The Domestic Sources of International Trust

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

January 5, 2026

Abstract

How do states overcome mistrust? Scholars argue that costly foreign policy signals 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 case studies of the Cold War's end and U.S.-South Korea trust-building post-coup, illuminating several phenomena the current literature struggles to explain: initial trust-building between enduring rivals, asymmetric trust-building, and trust-building through 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)

‡Professor, Department of Political Science, Pennsylvania State University (rum842@psu.edu). The order of the authors is alphabetical and the contributions are equal. This material is based upon work supported by the National Science Foundation under Awards No. 2342950 and 2342951. We thank Andrew Kydd, Soyoung Lee, Brandon Yoder, Terence Roehrig, Tyler Pratt, Charles Glaser, Alexandre Debs, our anonymous reviewers, and participants at ISA, Peace Science, and the Penn State-Pitt IR Workshop for helpful feedback.

Prominent research argues that states can use costly signals to overcome the security dilemma (e.g., Kydd 2000; Glaser 2010; Jervis 1978; Yoder 2019b; Haynes and Yoder 2020; Yoder and Haynes 2025). According to this literature, states reveal their genuine security-seeking intentions by pursuing *foreign* policies that greedy states would not, such as reducing arms or joining international agreements. 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). From 1985 to 1987, Mikhail Gorbachev released dissidents from internal exile, loosened emigration restrictions for persecuted Soviet Jews, and introduced *glasnost* and *perestroika*, all of which increased Western trust in Soviet intentions (Bartel 2022; Wilson 2014). More recently, after rebels formerly aligned with ISIS toppled the long-standing Syrian regime, President Biden's remarks indicated that domestic policies such as protection of religious minorities would be a key litmus test for U.S. cooperation (Biden 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 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 the above examples evidence how *both* liberal and illiberal—even violent or coercive—actions reassured the United States. Second, researchers believe that trust-building between international rivals must start small, for example via modest arms control agreements, since even security-seekers fear exploitation (Kydd 2005). Yet, as detailed below, Gorbachev initiated trust-building with the West via domestic choices that, in the Soviet context, were revolutionary. Finally, while existing literature focuses on reciprocal trust-building choices, some trust-building described above began with unilateral domestic actions.

To our knowledge, trust-building via domestic choices has previously been mentioned only

in passing (Kydd 1997). Yet the puzzles above suggest that how domestic choices build trust departs from our understanding of foreign policy trust-building. Thus, a systematic analysis linking domestic policies and international trust is needed. We take up this task, theorizing the unique features of domestic policies that facilitate unappreciated opportunities to build trust.

To begin, we construct a simple two-period model of international trust problems similar to Kydd (2005), Yoder and Haynes (2025), and others. We then advance two conceptual claims that differentiate domestic policy choices from international signals, which we introduce into the model. First, states' preferences over domestic policies are often correlated with foreign policy preferences. These correlations allow domestic policies to function as costly trust-building signals on the international stage. Second, 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 payoffs from a state's policy choice depend upon an international counterpart's choice. In the classic security dilemma model of costly international signaling, scholars assume that choices are highly dependent; one state's payoff greatly depends on the other's decision (Jervis 1978; Kydd 2005). We argue that domestic decisions tend to be more (but not necessarily entirely) independent, meaning that the benefits and costs that a state derives from them depend less on what their counterpart does. The relative independence of domestic choices allows them to spark trust-building between states that seek cooperation but consider international signals too risky.

These insights illuminate the empirical puzzles described earlier. First, our theory explains trust-building following illiberal domestic choices. While liberal policies can establish trust in some cases, our theory indicates that preference-compatibility is the key determinant of international cooperation. Thus, there are some situations—e.g., harsh anti-Communist repression—in which illiberal domestic choices could reassure even liberal states. Second, it explains the ability to build trust through grand gestures on the domestic level: When choices have high independence, security-seekers can undertake salient and reassuring domestic policies without fear of international exploitation. Finally, we identify a novel asymmetric trust-building equilibrium when initial

trust and independence are both lopsided. This explains historical cases where one side takes a unilateral first step, knowing the other side will not initially reciprocate.

We illustrate these insights in case studies of U.S.-Soviet trust-building at the Cold War's end and U.S.-South Korea trust-building during the Cold War. The former illustrates how independence allows trust-building to occur through grand gestures, even among highly distrustful rivals. The latter provides an illustration of illiberal and asymmetric trust-building, which we conjecture could be common in Cold War U.S.-client relations.

Our argument has several implications. First, we offer a general theory of domestic politics and international trust. The extensive trust literature has paid scant attention to domestic politics. We 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.

Second, given that political elites spend the majority of their time on domestic politics (Lindsey and Hobbs 2015), we expose avenues for trust-building that past scholarship may have overlooked. This 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), unethical government research (e.g., Joseph and Poznansky 2024), and discrimination against minorities and immigrants (e.g., Peters 2015). We provide a common framework for understanding why, whether, and under what conditions, any domestic choice facilitates international trust or mistrust.

Finally, we advance research on regime type and peace (e.g., Maoz and Russett 1993a,b; Russett and Oneal 2001). Our framework sheds new light on the liberal peace, providing a novel explanation for how democracies can identify each other's cooperative international preferences. Yet our approach goes beyond the democratic peace by explaining trust-building between autocra-

cies or between mixed-regime pairs. It explains 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.

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; Joseph 2026), and general problems of cooperation (Crescenzi 2018; Goldfien 2024). Policymakers care about trust problems. 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-type states: security-seeking and greedy (Glaser 2010). The fundamental difference between them is their value from exploiting each other. Security-seekers maximally benefit from reciprocal cooperation, while greedy states prefer to defect no matter what their partner does.[1] A trust problem arises because states are uncertain about each other's intentions; and the value they accrue from cooperation depends on what the other does.

We visualize a formalization of this problem in Table 1. Here, the trust problem focuses on two states that will soon confront a critical foreign policy choice, potentially the choice to comply with an arms control agreement or not; or to withdraw forward-deployed troops or not. In such situations, each state can choose to cooperate or defect. Each box in the 2×2 of Table 1 describes the payoffs players accrue given different combinations of those choices. Table 2 explains the parameters.

The critical feature across all trust models is that security-seekers and greedy types differ in their preference ordering over different combinations of choices. If A is greedy, then A's prefer-

[1]The sources of these motives are beyond our scope. For overviews, see Joseph (2021, 2026).

Table 1: Representation of the Trust Problem

		Player B: Cooperate ($b_t = c$)	Player B: Defect ($b_t = d$)
Player A	Cooperate ($a_t = c$)	Mutual cooperation $1, 1$	B cheats A $-k, e_B^m$
	Defect ($a_t = d$)	A cheats B $e_A^m, -k$	Mutual defection $0, 0$

Note: Each box details an informal description and player-specific payoffs given the four realizable combinations of potential decisions. The values for mutual cooperation and mutual defection are normalized to $1, 0$ respectively. Players' motivates are privately drawn given $pr(e_i^m = L) = p_i, pr(e_i^m = H) = 1 - p_i$.

Table 2: Substantive description of parameters

Parameter	Description
e_i^m	i's value from exploiting j. $m \in \{g, s\}$ represents i's motives. $e_i^s = L$ (low-value) and $e_i^g = H$ (high-value) are exploitation payoffs for security-seekers and greedy types, respectively ($H > 1 > L$).
$k > 0$	The cost of being exploited
$p_i \in [0, 1]$	Probability i is security-seeker. Determines j's initial trust in i.

ence over outcomes is: A cheats B $>$ mutual cooperation $>$ mutual defection $>$ B cheats A. If A is security-seeking, then it is: mutual cooperation $>$ A cheats B $\sim$ mutual defection $>$ B cheats A, where $\sim$ represents flexible ordering. B's ordering is symmetric. Note that both types receive the worst payoff if cheated (i.e., the sucker's payoff). Substantively, being cheated on major international agreements not only puts a state's national security at risk, but can damage a leader's international reputation and 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. Yet states are usually uncertain about each other's intentions, and this can complicate cooperation. 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 p_B. She believes that B is greedy with probability $1 - p_B$. Similarly, B believes that A is a security-seeker with probability p_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, and when does uncertainty cause them to defect? Scholars rationalize 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 trust A. Here, A also defects because A expects B to defect due to B's distrust. This second mechanism illustrates the difficulties of two-sided trust problems. Each side must trust the other and 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 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).

1.1 The Baseline of Foreign Policy Signaling

We treat extant research on trust-building through foreign policy signals as our baseline. In Appendix A.1, we introduce a two-period model that first privately draws each state's intentions (using probabilities p_A, p_B), then iterates the simultaneous-move trust-building model visualized in Table 1 over two periods. In later sections, we analyze the effects of domestic (or other independent) signals by manipulating features of the first-period choice based on our substantive arguments about why domestic policies are different.

We advance the simultaneous-move model for two reasons. First, it harnesses the power of incrementalism (Ashworth, Berry, and de Mesquita 2021, 58-59) by adding one parameter (independence, introduced next) to the framework common across much trust-building research (Jervis 1978; Kydd 2000, 2005; Schultz 2005; Yoder and Haynes 2025; Acharya and Ramsay 2013). Second, it captures the strategic problem we hope to study. 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., Glaser 2010; Waltz 1979; Braumoeller 2008). Thus, at every moment, each worries the other is cheating them, creating pressure to cheat first. If we instead sequenced choices, it would artificially resolve the trust problem for whoever moves second.[2]

In Appendix A.2, we analyze the baseline model for Perfect Bayesian Equilibria (PBE). Consistent with Kydd (2005) and others, we find many PBE—including mixed and pure strategy equilibria—in overlapping parameter ranges. We report them all in appendices. In the manuscript, we focus on equilibria that survive an efficiency refinement (PBE-ER). This rules out any equilibria if, under the same parameter ranges, all types of all states would prefer to play a different equilibrium (i.e., it is Pareto dominated given initial beliefs). This is substantively appealing for the cases that motivate us because security-seekers often want to end competition because it is inefficient. It also holds technical appeal. Empirical implications of all trust models are hampered by multiple

[2]We accept some substantive areas exist, including proliferation (Debs and Monteiro 2014), where one-sided exploitation is normal.

equilibria in overlapping parameter ranges. One reason we are confident in our main conclusion is that regardless of whether we apply a refinement or not, we cannot rationalize trust-building in the international signaling model when initial trust is low. Yet, as we show below, we can rationalize trust-building in these ranges when we model domestic signaling opportunities. Further, in the domestic signaling model, only trust-building equilibria survive refinement when initial trust is low, and they reliably do so, suggesting that these results are most robust in interesting parameter ranges.

