If $\omega_i^x \in [0,1]$, and*

$$y_i > \frac{a + (1+a)(1-\beta_j(1+a))}{(1+a)(1-\beta_j)}$$

then the following strategy profile is an equilibrium $s^A(pr(a_1 = c) = \omega_A^x | A_s, a_1 = d|A_g, a_2 = c|A_s, a_2 = d|A_g,).s^B(pr(b_1 = c) = \omega_B^x | B_s, b_1 = d|A_g, b_2 = c|B_s, b_2 = d|A_g)$.

In the second period, $\sigma_2^A | b_1 = c = 1$. Clearly, we can sustain $a_2 = c|A_s$ in this case. $\sigma_2^A | b_1 = d = \frac{y_B(1-\omega_B^z)}{y_B(1-\omega_B^z)+1-y_B}$. Thus, to sustain second period choices, it must be that $\frac{y_B(1-\omega_B^x)}{1-y_B\omega_B^x} > 1/(a+1)$, or

$$\omega_B^x < \frac{y_B(a+1)-1}{y_B a}$$

. Plugging in the value of ω_i^x gives us the equilibrium condition. There are no off-path actions in the first period, and thus no second period reversion is necessary.

Since second period strategies are not conditions, in the first period, the binding constraint is inducing i_S to mix. Because we conjecture second period strategies are not conditioned in first period strategies, it follows instantly that the $\omega_i^1 = \omega_i^x$ leaves player's indifferent. Finally, $\frac{1-\beta_j(1+a)}{(1+a)(1-\beta_j)} < \frac{a+(1+a)(1-\beta_j(1+a))}{(1+a)(1-\beta_j)}$, which implies we need not specify this condition. There are no off-path actions in the first period, and thus no second period reversion is necessary.

We now solve for a **trust building** mixing equilibria. It differs from the above in that it includes a contingent second period strategy. Define a first period mixing probability:

$$\omega_i^z = \frac{1-\beta_j(1+a)}{y_i(H+a-\beta_j(1+a))}$$

We'll prove that this leaves i_s indifferent given contingent second period strategies. Note that for $\omega_i^z \in [0,1]$, it must be that $\frac{1}{1+a} > \beta_j$. Further, $y_i > \frac{1-\beta_j(1+a)}{(H+a-\beta_j(1+a))}$, and in the baseline case, $y_i > \frac{1}{(H+a)}$, which is consistent with our claim that there is a low bound on trust to sustain cooperation absent any independence. What is more, the lower boundary on y_i exceedes the pure strategy trust building equilibrium.

Lemma A.18 *If*

$$\omega_i^z \in [0,1]$$

$$y_i > \frac{a+(1+a)(1-\beta_j(1+a))}{(1+a)(1-\beta_j)}$$

, then the following strategy profile is an equilibrium $s^A(pr(a_1 = c) = \omega_A^z|A_s, a_1 = d|A_g, a_2 = c|A_s\&a_1 = b_1 = c, a_2 = d|$ *otherwise),* $s^B(pr(b_1 = c) = \omega_B^z|B_s, b_1 = d|A_g, b_2 = c|B_s\&a_1 = b_1 = c, b_2 = d|$ *otherwise).*

In the second period, $\sigma_2^A = 1|b_1 = c$, $\sigma_2^B = 1|a_1 = c$, which assures we can sustain second period cooperation. Note we can sustain mutual defection given any set of beliefs, as desired.

Moving to the first period, first we prove ω_i^z holds i_s indifferent. Focusing on A_S, $\omega_B y_B(H + H) + (1 - \omega_B y_B)(\beta_A H + 1) = \omega_B y_B(H - a + 1) + (1 - \omega_B y_B)(\beta_A(H - a) + 1 - \beta_A + 1)$, solves for $\omega_B = \omega_i^z$. Trivially, if we can sustain $\omega_i^z \in (0,1)$ then we can sustain i_s's first period strategy. Greedy types cannot profitably deviate from $a_1 = d \to c$ if $\omega_B y_B(H + a + 1) + (1 - \omega_B y_B)(\beta_A(H + a) + 1 - \beta_A + 1) > \omega_B y_B(H + H + a) + (1 - \omega_B y_B)(\beta_A H + 1)$. Rearranging and plugging in $\omega_B = \omega_A^z$, this is satisfied if $\beta_A(a + 2H - 3) + 1 > 0$. This is always satisfied if $\mathcal{A}_2$ holds, as desired. There are no off-path beliefs. This completes the proof.