The results of our baseline analysis are summarized in Figure 1. Panel (b) describes the equilibria that can survive the PBE-ER. Two points are notable. First, all equilibria are symmetric, pure strategy equilibria. Second, given we start with the same baseline structure as others (e.g,, Kydd 2005), the three PBE-ER we find are consistent with past scholarship on trust-building through international costly signaling. We are most interested in the trust-building equilibrium. 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 each player learns the other's motives from first-period actions. When they both cooperate, they come to trust each other, facilitating second-period cooperation.[3]

Panel (a) plots the conditions under which the equilibria listed in panel (b) survive as a function of each player's initial trust levels. The conditions reflect findings in past work (Kydd 2005).[4] Here we emphasize the boundaries of the trust-building equilibrium (which we shade dark red in this and all future plots).[5] We find there is a minimal level of trust necessary to sustain trust-building. Crucially, this lower bound is not a function of the efficiency refinement.[6] Rather, we cannot sustain first-period cooperation in any PBE with low levels of initial trust because when at least one state initially believes the other is very likely greedy, that state defects in period one. The other side,

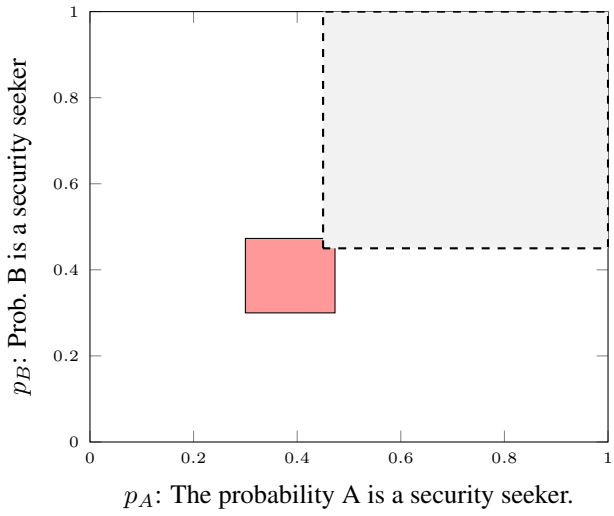
[3]This symmetric equilibrium is the only trust-building PBE-ER in the baseline model. After introducing our innovation below, we find both symmetric and asymmetric trust-building equilibria. Therefore, we later refer to this classic trust-building equilibrium as the "symmetric trust-building equilibrium."

[4]Notably, Kydd does not apply a Pareto refinement. He sustains inefficient equilibria we discuss in appendices.

[5]As Kydd (2005, 192) shows, the boundaries of this equilibrium will shift with different parameter values, but the core conclusions remain the same.

[6]The tragic (mutual defection) equilibrium is the least efficient. If states can rationalize cooperation through another equilibrium, it Pareto dominates the tragic equilibrium.

Figure 1: Equilibria that Survive Refinement in the Standard Trust-Building Game



(a) Conditions where PBE emerge

Motive	Period 1	Period 2 | Period 1			
		CC	CD	DC	DD
Trust-building equilibrium (dark red, solid)					
Security	C	C	D	D	D
Greedy	D	D	D	D	D
Suckers equilibrium, (gray, dashed)					
Security	C	C	D	D	D
Greedy	C	D	D	D	D
Tragic equilibrium (white space)					
Security	D	D	D	D	D
Greedy	D	D	D	D	D

(b) Equilibrium description

Note: Panel (a) plots three equilibria that can survive an efficiency refinement given $H = 1.5, L = -0.1, k = .9$. These equilibria uniquely survive refinement in their plotted parameter ranges. Panel (b) describes the strategy profiles. Since these are symmetric strategies, we present them for one state. We report inefficient pure and mixed strategy PBE in Appendix A.2, and A.2.1, respectively.

regardless of its true motives and level of trust, also defects because it understands exploitation is certain. Because it is impossible to build trust in the first period, this also assures we cannot sustain second-period cooperation. Therefore, when initial trust (p_i) is low, the unique PBE in the baseline model is mutual defection in both periods. This finding explains why trust-building is difficult, especially for enduring rivals or other highly distrustful states. 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?

2 Domestic Policies and International Signaling

We argue that when there is too much distrust for security-seeking states to risk international cooperation, they can still kick-start the trust-building process through costly signaling via domestic policy choices. When we say domestic choices, we do not primarily mean broad attributes of states, such as regime type (Russett and Oneal 2001). Regime type is 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 and Joseph 2023; Goldfien, Joseph, and Krcmaric 2023).

Instead, we focus generically on government decision-makers who are confronted with a domestic problem and can select among different policy options to resolve it. This might include when a government faces nationwide protests and must decide between repression or negotiation with demonstrators; or when it faces economic stagnation and 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 preferences over these options are related to its broader preferences.[7] As at 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. 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 signal international motives because states' domestic poli-

[7]Leaders usually have more than two possible responses to domestic problems, but that is also true of international problems. Like the international trust-building literature, we emphasize two choices that represent the options most favorable to security-seeking and greedy types.

cies are often related to their foreign policy goals. Goldfien, Joseph, and McManus (2023) 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 argue that preferences across the domestic and international policy decision-spaces are similarly correlated in certain trust-building scenarios. That is, the preference to exploit international cooperation is likely correlated with certain domestic policy preferences.

One version of our claim overlaps with the liberal peace. Some argue that domestic policies such as free elections and respect for ethnic minorities and human rights predict security-seeking international intentions (Kydd 1997; Tomz and Weeks 2020; Maoz and Russett 1993a; Tomz and Weeks 2013b). Indeed, our first case study illustrates how the U.S. interpreted Soviet liberalizing domestic reforms as evidence of international trustworthiness. But we argue that what really matters is compatible international preferences, not liberal actions per se.[8] 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.[9] For 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.[10]

Since the correlation between preferences for domestic policies and international cooperation is context-dependent, we cannot identify particular domestic choices that always promote trust. Indeed, democracies, Communist regimes, right-wing dictatorships, and theocratic regimes might draw the *opposite* trust inferences from the same domestic decisions. For example, European analysts are likely to interpret the Turkish government's push to expand Islam's role in public life as evidence of lower willingness to cooperate, but analysts in Islamist states may infer higher

[8]For research highlighting the importance of compatible preferences in other contexts, see Voeten (2021), Gartzke (1998), and Spaniel and Smith (2015).

[9]There are even cases where liberalizing reforms can induce mistrust. For example, much earlier liberalizing reforms in the Soviet Union under Khrushchev arguably worsened relations with China (Haynes and Yoder 2020)

[10]Domestic suppression 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.

willingness to cooperate.

Furthermore, we do *not* claim that security-seeking and greedy states hold reliably different preferences over all, or even most, domestic policy choices. Many domestic choices (e.g., the speed limit) have no bearing on world politics. Further, even when a domestic choice is correlated with international preferences, the correlation is 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.[11] 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 correlation is imperfect. 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.[12] We find domestic choices can induce trust and international cooperation with only a moderate correlation (defined precisely in Appendix A.5). However, the amount of information communicated is greater when the correlation is stronger.

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? This is the question we now consider.

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

[11] Even at the international level, this correlation likely varies. Yet existing trust-building models have not examined it (Kydd 2005).

[12] See Appendix A.5.

choice depend upon what a foreign counterpart does.[13] 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 its own preferences. Yet sometimes the payoff also depends, at least partially, 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 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 or negotiate with protesters. If A's value from either repressing or negotiating does not depend on anything B does, then A's decision about how to respond to the protests is fully payoff independent. If instead, A's payoff from this decision depends somewhat on B's response (e.g., whether B lodges diplomatic criticism or 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 is fully independent. But if the payoffs depend slightly, partially, or highly on what another 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 not (defects). This assumption is reasonable because the value states accrue from many international policy options previously studied are heavily determined by what rivals do.

In contrast, domestic choices often have greater independence. Examples of domestic choices that are plausibly near-fully independent include reforming social welfare, reducing domestic regulations, or changing land use policy. The associated costs and benefits have little to do with the choices foreign countries make. But choices about immigration policy and press freedom could be

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

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). This last example also highlights that States A and B need not be choosing among identical options in trust-building (or even among options that are both domestic or both international).

Overall, we argue that domestic choices tend to be more independent than foreign policy choices *on average*. We also believe few domestic choices are as dependent as the choice to cut or build arms that the trust-building literature typically focuses on. However, not all domestic (foreign) policy choices are fully independent (dependent), and some foreign policy choices are even more independent than some domestic choices. Table 3 summarizes how various domestic and international choices can vary in independence. To be clear, these are on-average codings, and we encourage researchers to apply contextual knowledge in determining how independent actions are in historical cases.

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 does (i.e., full dependence). If the payoff from taking an action is independent of what the other does, this fear diminishes, and security-seekers face little risk when they make choices that reflect their genuine cooperative preferences.[14] For example, a state shifting policy in response to a domestic crisis would not fear that a counterpart would be able to greatly exploit its choice.[15] At the same time, greedy states are more easily identified because they can no longer claim, "I am not cooperating because I am afraid that you will cheat me." 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.[16]

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

[15]If the counterpart is security-seeking, it would prefer to reciprocate cooperation anyway.

[16]This is also true of independent international choices, although we believe domestic choices tend to have greater

Table 3: Examples of Choices with Varying Degrees of On-Average Independence

Options	Independence	Explanation
Domestic: Low/high speed limit	High	Costs/benefits do not depend on other countries' choices.
Domestic: More/less social spending	High	Costs/benefits do not depend on other countries' choices.
Domestic: Permit/not press freedom	Moderate	Costs/benefits are primarily domestic, but costs of free media are higher if rival authorizes influence operations.
Domestic: Repress protesters/not	Moderate	Costs/benefits are primarily domestic, but rival's criticism or sanctions could increase repression's cost.
International: Travel to a summit meeting/not	Moderate	Greater benefits if rival makes policy concessions at summit, but benefits of demonstrating leadership may apply either way.
International: Institute tariffs/allow free trade	Low	The effect of either option on exporting firms depends greatly on whether other countries reciprocate.
International: Cut/build arms	Low	Cutting arms leaves state vulnerable if rival does not reciprocate. Arming is expensive, but confers military advantage if the other side does not build.

Note: High, medium, and low codings roughly approximate where choices fall on an independence continuum.

Crucially, independence *does not* mean that policy choices are costless. It 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, promoting behavior that reveals more information about type.

independence.

2.1.1 The Strategic Implications of Independence for Trust-Building

We introduce variation in payoff independence into the baseline two-period trust game described earlier. We conceptualize the level of independence of each state's choices using the continuous parameters $\beta_A, \beta_B \in [0, 1]$. 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, ensuring the second period is identical to the classic trust problem.[17]

We visualize the updated game with independence in Table 4. When β_A and β_B both equal 0, the model converges to the classic model of trust-building through international actions presented earlier (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.[18] Rather, A's first-period value from cooperation depends entirely on A's preferences.[19] When β_A and β_B are between these values, it means both states' trust-building payoffs are partially dependent.

Table 4: First-Period Payoffs in the Game with Independence

		Player B	
		Cooperate	Defect
Player A	Cooperate	1,1	$\beta_A - (1-\beta_A)k, e_B^m$
	Defect	$e_A^m, \beta_B - (1-\beta_B)k$	$\beta_A e_A^m, \beta_B e_B^m$

Note: These updated payoffs only apply to the *first period*. Second-period payoffs are as in Table 1. Both are equal when $\beta_A = \beta_B = 0$.

We then solve for all the Perfect Bayesian Equilibria (PBE) of our model in Appendix A.3

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

[18] 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.

[19] In each case, the effect on B's utility is symmetrical.