Finally, we **cannot support** any equilibria where greedy types mix in the first period. The binding constraint is the following strategy profile: $s^A(pr(a_1 = c) = \omega_A^g|A_g, a_1 = c|A_s, a_2 = d)$, $s^B(pr(b_1 = c) = \omega_B^g|B_g, b_1 = c|B_s, b_2 = d)$. In this case, we can hold A_g indifferent if: $y_B(H + a) + (1 - y_B)(\omega_B^g(H + a) + (1 - \omega_B^g)(\beta_A(H + a) + 1 - \beta_A)) = y_B H + (1 - y_B)(H\omega_B^g + (1 - \omega_B^g)\beta_A H$. This simplifies to, $1 - \beta_A(1 - a) - y_B(1 - a)(1 - \beta_A) = \omega_B^g(1 - y_B)(1 - \beta_A)(1 - a)$. If $a \geq 1$, the RHS is negative. If $a < 1$, the RHS is less than the left. Thus, $\omega_B^g > 1$. It follows that we cannot sustain a mixing probability that leaves A indifferent.

Mixing in both periods Finally, we solve for equilibria where i_s plays mix strategy in both periods. Both of these equilibria are **trust building equilibria**. A critical feature to note is that if i_s mixes in the second period, then no matter posterior beliefs σ_2^i that security seekers expect $\frac{H}{1+a}$ and greedy types expect $\frac{H+a-1}{1+a} + 1$. This will greatly simplify our analysis.

We begin with the case wherein security seekers condition their decision to mix in the second period if they observe first-period mixing because it imposes the fewest conditions on y_i.

$$\omega_i^\gamma = \frac{1 - \beta_j(a+1)}{y_i(\frac{H}{1+a} - \beta_j(a+1)}$$

As with all other mixing equilibria, $\omega_i^\gamma \in (0,1)$ imposes a minimum bound on $y_i > \frac{1-\beta_j(a+1)}{(\frac{H}{1+a}-\beta_j(a+1))}$, where $y_i > \frac{1+a}{H} > 0$, given full dependencies. Note this cannot be achieved for any y_i under $\mathcal{A}_2$, and for any β_j, the lower boundary on y_i exceeded the pure strategy trust building equilibrium. What is more, the lower boundary on y_i exceeded the pure strategy trust building equilibrium.

Lemma A.19 *If* $\omega_i^\gamma \in (0,1)$*, then the following strategy profile is an equilibrium* $s^A(pr(a_1 = c) = \omega_A^\gamma | A_s, a_1 = d | A_g, pr(a_2 = c) = \omega_A^* | A_s \& a_1 = b_1 = c, a_2 = d|$ *otherwise*$)$*,* $s^B(pr(b_1 = c) = \omega_B^\gamma | B_s, b_1 = d | B_g, pr(b_2 = c) = \omega_B^* | B_s \& a_1 = b_1 = c, b_2 = d|$ *otherwise*$)$*.*

We've shown in the second period, i_s cannot profitably deviate if $\sigma_2^i > \frac{1}{a+1}$. Note that $\sigma_2^i | j_1 = c = 1$, as desired. Also note that we can always support second period mutual defection, which we assert in all cases other than $a_1 = b_1 = 1 \& i_s$.

Working backwards, we solve for ω_i^γ. i_s is held indifferent given ω_i^γ. Focusing on A_s, $\omega_B y_B(H + \frac{H}{1+a}) + (1 - \omega_B y_B)(H\beta_A + 1) = \omega_B y_B(H - a + 1) + (1 - \omega_B y_B)(\beta_A(H - a) + 1 - \beta_A + 1)$, which gives us $\omega_B = \omega_B^\gamma$, as desired.

The greedy type cannot profitably deviate from $a_1 = d \to c$ if: $y_B \omega_B(H + a + 1) + (1 - y_B \omega_B)(\beta_A(H + a) + 1 - \beta_A + 1) > y_B \omega_B(H + \frac{H+a-1}{1+a} + 1) + (1 - y_B \omega_B)(\beta_A H + 1)$. This solves for $1 + \beta_A a - \beta_A > y_B \omega_B(\frac{H+a-1}{1+a} + (1 - a)(1 - \beta_A))$, subbing in ω_i^γ, the y_A cancels and the inequality is always solved, as desired. This completes the proof.

Finally, define a first period mixing probability,

$$\omega_i^\alpha = \frac{1 - \beta_j(a+1)}{y_i(a+1)(1 - \beta_j)}$$

As with all other mixing equilibria, $\omega_i^\alpha \in (0,1)$ imposes a minimum bound on $y_i > \frac{1-\beta_j(a+1)}{(a+1)(1-\beta_j)}$, where $y_i > \frac{1}{(a+1)} > 0$, given full dependencies.

Lemma A.20 *If* $\omega_i^\alpha \in (0,1)$*, and*

$$y_i > \frac{\frac{a}{(a+1)} + 1 - \beta_j(a+1)}{(a+1)(1 - \beta_j)} > \frac{1}{a+1}$$

hold, then the following strategy profile is an equilibrium $s^A(pr(a_1 = c) = \omega_A^\alpha | A_s, a_1 = d | A_g, pr(a_2 = c) = \omega_A^* | A_s, a_2 = d | A_g)$*,* $s^B(pr(b_1 = c) = \omega_B^\alpha | B_s, b_1 = d | B_g, pr(b_2 = c) = \omega_B^\alpha | B_s, b_2 = d | B_g)$*.*

The analyses above are enough to show that i_g cannot profitably deviate given that i_s will mix regardless. Thus, we analyze i strategy. We've shown in the second period, i_s cannot profitably deviate if $\sigma_2^i > \frac{1}{a+1}$. A's posterior belief is $\sigma_2^A | b_1 = d = \frac{y_b(1-\omega_B^\alpha)}{1-\omega_B^\alpha y_B}$. Plugging in ω_B^α, $\sigma_2^A | b_1 = d = \frac{y_B(a+1)(1-\beta_A)-1+\beta_A(a+1)}{a}$. To sustain the equilibrium, $y_B > \frac{\frac{a}{(a+1)}+1-\beta_A(a+1)}{(a+1)(1-\beta_A)}$, as written. We emphasize that $\frac{\frac{a}{(a+1)}+1-\beta_j(a+1)}{(a+1)(1-\beta_j)} > \frac{1}{a+1}$ to make clear that there must be so much initial trust, that i_s is still willing to mix even after being cheated in the first period.

Working backwards, we solve for ω_i^α. i_s is held indifferent given ω_i^α. Focusing on A_s, $\omega_B y_B H + (1 - \omega_B y_B) H \beta_A = \omega_B y_B(H - a) + (1 - \omega_B y_B)(\beta_A(H - a) + 1 - \beta_A)$, which gives us $\omega_B = \omega_B^\alpha$, as desired.

A.4 Robustness: Introducing salience

We now allow $\theta > 0$ to vary. Since our question is, does trust-building operate at different levels of salience, we focus on that equilibrium.

We define $\mathcal{C}_1 := \theta < \frac{(a-1)(1-\beta)}{H+a-1}, a > 1$. This condition helps us manage a corner condition on the Greedy type's incentives to defect in equilibrium condition 6.

Proposition A.21 *If*

$$y_i > \frac{1-\beta_j(a+1)}{\theta(H-1)+(a+1)(1-\beta_j)} \tag{5}$$

holds, and either $\mathcal{C}_1$ holds or $\mathcal{C}_1$ is violated and

$$\frac{1+\beta_j(a-1)}{\theta(H+a-1)-(a-1)(1-\beta_j)} > y_i \tag{6}$$

holds, then there is a pure strategy trust-building equilibrium with the same strategy profile as written in Proposition A.3.