Table 5: Equilibrium Changes from Introducing Independence

Description	Difference from Standard Model Caused By Independence (β)
	Changes to equilibria we can sustain in the baseline model
Trust-building	As independence increases, the equilibrium expands into low trust levels. Once the independence threshold is met ($\beta_i > \frac{k}{1+k-L}$), trust-building is possible even at the lowest initial trust levels.
Suckers	No change
Tragic	When either side meets the independence threshold, we cannot support this equilibrium, *even without the efficiency refinement* under any parameter ranges. Security seekers deviate from $a_1 = d \rightarrow c$.
	New equilibria that emerge when β_A, β_B both exceed threshold (Figure 2)
Mixed strategy trust-building	When $\beta_i = 0$, this was Pareto dominated by the suckers equilibria. As independence increases, the equilibrium expands into low levels of trust. The mixing inefficiencies reduce greedy states' incentives for first-period cooperation, assuring mixed trust-building survives for higher p_i values than pure strategy trust-building.
Semi-tragic	This replaces the tragic equilibrium as the least players can assure themselves because security seekers now prefer cooperation knowing they will be cheated.
	New equilibria that emerge when independence is lopsided (e.g., β_A low, but β_B high) (Figure 3)
Asymmetric equilibria	The highly independent security-seeking state cooperates without expecting reciprocation. The highly dependent security-seeker defects. When trust is also lopsided, we can support an asymmetric trust-building equilibrium. When trust is jointly high, we can only support asymmetric semi-tragic equilibria because the greedy A pools in period one, which prevents trust building. Asymmetric equilibria can be Pareto optimal in overlapping ranges with symmetric equilibria.

under the assumption that the level of independence can vary.[20] Treating the baseline model as the counterfactual, we preview the effects of introducing and increasing independence in Table 5. The top row explains how introducing independent signaling actions alters the conditions for equilibria found in the baseline model. The second and third rows explain why new PBE-ER emerge when we introduce independence in the choices of one or both sides. As the table indicates, introducing independence has many nuanced implications. We can no longer support the commonly-studied tragic equilibria when independence is high; the trust-building equilibrium originally identified now survives under lower levels of initial trust, and several new equilibria, including mixed strategy and asymmetric trust-building equilibria, emerge.

We explore all these nuances in the technical appendices. Here we focus on establishing our

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

primary claim: Introducing independence facilitates trust-building under conditions of deep initial mistrust. To establish this claim as clearly as possible, we initially focus on the symmetric pure strategy trust-building equilibrium (shaded dark red in Figure 1). This was the only trust-building PBE-ER established in the baseline model, and it continues to survive as we introduce variation in independence, allowing for a clear comparison.[21] Proposition A.6, reported in the Appendix, formally characterizes the conditions under which we can sustain this equilibrium given that payoff independence can vary. Here we describe the empirical implications of increasing independence by contrasting when trust-building occurs under high independence, relative to the baseline condition. Results 1a and 1b described below follow from a comparative static analysis of the trust-building equilibrium, which is generally presented in section A.3.2.

Result 1a: Independence and trust-building given initial mistrust. Starting at the baseline ($\beta_i = 0$), as independence increases, the minimum level of initial trust (p_i) necessary to sustain a trust-building equilibrium decreases.

In the baseline model, we could not sustain trust-building equilibria when initial trust was low because the security seeker faced two strategic problems that prevented her from taking the initial trust-building action. Increasing independence increasingly resolves both of these problems. The direct problem was that the security-seeking A worried that B would cheat her. As the independence of A's choice increases (β_A), A's value is increasingly determined by her type, and less by how B reacts. Since security seekers directly benefit from cooperation, security-seeking A's value for cooperation is increasing with independence, regardless of concerns that B is greedy. The indirect problem was that the security-seeking A understood that B was deeply suspicious of her. Thus, even though A was a security seeker, A also understood that B's suspicion of A would drive even security-seeking B to defection. Increasing the independence of B's choice (β_B), indirectly incentivizes A to cooperate because A now believes that security-seeking B will cooperate in spite of B's initial mistrust. As in the baseline model, both states weigh the tradeoffs of being exploited in the first period against the potential value of cooperation in both periods. But as independence

[21] Focusing on the pure strategy equilibrium assures we are conservative in reporting how important independence is for facilitating trust-building. If we included the mixed strategy equilibria also, we would report more far-reaching effects.

increases, mistrustful security seekers are increasingly willing to risk cooperation because they are less sensitive to the cost of being exploited (the direct effect), and believe they are less likely to be exploited by their rival (the indirect effect).

We visualize the effect of increasing two-sided independence in Figure 2. The figure assumes intermediate levels of independence for both states ($\beta_A = .6, \beta_B = .7$), but otherwise holds the k, H, L values at the same levels as in Figure 1, allowing for direct comparison. Substantively, these intermediate levels of independence could reflect a scenario where both sides face domestic reform opportunities. Yet because not all independent choices are domestic, it could also reflect that one side faced a domestic choice and the other faced a reasonably independent foreign policy decision.

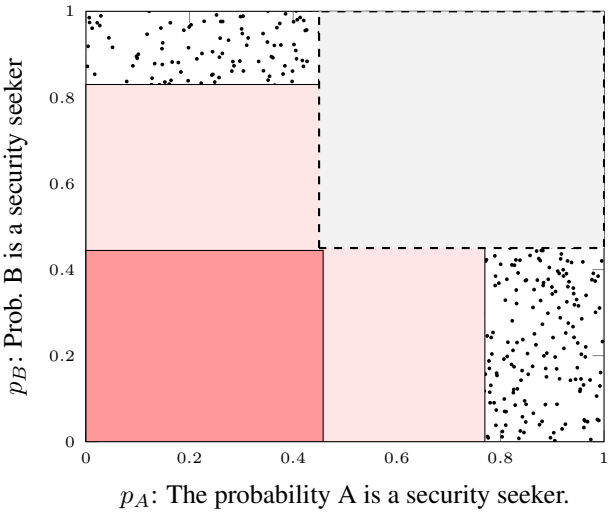
At the highest levels of initial trust, we still cannot sustain trust-building equilibria. The reason is that independence does not influence the greedy type's incentives to conceal his true preference, which is to cheat his rival in the second period. However, looking at the lowest levels of initial trust reveals a surprising result:

Result 1b: Independence threshold. When both players' choices are sufficiently independent (i.e., $\beta_A, \beta_B > \frac{k}{1+k-L}$), 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., $p_i \to 0$).

The independence threshold represents the point where the maximum payoff a security seeker can assure herself (i.e., her minmax) comes from selecting cooperation rather than defection.[22] Substantively, consider a case where a minor power A faces the decision to increase press freedom or not. State B (say the U.S.) could exploit this decision by meddling with a newly free press. But, in practice, B's policy choice has only a small effect on A relative to the domestic implications of allowing press freedom. Thus, A's value from promoting freedom does not depend much on how B reacts. Instead, A's decision hinges on whether A actually wants press freedom or not. In this case, we might say that A's independence threshold is met because even if A was so mistrustful of B that A was certain B would exploit A's new press freedom, that it would not influence A's decision. If A was the type that truly wanted these reforms, A would implement them and simply

[22]This also explains why we cannot sustain the tragic equilibrium.

Figure 2: Pareto Optimal Equilibria Given Two-Sided Independence



(a) Conditions where PBE emerge

Motive	Period 1	Period 2 | Period 1			
		CC	CD	DC	DD
Mixed strategy trust-building (light red)					
Security	C	C/D	D	D	D
Greedy	D	D	D	D	D
Semi-tragic (random dots)					
Security	C	D	D	D	D
Greedy	D	D	D	D	D

For the trust-building (dark red, solid line) and suckers (gray, dashed line) equilibrium descriptions, see Figure 1.

(b) Description of new equilibria

Note: As in Figure 1, we set $H = 1.5, L = -0.1, k = .9$ in panel (a). We additionally assume intermediate values $\beta_A = .6$, $\beta_B = .7$, which implies $\beta_A \sim \beta_B > \frac{k}{1+k-L}$. As in Figure 1, we shade trust-building dark red and suckers gray. Panel (b) describes the strategy profiles and color codings for the new PBE-ER that emerge. PBE-ER are unique in plotted parameter ranges.

accept that B would meddle.

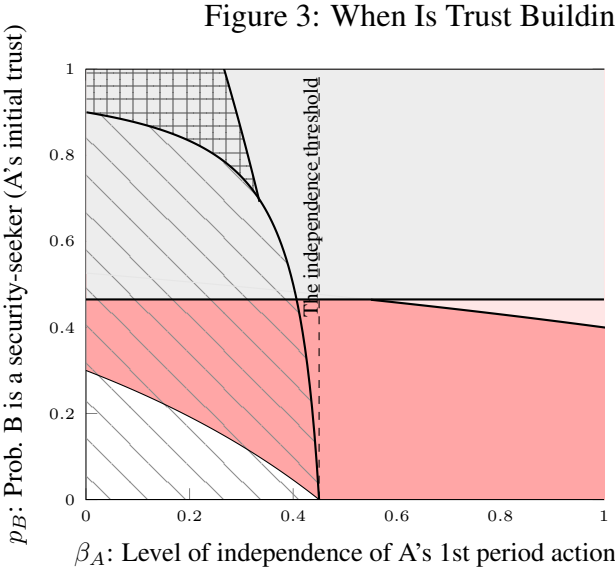
The threshold's existence demonstrates that Result 1a is not overshadowed by extreme initial mistrust. Rather, as independence increases, the strategic incentives to cooperate in the first period continuously increase for low-trust types until they dominate the concerns that cause security seekers to defect. The threshold is decreasing in the cost of being exploited (k) and the benefit a security seeker accrues from cheating her rival (L). While this threshold can vary, Figure 2 demonstrates that, under plausible costs and benefits, we meet it for even intermediate levels of independence.

Figure 2 assumes both sides' trust-building actions meet the independence threshold. In practice, one state may have an independent trust-building opportunity, but the other may not. This could arise if A is more sensitive to exploitation than B. For example, democracies can be exposed to election rigging by promoting press freedom, but autocracies cannot. It could also arise because different states have different available trust-building options. If one state starts with free markets, they cannot use market reform to engender trust. This might leave that state with only more dependent options, like arms reductions, to signal their security-seeking preferences. Can trust-building survive when independence is lopsided? Figure 3 considers this case. The plot assumes that B's trust-building action is moderately independent and B is moderately trusting of A (β_B, p_A are moderately high). Meanwhile, it allows A's level of independence (β_A) and initial trust in B (p_B) to vary. On the furthest left side of the plot, we observe the counterfactual to the baseline where, like in the baseline, $\beta_A = 0$. But unlike in the baseline, β_B is higher. As the plot moves to the right side, β_A increases, and the parameter ranges cross over those analyzed in Figure 2.

We note two points. First, as with prior plots, the symmetric trust-building equilibrium arises in the dark-red shaded area. Examining how this equilibrium expands as a function of β_A visualizes Result 1a in action. When A's decision is fully dependent ($\beta_A = 0$), there is a lower bound on the symmetric trust-building equilibrium. As β_A increases, this lower bound diminishes. A reaches the independence threshold when $\beta_A = .45$. From this level of independence onward, we can sustain symmetric trust-building at the lowest levels of initial trust.