We start with A_s's strategy. Because this is a complete separating equilibrium, $\sigma_2^A|b_1 = d = 0, \sigma_2^A|b_1 = c = 1$. By Lemma A.1, we can support second-period cooperation. Turning to the first period, A_s prefers to cooperate, rather than defect if:

$$y_B(H+H\theta)+(1-y_B)(\beta_A H+\theta) > y_B(H-a+\theta)+(1-y_B)(\beta_A(H-a)+1-\beta_A+\theta)$$

This assumes that if A defects, A gets the second-period value $\theta \times 1$ from $b_2 = a_2 = d$. This re-arranges to equilibrium condition 3, as desired.

Turning to A_g's strategy. We've already shown we can support second-period defection for any set of beliefs. Turning to the first period, A_g prefers to defect, rather than cooperate if:

$$y_B(H+a+\theta)+(1-y_B)(\beta_A(H+a)+1-\beta_A+\theta) < y_B(H+\theta(H+a))+(1-y_B)(\beta_A H+\theta)$$

$$y_B(\theta(H+a-1)-(a-1)(1-\beta_A)) < 1+\beta_A(a-1)$$

If $\theta(H+a-1)-(a-1)(1-\beta_A) < 0$, this is always satisfied. This gives us $\mathcal{C}_1$. The basic logic here is that if the second period does not matter, then B holds little incentive to trick A, and instead relies on her direct incentive. In the case that $\mathcal{C}_1$ is violated, then B's value for tricking A increases. In this case, B must reckon with expectations about A's type. B's incentives rearrange to inequality 6 as desired. There are no off-path beliefs. This completes the proof.

A.4.1 Result 2: Implications of salience

Result 2: If the independence threshold described in Result 1a ($\beta > \frac{1}{1+a}$) holds, then there is no Goldilocks problem.

1. There is no upper bound on the salience of domestic choices for which we can support the trust-building equilibrium. The security-seeker prefers first-period cooperation for any level of salience, even at very low levels of initial trust.

2. There is a lower bound on the salience of domestic choices for which we can support the trust-building equilibrium. The greedy type can profitably deviate to first-period cooperation if first-period salience is low.

We are interested in the interaction of three parameters: θ, β, y. We start with the security-seeker's incentives.

In one critical way, salience does not affect the security-seeker's ICC. Absent salience, A_s strictly preferred first-period cooperation if $\beta_i > \frac{1}{a+1}$. Introducing θ does not affect this threshold condition. It follows that this threshold still represents the minimum level of independence necessary to sustain trust-building for any initial level of trust.

For β below this threshold, we get a slightly adjusted condition:

$$\beta > \frac{1 - y[\theta(H-1) + (a+1)]}{(a+1)(1-y)}$$

Returning to how it is written in condition 6, we see that we can support it for lower levels of initial trust if θ is larger.

Salience has more substantial implications on the greedy type's incentives in a way that strengthens our core result. Notice, that the dynamics of the equilibrium are different when θ is so low that $\mathcal{C}_1$ is satisfied. In this case, we can always satisfy the greedy type's ICC.

Condition 6 governs the case when $\mathcal{C}_1$ is violated. It shows two things: (a) When $\mathcal{C}_1$ is violated, the incentives for the greedy type to profitably deviate from the trust-building equilibrium are increasing in θ. However, (b) for any θ, and $\beta > \frac{1}{a+1}$ there exists a y sufficiently small to support the trust-building equilibrium.

Putting these together with the security-seeker's incentives we can support the following claim in the manuscript.

Remark Grand Gestures: So long as $\beta > \frac{1}{a+1}$, then we can support the trust-building equilibrium for any $\theta < \frac{1+\beta(a-1)+y(a-1)(1-\beta)}{H+a-1}$.