Second, the bottom left corner of the plot reveals a surprising equilibrium:

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



(a) Conditions where PBE emerge

Player/Motive	Period 1	Period 2 | Period 1 CC	CD	DC	DD
Asymmetric trust-building (diagonal, striated)					
A/Security	D	C	D	C	D
A/Greedy	D	D	D	D	D
B/Security	C	C	D	C	D
B/Greedy	D	D	D	D	D
Asymmetric semi-tragic (thatched)					
A/Security	D	D	D	D	D
A/Greedy	D	D	D	D	D
B/Security	C	D	D	D	D
B/Greedy	D	D	D	D	D

(b) Equilibrium description

Note: Continues to assume $k = .9, H = 1.5, L = -.1$. Assumes B's decision is moderately independent $\beta_B = .7$, and B has moderate initial trust $p_A = .5$. Plots PBE-ER given variation in the independence of A's choice and initial trust. As in Figures 1 and 2, we shade trust building dark red, mixed strategy light red, and suckers gray. The new thatched space is the asymmetric semi-tragic equilibrium (panel b). The new diagonal striated space is the asymmetric trust-building equilibria (panel b). Where the lines and shading intersect, there are multiple PBE-ER. Inefficient (including mixed strategy) equilibria are reported in the Appendix.

Result 1c: Asymmetric trust-building. When trust and independence are lopsided (e.g., β_A, p_B small but β_B, p_A large) we can sustain asymmetric trust-building equilibria. When the scope of lopsidedness is extreme ($\beta_A, p_B \to 0$), this equilibrium is unique.

This is a trust-building equilibrium because security-seeking A conditions second-period cooperation on B's first-period choice, and A infers B is trustworthy if B cooperates in the first period. However, it carries several interesting properties. First, B knows that A always defects in the first period. Still, if B is a security seeker, B accepts cooperation with certain exploitation. Second, first-period exploitation is necessary to forge trust and cooperation in the second period. If B does not accept exploitation, A infers that B is greedy, and second-period mutual defection is assured.

Sustaining this equilibrium requires lopsidedness in both independence and trust. State A must have a sufficiently dependent choice and be sufficiently distrustful that it will never cooperate in the first period. In contrast, B's choice must be sufficiently independent that security-seeking B will tolerate certain exploitation. B must also have at least moderate levels of initial trust. The reason is that B does not learn anything from A's first-period decision because A defects regardless

of A's type. Thus, if B starts out too suspicious, then B is unwilling to cooperate in the second period. But B's initial trust cannot be too high either. If it is, then the suckers dynamic takes hold and greedy types play cooperation to cheat A in the second period.

So far, we have found unique PBE-ER. But Figure 3 shows some overlap between efficient asymmetric and symmetric equilibria. The reason is that while both sides do equally well in a symmetric equilibrium, all asymmetric equilibria assure one side initially exploits the other. The exploiter prefers asymmetric equilibria and the exploited prefers symmetric equilibria.

Overlapping equilibria can complicate empirical predictions. Here they strengthen our overall claim. To summarize what we have found, when trust-building decisions are dependent (as assumed in previous work), we cannot sustain trust-building PBE if at least one side has low levels of initial trust. However, when trust-building decisions are at least moderately independent, not only can we sustain trust-building equilibria at low levels of initial trust, but only trust-building equilibria survive refinement. When both sides face independent choices, there is a clear prediction given low initial trust: symmetric trust-building. When the trust-building options available to both sides are lopsided, the result is more complicated. We often sustain both asymmetric and symmetric trust-building equilibria. But we do not find non-trust-building equilibria survive refinement.[23] Thus, either way, our model predicts some form of trust-building even in lopsided cases with low levels of initial trust.

2.2 Salience of the Trust-Building Exercise

Existing trust-building work considers variation in the salience of different international policies. 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,

[23] The only other equilibrium is substantively unappealing. Even though all types reveal their true motives in period one, mutual defection is assured in period two.

in principle, can always find a right-sized signal to build trust when starting from any initial trust level.

Skeptics argue it is difficult to finely calibrate international gestures to satisfy this "Goldilocks" problem (Lieber 2011; Rosato 2014; Montgomery 2006). Goldilocks problems can be exacerbated by problems with interpretation. It is often hard to notice if a rival dismantles a few missiles given regular changes in force posture, and even harder to interpret the salience of such a gesture. These interpretation problems likely impede trust-building between enduring rivals because the calibration window is small when initial trust is low (see Yoder and Haynes 2021).

We now show that trust-building through independent (usually domestic) choices does not succumb to calibration problems. We introduce a salience parameter, $\theta > 0$, into the model. Larger values of this continuous parameter indicate that the second-period choice is increasingly important relative to the first-period choice. For example, when $\theta = 2, 1/2, 1$, it means that the second-period choice is twice as, half as, and equally important as the first-period choice, respectively. 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 2 If the independence threshold described in Result 1a ($\beta_i > \frac{k}{1+k-L}$) holds, then there is no Goldilocks problem because the security-seeker prefers first-period cooperation for any level of salience, even at very low levels of initial trust (i.e., there is no upper bound on the salience of first-period choices).

This finding has two substantive implications that contrast international and domestic choices. First, trust-building via domestic (or other independent) choices entails an easier calibration task than typically-studied international trust-building. Given sufficient independence, no first-period decision 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 classic international signaling, states can potentially build trust quickly through high-stakes domestic choices. Second, domestic and international actions can work together. States can build initial trust via domestic choices when calibration problem make international signaling initially intractable. Since domestic choices are imperfectly correlated with international preferences, they may not com-

pletely resolve the trust problem. But they may increase trust enough to make international signals, such as arms control agreements, viable.

3 Implications for Ending Enduring Rivalry: The Cold War

Our theory of trust-building is broadly applicable. But to test it, we focus on episodes that existing research struggles to explain: ending enduring rivalries; and illiberal and asymmetric trust-building. We first consider enduring rivalries. Existing research lacks a satisfying explanation for how deeply mistrustful rivals can find the right-sized gestures to initiate trust-building. We illustrate empirically how independent domestic choices help rivals overcome this problem by detailing the end of the Cold War. In the next section, we examine asymmetric trust-building via illiberal domestic choices.

The Cold War case has three notable advantages. First, it is historically consequential. Second, successful trust-building was unexpected at the time. 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 staunch 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.

Third, previous research has used the Cold War 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 "first important costly signal" that Moscow had security-seeking intentions (Kydd 2005, 227). In this account, the process of reassurance was initiated by INF and reinforced primarily by additional Soviet foreign policy choices. While acknowledging the impor-

tance of INF, we argue that Soviet domestic policies, such as improved treatment of dissidents, *glasnost*, and *perestroika*, kick-started the trust-building process even earlier. Analyzing this case illuminates how domestic and international policies work together, with domestic policies playing a potentially necessary preliminary role. Like Kydd, we focus on Soviet trust-building actions and Washington's trust of Moscow. Section 3.4 briefly considers Western trust-building actions and Soviet trust.

Following best practices in case evaluation of models (Gerring 2004), the next section argues that the initial conditions of the case match the symmetric trust-building equilibrium's parameter ranges. That is, we could not rationalize trust-building through highly dependent international choices in this case because initial trust was too low, but we can rationalize trust-building through more independent choices, notably Soviet domestic reforms. We then trace critical elements of our causal mechanism through the case (Lorentzen, Fravel, and Paine 2017; Joseph, Poznansky, and Spaniel 2022).

3.1 Mapping the Model to the Case

In the Cold War case, we code the goal of security seekers as rapprochement. Thus, second-period mutual cooperation would involve arms control and other security and economic agreements (ending the Cold War). Defection would involve continuing or intensifying the arms race or security competition, to include reneging on arms control or other agreements. We argue that the case conditions match the parameter ranges for symmetric trust-building (Appendix B maps the model to the case in tabular form). We code both sides as genuinely security seeking by the late Cold War, consistent with classic research on the effects of both "New Thinking" and structural constraints (Risse-Kappen 1994; Brooks and Wohlforth 2000; English 2000).

Even though both states desired rapprochement, we code trust between the U.S. (p_{USSR}) and Soviet Union (p_{US}) as very low in the early 1980's.[24] In his first press conference as commander-in-chief, Reagan stated that the Soviets "reserve unto themselves the right to commit any crime, to

[24]This matches Kydd (2005).

lie, to cheat," to achieve their goals (Gwertzman 1981). Soviet General Secretary Leonid Brezhnev, in 1980, labeled the U.S. an "unreliable partner." (FRUS 2018, 1977–1980, Volume XII, No. 166). Even the reform-minded Gorbachev was dismayed by U.S. defense policy, claiming in 1985 that the Americans "promise the world stability but in reality strive to wreck the military balance" (Eaton 1985).

Given mutual distrust between states that genuinely desire second-period mutual cooperation, our theory predicts that a symmetric trust-building equilibrium can emerge if both rivals have sufficiently independent and salient trust-building choices to make. By contrast, if only dependent trust-building choices were available, then they would not make them. Although Washington and Moscow had discussed arms control from the early 1980's onward, the benefits had low independence (low β_i), creating an impediment to progress under low trust. In contrast, the Soviets' domestic policy choices had moderate-to-high independence (moderate-to-high β_{USSR}). The payoffs from more lenient treatment of dissidents, greater free speech, and reduced central planning of the economy (the Soviets' first-period choice in our model) were not greatly dependent on Western policies. For example, the main benefits that Soviet leaders expected from *glasnost*, a policy of greater openness and freedom of speech, were reduced corruption and increased efficiency. These benefits did not depend much on any Western responses. We also classify Soviet domestic choices as high salience. 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 Kremlin's governing ideology and political control.

Finally, 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 Bloc. Soviet domestic reforms reflected a shift in intrinsic values (English 2000). Though the Reagan Administration would have preferred the USSR to become a true democracy, even liberalizing reforms within an autocratic context indicated greater compatibility in international preferences. 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 the characteristics necessary to launch trust-building.

3.2 Initial Trust-Building through Soviet Domestic Choices

East-West trust in the early 1980s could hardly have been lower. President Reagan 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). Similarly, in 1983, Secretary of State George Shultz stated that "Soviet actions have come into conflict with many of our objectives" and lamented Moscow's penchant for "stretching a series of treaties and agreements to the brink of violation and beyond" (FRUS 2021, 1981–1988, Volume IV, Soviet Union, January 1983–March 1985, No. 61).

In 1985, Mikhail Gorbachev became General Secretary of the Communist Party of the Soviet Union. Gorbachev and his advisors desired to end the costly U.S.-Soviet competition (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 1980s were often more about public relations than genuine attempts to build trust (Colbourn 2022). Though the Soviets, especially by the time Gorbachev reached office, greatly desired reductions in defense spending, unilateral arms reductions would have raised security concerns and potentially damaged Soviet prestige. Therefore, arms control could not proceed.

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 trustworthy partner for the West. A notable case involved Andrei Sakharaov,

the Soviet nuclear physicist and Nobel Peace Prize-winning dissident. According to Shultz, the U.S. point person on Soviet diplomacy, Sakharov's release from internal exile in 1986, "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 reinforced 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 views:

> [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 the CIA 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 eventually pursued broader reforms, *glasnost* and *perestroika*, to increase 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, a "bombshell" that would "change our social system" (Taubman 2017, 248). 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).

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, made him hopeful about Washington-Moscow relations (Matlock 2004, 294).

3.3 International-Level Trust-Building Begins

Soviet domestic reforms not only built trust, but facilitated international cooperation requiring more dependent choices, including the INF Treaty. The conclusion of the INF Treaty reflected, to be sure, important concessions by Gorbachev (e.g., including the SS-23 in the deal, de-linking missile defense). However, its signing was aided by the trust 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). To whet Reagan's appetite for a summit in Washington to sign the INF Treaty, Shultz reported from Moscow that "the Soviet Union is changing" (Leffler 2007, 399).

When Gorbachev eventually came to Washington to sign the INF Treaty in December 1987,

[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 Reagan officials' initial certainty of malign Soviet intentions.

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, then-U.S. ambassador to Russia, was "electrified" when he read it (Savranskaya et al. 2010, 110). The following day, the ambassador told Reagan that "the Soviet Union will never be the same" (Matlock 2004, 296). Just days later, in Red Square, Reagan declared that he no longer saw the Soviet Union as an "evil empire."

The 1988 19th Party Conference proved another milestone in the Kremlin's reform program. Though falling far short of democracy, the conference resulted in political liberalization unprecedented in the Soviet context. Gorbachev 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 cemented his reforms by removing several conservatives from high-ranking posts. 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 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 reform was perhaps the earliest indicator 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. By increasing trust, Soviet domestic reforms thus contributed to an unprecedented level of East-West cooperation.

3.4 View from the Soviet Union

On the other side, how did U.S. and Western policies contribute to Soviet trust? Here we briefly highlight one important choice: The Western response to Soviet economic and political reforms.[26] Gorbachev came to power in 1985 with security-seeking intentions, but nonetheless worried that leaders like Reagan were too stuck in the Cold War mindset to be partners in peace.[27] When Gorbachev initiated his domestic reforms to revitalize the Soviet Union, the West might have wholly dismissed these efforts. Instead, Western governments largely encouraged these reforms, which reassured the Soviets of their trustworthy intentions.

We classify this Western choice as moderately independent (β_{US}). The choice to respond positively to Gorbachev would bring more benefits if Gorbachev followed through on reforms. However, even if he had not, the U.S. and its partners would still have gained important benefits. They would have seemed magnanimous on the world stage, demonstrating that they were not implacably hostile to the Eastern Bloc. In addition, anti-Communist leaders like Reagan and Thatcher would, in particular, benefit from showing they were not recklessly hawkish (Schultz 2005; Goldfien 2025). Indeed, public approval for Reagan's Soviet policy surged when he pivoted to a more moderate stance (Nincic 1988). Therefore, this choice—while more dependent

[26]This was an international choice, which we argued above can sometimes be sufficiently independent to support trust-building under low initial trust. This highlights that, ultimately, independence rather than domestic policy is key.

[27]Gorbachev initially saw Reagan as "stubborn" and a "dinosaur" (Matlock 2004, 169).

than Soviet domestic reforms—was still independent enough that Western leaders like Reagan and Thatcher were willing to make it even in the context of substantial distrust. We also classify the choice as moderately salient (θ_{US}) because it concerned the West's orientation toward its primary geopolitical rival and affected the political standing of Western leaders. Since this was an international choice, the Soviet leadership could also be confident that it correlated strongly with future willingness to cooperate internationally.

Ultimately, the choice to encourage, rather than dismiss, Soviet domestic reforms succeeded in reassuring the Kremlin that the U.S., U.K., and other Western countries could be partners in international cooperation. For example, as Gorbachev considered further liberalization of the Soviet economy in early 1987, George Shultz—who had earned a PhD in Economics from the University of Chicago and run Bechtel engineering company—offered counsel. According to one historian, "Distinctions between Milton Friedman and John Kenneth Galbraith were less important than an American secretary of state's message of two states confronting common challenges, aspiring for common results, and establishing greater trust. Gorbachev and those around him began to sense that the most conservative and anticommunist presidential administration of the Cold War was actually out to help them," not out to get them (Wilson 2014, 133).

The West's other arch-capitalist leader, Margaret Thatcher, also supported Gorbachev's reforms (Brown 2020). Following a visit to Moscow in early 1987, Thatcher wrote to Reagan, "I am firmly convinced it is in our interest to encourage him [Gorbachev], especially in his endeavours to create a much more open society."[28] This attitude also contributed to Gorbachev's confidence in Western intentions. Writing about Thatcher's contributions to the end of the Cold War, Gorbachev reflected that she "was genuinely interested in what was happening in our country. She closely, and with astonishing command of detail, followed perestroika and glasnost, and sincerely wished for our process of change to succeed." This attitude led Gorbachev to assess that Thatcher—echoing the British prime minister's famous judgment of him—was a "person one can deal with" (Gorbachev 2013).

[28] See "Letter to Reagan from Thatcher About Her Meetings with Gorbachev in Moscow. April 1, 1987," accessed via the National Security Archive, `https://nsarchive.gwu.edu/document/21546-document-09`.

Our analysis of the U.S.-Soviet case shows that it is possible for domestic and international choices to work in tandem to build trust, even as the role of domestic choices is particularly crucial. It also offers a rejoinder to accounts that critique the U.S. for being too slow to build trust and cooperate with Gorbachev (i.e., Braumoeller 2013). Our concept of independence explains why trust-building likely required starting with reforms in the Soviet Union rather than ambitious arms control agreements. Yet we show that the U.S. was not simply passive before the INF Treaty was signed in late 1987. Rather, U.S. and Western encouragement of Gorbachev's domestic reforms was a meaningful and conscious policy choice that effectively reassured the Soviet leader and promoted further East-West cooperation.

4 Asymmetric and Illiberal Trust-Building: U.S.-South Korea Relations following the 1961 "Military Revolution"

We now use our model to illuminate patron-client relations, analyzing trust-building between the U.S. and South Korea following Park Chung-hee's May 1961 coup, after which the new government in Seoul reassured Washington of its pro-Western alignment via a crackdown on domestic communists and an authoritarian modernization program. This case differs from the U.S.-Soviet case in two main ways. First, whereas the U.S.-Soviet case fit the parameter ranges for symmetric trust-building, the initial conditions in this case support the asymmetric trust-building equilibrium. While the new South Korean regime had little doubt about the U.S. commitment to supporting anti-Communist governments, Washington was uncertain whether the newly established regime would tilt toward capitalism or Communism. Second, whereas the U.S.-Soviet case featured reassurance through liberalization, the U.S.-South Korea case illustrates the potential for illiberal policies to reassure.

While focusing only on the South Korea case, we conjecture that authoritarian client states frequently reassured the U.S. of their reliability as Cold War allies with illiberal domestic policies. As historians such as David Schmitz have highlighted, American policymakers often saw right-

wing dictatorships that would keep leftists in check as an attainable outcome far preferable to Communist rule (see, e.g., Schmitz 1999, 2006).

4.1 Mapping the Model to the Korea Case

When a group of military officers ousted the democratically elected government of South Korea in early 1961, it threw U.S.-South Korea relations into uncertainty. To establish the potential for trust-building to resolve this uncertainty, we begin by mapping the situation to our model (see Appendix B for a more formal mapping of how the case corresponds with the equilibrium plotted in the bottom-left of Figure 3). Since the U.S. was staunchly anti-Communist, we define security-seeking intentions as a shared anti-Communist orientation. Therefore, the goal for cooperation among security seekers is blunting the spread of Communism.

We code U.S. trust in the new South Korean regime as low (low p_{ROK}). Even before the coup, Washington had worried about societal malaise and Communist influence in South Korean. The coup plotters were not immediately seen as preferable to the deposed government of Prime Minister Chang Myon (Brazinsky 2009). Indeed, their intentions were questioned. The coup leader, Park Chung-hee, had prior Communist ties, leading one U.S. 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 1996, 1961–1963, Volume XXII, Northeast Asia, No. 224). However, trust was lopsided. Park's military government believed that it could trust in U.S. support provided that it was seen as anti-Communist and committed to real reform and modernization (high p_{US}), even as it understood that the U.S. did not yet trust it (Kennedy 1988, 214).

The independence of the choices available to the two sides was also lopsided. The most meaningful way the United States could signal its willingness to cooperate was to give diplomatic recognition and aid to the new government. This choice was highly dependent (low β_{US}) because it would be embarrassing for the U.S. and a waste of resources if Seoul ultimately joined the Communist camp. In contrast, the main choices that the new Korean government had available to signal their anti-Communist credentials were domestic. Because Park's regime would reap the domestic

rewards of these policies regardless of what the U.S. did, they can be considered to have moderate-to-high independence (moderate-to-high β_{ROK}). For both sides, the choices available were quite salient and likely to be strongly correlated with a future desire for anti-Communist cooperation.

Given the lopsided nature of both trust and the independence of choices, our theory predicts that an asymmetric trust-building equilibrium would be most likely to emerge in this case. Specifically, we would expect Seoul to cooperate by pursuing anti-Communist domestic policies, while Washington defected by withholding recognition and aid in the first round. This initial asymmetric cooperation would build enough trust to facilitate mutual cooperation in the subsequent round.

4.2 Asymmetric Trust-Building and International Cooperation

Consistent with expectations, an asymmetric trust-building period ensued after the coup, in which the U.S. withheld official recognition and support from the new regime, while the Park government engaged in domestic reforms that proved appealing to Washington. The U.S. understood that recognition and support for the new military government would confer legitimacy, the benefits of which would be highly dependent on the character of the new regime. Thus, American officials in Korea quickly disavowed any support for the coup (FRUS 1996, 1961–1963, Volume XXII, Northeast Asia, No. 213), and Washington adopted a "cautious attitude of wait-and-see" (FRUS 1996, 1961–1963, Volume XXII, Northeast Asia, No. 216).

For the Park regime, anti-Communism at home and reforms aimed at developing the economy and rooting out corruption were generally independent. The benefits of these reforms largely derived from the Park government's own values. Though the Park government sought American support, it also had strong views of its own about how South Korea should be organized (Brazinsky 2009). Park and his government undertook anti-Communist policies and swiftly, if undemocratically, implemented modernizing reforms across South Korea's economy and society.

Park's early moves greatly encouraged Washington. In a telegram from the U.S. embassy in Seoul in October 1961, the reassurance felt by American officials is palpable. Ambassador Samuel Berger wrote that the regime "has taken hold with energy, earnestness, determination and imag-

ination, albeit with certain authoritarian and military characteristics" (FRUS 1996, 1961–1963, Volume XXII, Northeast Asia, No. 244). Further, "[v]igilence against communist subversion and quality and volume of anti-communist propaganda have greatly improved." Berger concluded that the new regime "offers much hope" (FRUS 1996, 1961–1963, Volume XXII, Northeast Asia, No. 244). While it is unclear whether the U.S. ultimately preferred the Park government to its predecessor, it is clear that Park's actions increased U.S. trust relative to the beginning of his tenure.