For low levels of independence, A_s's incentives to cooperate in the first period are increasing in θ. This would work against the grand gesture by imposing a lower limit on θ. However, once we surpass, $\beta > \frac{1}{a+1}$, A_s strictly prefers cooperation for any θ. What is more, the greedy type's ICC guarantees she cannot profitably deviate from trust-building given a θ sufficiently low.

A.5 Robustness: Similarity

This section explores the impact of variation the similarity, or correlation, of domestic and international preferences.

We introduce similarity as a random variable $\alpha > 0.5$. We draw $pr(\alpha_i = 1) = \alpha$. If $\alpha_A = 1$ then the payoffs are as they are in Table 3 for player A. If $\alpha_A = 0$, then player A's first-period payoffs are reversed. The greedy type gets the security-seeking type's payoffs and the security-seeking type gets the greedy type's payoffs. B's payoffs are defined the same way. Here α represents how similar the two choices are in that when $\alpha = 1$ players are certain that first- and second-period preferences are aligned. When $\alpha = 0.5$ it means that there is an even chance that payoffs align or do not align across periods. In other words, 0.5 is the value of α at which domestic choices provide the least information.

The sequence of moves is as follows:

- Nature draws player types i.i.d from y_i (private)
- Nature draws α_i i.i.d. (private)
- A first trust problem arises in which A and B simultaneously select $s_{i1} = c, d$.
- A second trust problem arises in which A and B simultaneously select $s_{i2} = c, d$.
- Payoffs are realized.

A.5.1 Analysis

Our goal is to show that the trust-building equilibrium can survive under this condition. Thus, we solve for equilibria that are close to the pure strategy trust-building equilibria reported in Proposition A.21. Specifically, we are looking for equilibria where A_s plays $a_2 = c|b_1 = c$, and defects otherwise. There are two. In one, A_s follows her direct first-period incentives, in the other A_s always cooperates in the first period no matter her direct incentives. First, we solve for the latter.

Proposition A.22 *When*

$$y_B > \frac{1-\alpha_B}{1+a-\alpha_B} \tag{7}$$

and

$$\frac{a-\alpha(1-\beta)(a-1)}{\theta(H-1+a)-\alpha(1-\beta)(a-1)} > y > \frac{(\theta+a)(1-\alpha)+\alpha(1+\beta(a-1))}{\theta H-\alpha(\theta+a)+\alpha(1+\beta(a-1))} \tag{8}$$

holds. Then A can support the following strategies in a symmetric PBE. $s^{A_g}(a_1 = c|1-\alpha, a_1 = d|\alpha, a_2 = d)$, $s^{A_s}(a_1 = c, a_2 = d|b_1 = d, a_2 = c|b_1 = c)$. *B's condition and strategies are defined symmetrically.*

In this variant of trust-building, A_s cooperates no matter what A's first-period motivations are. But A_g's strategy depends on α.

We start with second-period strategies. We showed in Lemma A.1 that A_g always defects (as desired), and A_s cooperates iff $\sigma_2^A > \frac{1}{1+a}$. In equilibrium, A's posterior beliefs after observing $b_1 = 1$ are:

$$\sigma_2^A|b_1 = c = \frac{y_B}{y_B + (1-\alpha_B)(1-y_B)}$$

Setting $\sigma_2^A > \frac{1}{1+a}$ this solves for equilibrium condition 7, as desired.

We now turn to first-period strategies. From A_s's perspective, clearly first-period cooperation is hardest to sustain in the $1-\alpha$ case (rather than α case).Focusing on the $1-\alpha$ case, A_s prefers first-period cooperation to defection iff:

$$y_B(H+H\theta)+(1-\alpha_B)(1-y_B)H+(1-y_B)\alpha_B(\beta_A H+\theta) > y_B(H+a+\theta)+(1-\alpha_B)(1-y_B)(H+a+\theta)+(1-y_B)\alpha_B(\beta_A(H+a)+1-\beta_A+\theta)$$

This solves for the RHS of the equilibrium condition 8.