Efforts to improve economic performance—even by non-democratic means—and a demonstrated anti-Communist orientation not only built trust with the U.S. but facilitated mutual cooperation going forward. The positive impression that the military government had made early in its tenure led the Kennedy administration to host Park at the White House in late 1961, which "played a significant role in stabilizing and legitimizing the South Korean regime" (Brazinsky 2009, 120-21). Notably, this is the sort of high-dependence action that can be hard to contemplate under low trust. Looking ahead, the U.S. could—despite the Park regime's "shortcomings," i.e., authoritarian style—expect that U.S. aid would "be more effectively used than by any previous government" and that the administration could "go to Congress this spring in good conscience. . . that our continuing massive support is well justified" (FRUS 1996, 1961–1963, Volume XXII, Northeast Asia, No. 244). For its part, the Park regime cooperated with the U.S. on the international stage by, for example, hosting U.S. troops on its soil and sending South Korean soldiers to support the American war effort in Vietnam. Thus, initial asymmetric trust-building led to eventual mutual cooperation.

5 Conclusion

We argued that domestic policy choices can operate as costly signals that engender international trust, and even 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 on situations where initial trust is so low that security seekers are unwilling to signal

their motivations via international choices because they fear exploitation. Given this 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 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 several contributions. First, it contributes to our understanding of 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 the potential for asymmetric trust-building and how even illiberal actions, such as suppressing left-wing groups, can increase trust in some circumstances, a puzzle that is not explained by the democratic peace approach.

Our findings also speak to policy issues, including evolving Sino-American competition. 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; Joseph 2026). This is concerning because major power rivalries are often long and very costly (Thompson 2001). Tragically, it is hard to find a path back to peace even if both sides tire of competing. We identify domestic policy reforms as a mechanism the U.S. and China may eventually use to kick-start trust-building.

We also clarify how the United States can identify which states are trustworthy international

partners. As rising populism brings new forms of government to power and political polarization calls into question the intentions and abilities of democracies (Myrick 2021; Joseph, Chung, and Park 2026), the assumption that the U.S. should trust states that ascribe to liberal values and mistrust those that do not may generate both misplaced trust in democracies and misplaced competition with cooperative autocratic regimes. When considering both formal and informal alliances (Kenwick and McManus 2021; McManus and Nieman 2019), our theory explains how the U.S. can make a fine-grained analysis of the international reliability of new regimes based on their domestic policies. We recommend using this approach to consider arms sales, diplomatic recognition, and military support for middle powers, which could be risky as competition with China intensifies.

References

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

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

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.

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.

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.

Braumoeller, B. F. (2013). *The Great Powers and the International System*. New York: Cambridge Univeristy Press.

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. (2020). *The human factor: Gorbachev, Reagan, and Thatcher, and the end of the Cold War*. Oxford University Press.

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.

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.

Eaton, W. (1985, April). Gorbachev agrees to summit talks: Stops deploying soviet missiles until november. *Los Angeles Times*. Accessed via Los Angeles Times online archives.

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

English, R. (2000). *Russia and the Idea of the West: Gorbachev, Intellectuals, and the End of the Cold War*. Columbia University Press.

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 (1996). *Foreign Relations of the United States*. Washington, DC: United States Government Printing Office.

FRUS (2018). *Foreign Relations of the United States*. Washington, DC: United States Government Printing Office.

FRUS (2021). *Foreign Relations of the United States*. Washington, DC: United States Government Printing Office.

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., M. Joseph, and D. Krcmaric (2023, 8). When do leader backgrounds matter? evidence from the president's daily brief. *Conflict Management and Peace Science*.

Goldfien, M. A. (2024). Just patronage? familiarity and the diplomatic value of non-career ambassadors. *Journal of Conflict Resolution 68*(7-8), 1417–1442.

Goldfien, M. A. (2025). To agree or not to agree: Hawks, doves, and regime type in international rivalry and rapprochement. *International Security 50*(2), 162–192.

Goldfien, M. A. and M. F. Joseph (2023, 3). Perceptions of leadership importance: Evidence from the cia's president's daily brief. *Security Studies 32*, 205–238.

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.

Gorbachev, M. (2013). Mikhail gorbachev: the margaret thatcher i knew. *The Guardian*.

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.

Joseph, M. F. (2021). A little bit of cheap talk is a dangerous thing: States can communicate intentions persuasively and raise the risk of war. *Journal of Politics 83*(1).

Joseph, M. F. (2026). *The Origins of Great Power Rivalries: A Rational Theory of Principled Motivations, and Historical Context*. Cambridge University Press.

Joseph, M. F., J. H. Chung, and H. S. Park (2026). Do elite partisan disagreements degrade the probability of victory in war? evidence from south korean military battle experiments. *American Political Science Review*.

Joseph, M. F. and M. Poznansky (2024, 11). Secret innovation. *International Organization 78*, 766–799.

Joseph, M. F., M. Poznansky, and W. Spaniel (2022, 4). Shooting the messenger: The challenge of national security whistleblowing. *The Journal of Politics 84*, 846–860.

Kennedy, C. (1988). Oral history interview with marshall green. Association for Diplomatic Studies and Training, Foreign Affairs Oral History Project.

Kenwick, M. R. and R. W. McManus (2021). Deterrence theory and alliance politics. In S. M. Mitchell and J. A. Vasquez (Eds.), *What Do We Know about War? 3rd Edition*. Lanham, Maryland: Rowman and Littlefield.

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.

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.

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

McManus, R. W. and M. D. Nieman (2019). Identifying the level of major power support signaled for protégés: A latent measure approach. *Journal of Peace Research 56*(3), 364–378.

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.

Myrick, R. (2021, 4). Do external threats unite or divide? security crises, rivalries, and polarization in american foreign policy. *International Organization 75*, 921–958.

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

Nincic, M. (1988). The united states, the soviet union, and the politics of opposites. *World Politics 40*(4), 452–475.

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

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

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

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.

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.

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

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

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.

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.

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.

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.

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 two sequential trust games 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 can hold one of two motivations $m \in \{s, g\}$. Where $m = g$ is greedy and $m = s$ is security-seeking. We sometimes notate i_m meaning player i has motive m. We will discuss payoffs in more detail in a moment, but motivations determine the value states accrue from exploiting the other side. The greedy type gets a high value from exploitation $e_i^g = H$, and the security seeker gets a low value from exploitation $e_i^s = L$ relative to the value of mutual cooperation (we normalize the value of mutual cooperation to 1). Define $p_i = pr(m = s \implies e_i = L)$ as the probability player i is a security-seeker, and $1 - p_i = pr(m = g \implies e_i = H)$ as the probability i is greedy.

The sequence of moves is as follows.

- Nature draws player types i.i.d.,
- Period 1: players simultaneously choose between cooperation and defection.
- Players observe the result of Period 1.
- Period 2: players simultaneously choose 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, $\sigma_1^i = p_j$.

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 to capture the commonly studied trust problem. First, all types prefer mutual defection to being cheated (i.e, k is the cost of being cheated):

$$\mathcal{A}_1 \qquad k > 0$$

Second, the greedy type prefers to exploit a rival over mutual cooperation, however the security seeker prefers mutual cooperation over exploitation

$$\mathcal{A}_2 \qquad H > 1 > L$$

The first-period payoffs are represented in Table 4. Note the values in Tables 1 and 4 converge when $\beta_i = 0$.

To understand the role of the dependence parameter, we detail some payoffs. 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 will 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 $\sigma_2^j | i_1 = d \leq p_i$. In any equilibrium strategy profile with an on-path pooling first period action $i_1 = d$, restrict off-path beliefs $\sigma_2^j | i_1 = c \geq p_i$.

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.

Finally, while we will solve for all PBE, we will ultimately restrict our attention to efficient equilibria. There are lots of ways to apply efficiency criteria for staged, incomplete information games. We eliminate an equilibrium if at least one type of one player prefers to deviate to another equilibrium, while no player does worse given their ex-ante total expected utilities.

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 $\omega_B^* = pr(b_2 = c) = \frac{k}{\sigma_2^A(1-L+k)}$, and $\omega_A^* = pr(a_2 = c) = \frac{k}{\sigma_2^B(1-L+k)}$.

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* $\sigma_2^i > \frac{k}{1+k-L}$ *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* $\sigma_2^i > \frac{k}{1+k-L}$ *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 $H > 1$. Conjecture $a_2 = c, b_2 = d$. A_g can profitably deviate to defect if $0 > -k$.

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 $-k < 0$. 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 $\sigma_2^A + (1 - \sigma_2^A)(-k) > \sigma_2^A L$, 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

$$\sigma_2^A \omega_B + (1 - \sigma_2^A \omega_B)(-k) = \sigma_2^A \omega_B L$$

This solves for $\omega_B = \frac{k}{\sigma_2^A(1-L+k)}$. This can be solved within 0, 1 so long as $\sigma_2^A > \frac{k}{1+k-L}$. 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. $EU_2^A = 0$
2. $EU_2^A|A_s = \sigma_A - k(1-\sigma_A)$, $EU_2^A|A_g = \sigma_A H$
3. $U_2^A|A_s = \frac{Lk}{1-L+k}$, $U_2^A|A_g = \frac{Hk}{1-L+k}$

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.

Later we will exclude Pareto dominated equilibria given interim beliefs.

Remark If $\sigma_2^i \geq \frac{k}{1+k-L}$, then the strategy profile characterized in bullet 2 is Pareto dominant. Otherwise, the strategy profile characterized in bullet 1 is unique.

Note, all expected utilities are increasing in σ_2^i if $\sigma_2^i = \frac{k}{1+k-L}$ is non-negative. Now note that given the equilibrium condition, the utilities in bullet 3 are positive. This assures that if we can support the mixed strategy equilibria, that the strategy profile characterized in bullet 3 dominates 1. Turning to the contrast between 2 and 3. Subbing in the boundary condition $\sigma_2^i = \frac{k}{1+k-L}$, A_s, A_g's utilities are identical in bullet's 2 and 3. This assures that 2 dominates 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 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* $p_i > \frac{k}{1+k-L} \sim 1 - \frac{1}{H}$ *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:*** *If* $p_i > \frac{k}{1+k-L}$ *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* $p_i > \frac{k}{1+k-L}$ *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: $1 + p_B H + (1 - p_B)0 > H + 0$. This solves for $p_i > 1 - \frac{1}{H}$, 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, $\sigma_2^A = p_B$. Thus, A prefers cooperation to defection in period 2 if $p_i > \frac{k}{1+k-L}$. This requires that $1+k-L > 0$. Working backwards, in the first period, A_s prefers cooperation over deviating to defection if: $1 + p - k(1 - p) > L$. This re-arranges to $p > \frac{L+k-1}{1+k}$. Note, $\frac{L+k-1}{1+k} > \frac{k}{1+k-L} \equiv (k + 1)(k + (L - 1)^2)(1 + k - L) < 0$ which cannot be satisfied if $1 + k - L > 0$. Note we can sustain the equilibrium for any off-path beliefs. The reason is that the only off-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 $\sigma_{2A} = p_B$. Lemma A.1.2 solves the second-period strategy. This gives is the condition $p_i > \frac{k}{1+k-L}$. 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 $k > 0$. B's strategy is symmetric. This completes the proof.

We note a second no learning equilibrium held together by a different off-path punishment: $s^A(a_1 = d, a_2 = c|A_s \;\&\; a_1 = b_1 = d, a_2 = d|$ otherwise$), s^B(b_1 = d, b_2 = c|B_s \;\&\; a_1 = b_1 = d, b_2 = d|$ otherwise$)$. Focusing on A_s, the same second period constraint binds. In the first period, there is no profitable deviation if $1 + p_B - k(1 - p_B) > -k \equiv p_B(1 + k) > 0$.

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 $p_B H > p_B - (1-p_B)k \implies H > 1 - k(1-p_B)/p_B$, always true. A_s prefers cooperation if $p_B - k(1-p_B) > p_B L$, which solves for the condition, 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*

$$\frac{k}{1+k} \geq p_i \geq \frac{k}{2+k-L} \tag{1}$$

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:

$$p_B H \geq p_B(1 + H) - k(1 - p_B) \equiv p_B \leq \frac{k}{1+k}$$

The security A plays on path if:

$$p_B(1 + 1) - k(1 - p_B) \geq p_B L \equiv p_B \geq \frac{k}{2+k-L}$$

Turning to existence, note, $\frac{k}{1+k} > \frac{k}{2+k-L}$ if $1 > L$, true by assumption. 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: $s^A(a_1 = c|A_S, a_1 = d|A_g, a_2 = c|A_s \& a_1 = c \& b_1 = d, a_2 = d$ otherwise), $s^B(b_1 = d, b_2 = c|B_s \& a_1 = c \& b_1 = d, b_2 = d$ otherwise). 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: $-k + p_B - k(1 - p_B) > 0$, solves for $p_B > \frac{2k}{1+k}$. A_g cannot profitably deviate from $a_1 = d \to c$ if: $0 > -k + p_B H$, solves for $\frac{k}{H} > p_B$. These are jointly solvable if $k > 2H - 1$. Setting $k = 2H - 1$, $\frac{2H-1}{H} > p_B \implies 2 - 1/H > p_B$, which cannot hold because $H > 1$.

Profile 2 is: $s^A(a_1 = d, a_2 = c|A_s \& b_1 = c, a_2 = d$ otherwise), and $s^B(b_1 = c|B_S, b_1 = d|B_g, b_2 = c|B_s \& b_1 = c \& a_1 = ., b_2 = d$ otherwise). Here $a_1 = .$, emphasizes that b_2 holds even given A's off-path deviation. The ICC is identical to the first profile.

A.2.1 A comment on mixed strategy equilibria

To be clear, there are many mixed strategy equilibrium and even mixed strategy trust-building equilibria. We omit them from this analysis of the special case of $\beta_i = 0$ because (a) they do not alter our basic conclusions (which relates to trust-building when p_i is low); (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 that survive the same parameter ranges. In section A.3.6, we will fully specify all the mixed strategy equilibria for the complete model (note the analysis thus far has only considered the special case $\beta = 0$, but the proof of the main model below $\beta_i \in [0, 1]$ will include this special case). Here we provide a preliminary analysis to demonstrate why they are both inefficient and also cannot alter our core result.

Recall, our main claim is that trust-building equilibria do not arise when p_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 p_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 we cannot support them for levels of p_i that are lower than the trust building equilibrium above. The reason is that the lower bound on p_i to support the contingent equilibrium is defined by the security-seeker's preference to engage in trust-building. When $p_B < \frac{k}{2+k-L}$, 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 p_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 p_i.

Proposition A.5 *If* $1 - L + k > 0$*, and*

$$\frac{k(1-L+k)}{k(2+k-L)-(H-1)(1-L)} > p_i > \frac{k(1-L+k)}{k^2-2kL+3k+L^2-2L+1} \tag{2}$$

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, $\sigma_{2A} = 1|b_1 = c$, $\sigma_{2A} = 0|b_1 = d$. This implies, $\omega_A^* = \frac{k}{(1-L+k)}$. The security-seeker remains on the path if, $p_B(1+\omega_B^*) - k(1-p_B) > p_B L$. This solves for $p_B > \frac{k(1-L+k)}{k^2-2kL+3k+L^2-2L+1}$. The greedy type remains on the path if, $p_B H > p_B(1+\omega_A^* H) - k(1-p_B)$. This solves for $p_B < \frac{k(1-L+k)}{k(2+k-L)-(H-1)(1-L)}$, as stated in the equilibrium.

Contrasting the lower bounds of inequalities 2 and 1,

$$\frac{k(1-L+k)}{k^2-2kL+3k+L^2-2L+1} > \frac{k}{2+k-L}$$

collapses to $L < 1$. This assures we can support 1 at lower levels of p_i, as desired.

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. Third, we solve for all mixed strategy equilibria and rule out those we cannot support. Finally, we apply an iterative efficiency refinement.

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

Proposition A.6 *If*

$$p_i > \frac{k - \beta_j(1+k-L)}{1+(1-\beta_j)(1+k-L)} \tag{3}$$

and

$$\frac{\beta_j(H-1-k)+1}{\beta_j(H-1-k)+1+k} > p_i \tag{4}$$

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, $\sigma_2^A|b_1 = d = 0, \sigma_2^A|b_1 = c = 1$. By Lemma A.1, we can support second-period cooperation. Turning to the first period, A_s prefers to cooperate, rather than defect if:

$$p_B(1+1) + (1-p_B)(\beta_A - (1-\beta_A)k) > p_B L + (1-p_B)\beta_A L$$

This assumes that if A defects, A gets the second-period value 1 from $b_2 = a_2 = d$. This rearranges to equilibrium condition 3, as desired. It will help later to express it as $\beta_A > \frac{k - p_B(2+k-L)}{(1-p_B)(1+k-L)}$.

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:

$$p_B H + (1-p_B)\beta_A H > p_B(1+H) + (1-p_B)(\beta_A - (1-\beta_A)k)$$

This rearranges to inequality 4 as desired. It will help later to express it as $\beta_A > \frac{p_B - k(1-p_B)}{(1-p_B)(H-1+k)}$.

There are no off-path beliefs. The strategies are symmetric. This completes the proof.

Later we will consider efficiency, and so it is useful to characterize total expected utilities.

Remark In the trustbuilding equilibrium, first period total expected utilities are:

$$EU_1^A|A_s = \beta_A - (1-\beta_A)k + p_B(2 - \beta_A + (1-\beta_A)k, EU_1^A|A_g = H(\beta_A + p_B - \beta_A p_B)$$

which are both increasing in β_A.

A.3.2 Establishing Result 1a,1b

Result 1a: When both players' choices are sufficiently independent (i.e., $\beta_A, \beta_B > \frac{k}{1+k-L}$), 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., $p \to 0$).

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, p.

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 p (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 p 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 $p \to 0$) given,

$$\beta_i > \frac{k}{1+k-L} \in (0,1)$$

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

$$\beta_i > \frac{k - p_j(2 + k - L)}{(1 + k - L)(1 - p_j)}$$

wherein the right hand-side is strictly decreasing in p_j.[29]

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. Note the denominator of 3 must be positive for all $\beta_j \in [0, 1]$. and the numerator is decreasing in β_j. The threshold sets the numerator to 0 and re-arranges. The second claim comes from simply re-arranging 3 as a function of β. Taking the derivative of the RHS, $\frac{L-2}{(1+k-L)(1-p_j)^2}$, which must be negative. 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 p. 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 $\beta_j > \frac{p_i - (1-p_i)k}{(H-k-1)(1-p_i)}$.

Putting both type's of incentives together, notice that

Remark For any set of parameters H, k, L, (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 $\frac{H-k}{H} > p_i$.

A.3.3 Other symmetric 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.7 *There is a **tragic equilibrium** if and only if the dependence threshold is not met: $\beta_i < \frac{k}{1-L+k}$. 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 $\beta_A L + EU_2 < \beta_A - (1 - \beta_A)k + EU_2$. 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.

[29] recall, subscript j means not i.

Proposition A.8 *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.9 *There is a **No learning** PBE if $p_i > \frac{k}{1+k-L}$ and $\beta_i < \frac{k}{1-L+k}$. 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: $\beta_A L + p_B - k(1-p_B) > \beta_A - (1-\beta_A)k + p_B - k(1-p_B)$. This solves for $\beta_A < \frac{k}{1-L+k}$, 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.10 *If $p_i > \frac{k}{1+k-L}$ and $\beta_i < \frac{p_j(1+k)}{1+k-L}$. 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: $\beta_A L + p_B - k(1-p_B) > \beta_A - (1-\beta_A)k$.

Proposition A.11 *There is a **Semi-tragic** PBE if and only if $\beta_i > \frac{k(1-p_j)-p_j(1-L)}{(1-p_j)(1+k-L)}$. 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: $p_B + (1-p_B)(\beta_A - (1-\beta_A)k) > p_B L + (1-p_B)(\beta_A L)$. This solves for $\beta_A > \frac{k(1-p_B)-p_B(1-L)}{(1-p_B)(1+k-L)}$ as stated. B's incentives are symmetric. Rearranging gives the minimum boundary on p, $p_j > \frac{k-\beta_i(1+k-L)}{(k+1-L)(1-\beta_i)}$. This completes the proof.

A.3.4 Asymmetric equilibria

When $\beta_i = 0$, we ruled out the possibility of asymmetric equilibria entirely. Once we add in independent first period actions, we can rationalize asymmetric equilibria where players play different strategies in the first period. The intuition behind this result is that the security seeker's minmax changes. When $\beta = 0$, the largest amount both types could guarantee themselves follows from $a_1 = a_2 = d$. We've shown that when the independence threshold is reached $\beta_i > \frac{k}{1+k-L}$ A_s strictly prefers $a_1 = c$ even if $b_1 = d$. This assures that even if A_s knows that B will cheat for certain in the first period, A_s will still cooperate in the first period. Because different types now how different minmax strategies, the opportunities for screening also change.

When $\beta_A, \beta_B > \frac{k}{1+k-L}$ then the incentives for both players are the same. However, when dependencies are lopsided (β_A is high and β_B is low), then security seekers face different minmax strategies. As we will show the dependence threshold is a binding constraint, even when it is not met, lopsided initial trust (p_A high, p_B low) is another critical factor for inducing different asymmetric equilibria.

Asymmetric trust-building equilibria Two asymmetric, pure strategy, trust-building equilibria emerge.

Proposition A.12 *If* $p_B > \frac{1}{1-k+L}$,

$$\beta_A > \frac{p_B H - k}{H - 1 - k} \sim \frac{k - p_B + k(1-p_B)}{1 + k - L} \sim 0$$

$$\beta_B < \frac{k - p_A(1 - L + k)}{(1-p_A)(1 - L + k)} \tag{5}$$

then, there is an asymmetric, pure strategy, trust building equilibrium: $s^A(a_1 = c|A_S, a_1 = d|A_g, b_2 = c|A_s \& a_1 = c \& b_1 = ., a_2 = d$ *otherwise*$)$, $s^B(b_1 = d, b_2 = c|B_s \& a_1 = c \& b_1 = ., b_2 = d$ *otherwise*$)$. *There is an equivalent 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 $p_B > \frac{k}{1+k-L}$.