Turning to A_g's incentives. In equilibrium $a_1 = c|1-\alpha$, and $a_1 = d|\alpha$. Clearly, it is easier to sustain $a_1 = c|1-\alpha$ because this type gets the maximum expected value in the second period, and loses a from defection in the first period. Focusing on the α case, A_g prefers first-period defection iff:

$$(y_B+(1-y_B)(1-\alpha_B))(H+a+\theta)+(1-y_B)\alpha_B(\beta_A(H+a)+1-\beta_A+\theta) > y_B(H+\theta(H+a))+(1-y_B)(1-\alpha_B)(H+\theta)+(1-y_B)\alpha_B(\beta_B H+\theta)$$

This solves for the LHS of the equilibrium condition 8. There are no off-path beliefs.

We now solve for the former

Proposition A.23 *When*

$$y > \frac{1-\alpha}{1+\alpha(a-1)} \tag{9}$$

and

$$\frac{a-\alpha(1-\beta)(a-1)}{\alpha[\theta(H-1+a)-2(1-\beta)(a-1)]+(1-\beta)(a-1)} > y > \frac{\alpha[(1-\beta)(a+1)-\theta]+\theta-a}{\alpha[\theta H-2\theta+2(1-\beta)(a+1)]-(1-\beta)(a+1)} \tag{10}$$

and

$$\frac{\theta-a+\alpha(\beta a+1-\beta+a-\theta)}{\theta-(a(\beta+1)+1-\beta)+\alpha[\theta(H-2)+a(\beta+1)+1-\beta]} > y \tag{11}$$

holds. Then A can support the following strategies in a symmetric PBE. $s^{A_g}(a_1 = c|1-\alpha, a_1 = d|1-\alpha, a_2 = d)$, $s^{A_s}(a_1 = c|\alpha, a_1 = d|1-\alpha, a_2 = d|b_1 = d, a_2 = c|b_1 = c)$. *B's condition and strategies are defined symmetrically.*

We showed in Lemma A.1 that A_g always defects (as desired), and A_s cooperates iff $\sigma_2^A > \frac{1}{1+a}$. In equilibrium, A's posterior beliefs after observing $b_1 = 1$ are:

$$\sigma_2^A|b_1 = c = \frac{y_B\alpha}{y_B+(1-\alpha_B)(1-y_B)}$$

Setting $\sigma_2^A > \frac{1}{1+a}$ this solves for equilibrium condition 7.

We now turn to first-period strategies. On path, A_s cooperates in the α condition. A^s cannot profitably deviate to defect in this condition iff:

$$y_B\alpha(H+H\theta)+(1-\alpha_B)(1-y_B)H+[(1-y_B)\alpha_B+y_B(1-\alpha)](\beta_A H+\theta) > [y_B\alpha+(1-y_B)(1-\alpha)](H-a+\theta)+[(1-y_B)\alpha_B+y_B(1-\alpha)](\beta_A(H-a)+1-\beta+\theta)$$

This solves for the RHS of the equilibrium condition 10.

On path, A_s defects in the $1-\alpha$ condition. A^s cannot profitably deviate to cooperate in this condition iff:

$$[y_B\alpha+(1-y)(1-\alpha)](H-a+\theta)+[(1-y_B)\alpha+y_B(1-\alpha)](\beta(H+a)+1-\beta+\theta) > y_B\alpha(H+\theta H)+(1-y)(1-\alpha)H+[(1-y_B)\alpha+y_B(1-\alpha)](\beta H+\theta)$$

This solves for equilibrium condition 11.

Turning to A_g's incentives. In equilibrium $a_1 = c|1-\alpha$, and $a_1 = d|\alpha$. Focusing on the α case, A_g prefers first-period defection iff:

$$[y_B\alpha+(1-y)(1-\alpha)](H+a+\theta)+[(1-y_B)\alpha+y_B(1-\alpha)](\beta(H+a)+1-\beta+\theta) > y\alpha(H+\theta(H+a))+(1-y)(1-\alpha)(H+a+\theta)+[(1-y_B)\alpha+y_B(1-\alpha)](\beta H+\theta)$$

This solves for the LHS of the equilibrium condition 10. There are no off-path beliefs.