In the first period, A_s cannot profitably deviate from $a_1 = c \to d$ if: $\beta_A - k(1-\beta_A) + p_B - k(1-p_B) > \beta_A L \equiv \beta_A > \frac{k - p_B + k(1-p_B)}{1+k-L}$, as desired. We can re-write it as $p_B > \frac{2k - \beta_A(1+k-L)}{1+k}$.

A_g cannot profitably deviate from $a_1 = d \to c$ if: $\beta_A H > \beta_A - (1-\beta_A)k + p_B H$. We can write it as $\frac{\beta_A(H-1-k)+k}{H} > p_B$.

B's first period binding constraint is B_s's ICC. B_s cannot profitably deviate from $b_1 = d \to c$ if: $p_A(L+1) + (1-p_A)(\beta_B L) > p_A(1+1) + (1-p_A)(\beta_B - (1-\beta_B)k)$, solves for $p_A < \frac{\beta_B(1-L+k)-k}{\beta_B(1-L+k)-k+L}$, as desired.

These conditions directly imply:

Remark Asymmetric trust building requires:

- One side has a highly independent trust-building action (β_A must be positive and sufficiently large) and moderate-to-high initial trust (p_B is only bound from below if β_A is sufficiently large).
- The other has lowly independent trust-building action (β_B cannot be too high, and must be lower than the independence threshold) and low initial trust (p_A cannot be too high).

Finally, B has one off-path deviation $b_1 = d \to c$. We claimed that this does not effect second period strategies (critically here $a_2 = c|A_s$). This follows instantly given our assumption of feasible off-path beliefs match on-path beliefs in this case.

We now turn to a second equilibria that deviates only in this off-path case. Rather, than assume players are insensitive to B's off-path action, we now assume players revert to the punishment $a_2 = b_2 = d|b_1 = c$.

Proposition A.13 *Replacing condition 5 with*

$$\beta_B < \frac{k - p_A(k - L)}{(1-p_A)(1 - L + k)}$$

then, there is a second 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 equilibrium swapping A and B.*

Trivially, we can support $a_2 = b_2 = d$ for any parameters. Thus, the only thing that changes is B's first period incentive to deviate. B_s imposes the binding constraint. B_s cannot profitably deviate from $b_1 = d \to c$ if: $p_A(L+1) + (1-p_A)(\beta_B L) > p_A + (1-p_A)(\beta_B - (1-\beta_B)k)$.

Later, we will analyze Pareto efficient equilibria. So we emphasize,

Remark This latter equilibria is Pareto dominated by the former across all the parameter ranges we can sustain it.

Other Asymmetric pure strategy equilibria There is an equivalent equilibrium swapping the As and Bs.

Proposition A.14 *If $\beta_A > \frac{k(1-p_B)-p_B(1-L)}{(1-p_B)(1+k-L)}$, and $\beta_B < \frac{k(1-p_A)-p_A(1-L)}{(1-p_A)(1+k-L)}$ 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. B_s ICC for first period cooperation over defect is $p_A + (1-p_A)(\beta_B - (1-\beta_B)k) < p_A L + (1-p_A)(\beta_B L)$. Note $\frac{k(1-p_B)-p_B(1-L)}{(1-p_B)(1+k-L)} < \frac{k}{1+k-L}$.

Remark We cannot sustain this equilibrium if both states have met their independence thresholds.

It is also useful to solve the ICC for $p_A < \frac{k-\beta_B(1+k-L)}{(1-\beta_B)(1+k-L)}$. As this illustrates, sustaining B's incentive requires low initial trust. If p_A was higher, B_s could profitably deviate to cooperation. But it also assures that if p_B is large, we can sustain this equilibrium even if $\beta_A < \frac{k}{1+k-L}$,

Remark We can sustain this equilibrium if neither states has met its independence thresholds.

A.3.5 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: $p_B(\beta_A - (1-\beta_A)k) + (1-p_B) > p_B(\beta_A H + H) + (1-p_B)H$. This re-arranges to, $p_B(-\beta(H-1) - (1-\beta)k - 1) > H - 1$. Note that the LHS must be negative and the RHS must be positive, which assures no p_B exists to satisfy it. It follows that for any contingent second period strategy profile we could support, A_g always has a profitable first period deviation from $a_1 = c \to d$.

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 $i_1 = c \to d$. 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.3.6 Mixed strategy equilibria

We now solve for mixed strategy equilibria. As will become clear, each is a variant of a pure strategy PBE we have already characterized. Because we will later apply a Pareto refinement it is important to understand that each mixed strategy equilibria is Pareto dominated by its respective pure strategy equivalent. They also arise generally in overlapping parameter ranges with their respective pure strategy equivalents. In fact, the mixed strategy equivalents of the symmetric suckers, no learning, and semi-tragic equilibria form a proper subset of their pure strategy equivalents. However, while the mixed strategy trust-building equilibria are dominated by the pure strategy trust building equilibrium, they can be sustained at higher levels of initial trust than the pure strategy equivalent.

Mixing in second period only There are three 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:

$$\omega_i^* = pr(i_2 = c) = \frac{k}{\sigma_2^j(1 - L + k)}$$

Which produced second period expected utilities:

$$U_2^A|A_s = \frac{Lk}{1 - L + k} \qquad U_2^A|A_g = \frac{Hk}{1 - L + k}$$

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

First,

Lemma A.15 *If*

$$p_i > \frac{k}{1 - L + k}$$

$$\beta_A < \frac{k(1 + k)}{(1 + k - L)^2}$$

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

Given first period pooling, $\sigma_2^i = p_j$. 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. From what we've shown, A_g trivially cannot profit from $a_1 = d \rightarrow c$ under any condition. A_s's ICC is: $\beta_A L + \frac{Lk}{1-L+k} > \beta_A - (1 - \beta_A)k$, which solves for $\beta_A < \frac{k(1+k)}{(1+k-L)^2}$, as desired.

Remark This equilibrium is Pareto dominated by the pure strategy no learning equilibrium, and is contained within its parameters.

Second,

Lemma A.16 *If*

$$p_i > \frac{k}{1 - L + k}$$

$$H < \frac{1-L+k}{1-L}$$

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, $\sigma_2^i = p_j$. This gives first condition 1. A_s ICC is $1 + \frac{Lk}{1-L+k} > L$, always satisfied. A_g's ICC is $1 + \frac{Hk}{1-L+k} > H$, which gives us $H < \frac{1-L+k}{1-L}$.

Remark This equilibrium is Pareto dominated by the pure strategy suckers equilibrium, and is contained within its parameters.

Third, there is a trust-building equilibria with perfect separation in the first period, and mixing in the second.

Lemma A.17 *If*

$$\frac{k-\beta(1-L+k)}{(1-\beta)(1-k+L)+L\omega^*} < p < \frac{k+\beta(H-1-k)}{k+1+\beta(H-k-1)-H(1-\omega^*)}$$

then the following mixed strategy trust-building strategy profile is an equilibrium $s^A(a_1 = c|A_s, a_1 = d|A_g, a_2 = \omega_A^*|(a_1 = b_1 = c\&A_s), a_2 = d$ *otherwise*$)$. *This assures on path beliefs* $\sigma_i^* = 1|j_1 = c$ *and mixing probability* $\omega_i^* = \frac{k}{1-L+k}$.

Since the first period is perfectly separating, $\sigma_2^i \in \{0,1\} \implies \omega_i^* = \frac{k}{1-L+k}$. We now turn to the first period. Greedy type cannot profitably deviate from first-period defection if: $pH+(1-p)\beta H > p(1+\omega^* H) + (1-p)(\beta - (1-\beta)k)$.

$$p < \frac{k+\beta(H-1-k)}{k+1+\beta(H-k-1)-H(1-\omega^*)}$$

Security seekers cannot deviate from first period cooperation if

$$p(1+L\omega^*) + (1-p)(\beta-(1-\beta)k) > pL + (1-p)\beta L$$

, which solves for

$$p > \frac{k-\beta(1-L+k)}{(1-\beta)(1-k+L)+L\omega^*} = \frac{k-\beta(1-L+k)}{(1-\beta)(1-k+L)+\frac{kL}{1-L+k}}$$

Remark This equilibrium is Pareto dominated by the pure strategy trust-building equilibrium. The lower bound in inequality 3 strictly binds the mixed strategy equilibrium. But the upper bound does not. That is, we can always find $\frac{\beta_j(H-1-k)+1}{\beta_j(H-1-k)+1+k} < p_i < \frac{k+\beta(H-1-k)}{k+1+\beta(H-k-1)-H(1-\omega^*)}$.

Remark This equilibrium is Pareto dominated by the pure strategy suckers equilibrium, and when $\beta_i =$ is contained within its parameters.

In the trust-building equilibrium, the lower bound arises because the security seeker weighs the concerns about being cheated against the value of second period cooperation. Mixed strategies reduce the security seeker's value of cooperation in the second period. Thus, the security seeker is willing to initially cooperate under fewer conditions. The upper bound arises because the greedy type is not tempted to cheat in the first period. By reducing their value of waiting to cheat in the second period, we increase their incentives to defect in the first (without altering their value from on-path play).

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:

$$\omega_i^x = \frac{k - \beta_j(1 + k - L)}{p_i(1 + k - L)(1 - \beta_j)}$$

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 $p_i > \frac{k-\beta_j(1+k-L)}{(1+k-L)(1-\beta_j)}$, with the special case $p_i > \frac{k}{(1+k-L)}$ given $\beta_i = 0$. It also requires $\frac{k}{1+k-L} > \beta_j$. Note that these conditions are a subset of the pure strategy no learning equilibria.

The first equilibrium is:

Lemma A.18 *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, $\omega_B p_B + (1 - \omega_B p_B)(\beta_A - (1 - \beta_A)k) = \omega_B p_B L + (1 - \omega_B p_B)\beta_A L$, 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.

Remark This equilibrium is Pareto dominated by the symmetric semi-tragic equilibrium, and is contained within its parameter ranges.

The second is,

Lemma A.19 *If $\omega_i^x \in [0, 1]$, and*

$$p_i > \frac{1 + k - (k + \beta_j)(1 + k - L)}{(1 + k - L)(1 - \beta_j)}$$

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. $\sigma_2^A | b_1 = d = \frac{p_B(1-\omega_B^z)}{p_B(1-\omega_B^z)+1-p_B}$. Thus, to sustain second period choices, it must be that $\frac{p_B(1-\omega_B^x)}{1-p_B\omega_B^x} > k/(1 + k - L)$. 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. Note the equilibrium constraint on p_i assures $\omega \in [0,1]$ given $\frac{k-\beta_j(1+k-L)}{(1+k-L)(1-\beta_j)} < \frac{1+k-(k+\beta_j)(1+k-L)}{(1+k-L)(1-\beta_j)}$. There are no off-path actions in the first period, and thus no second period reversion is necessary. Note that with the additional condition, this is a subset of the pure strategy suckers equilibrium.

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:

$$\omega_i^z = \frac{k - \beta_j(1+k-L)}{p_i(1+(1-\beta_j)(1+k-L))}$$

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 $\frac{k}{1+k-L} > \beta_j$. Further, $p_i > \frac{k-\beta_j(1+k-L)}{(2-L+k-\beta_j(1+k-L))}$, and in the baseline case, $p_i > \frac{k}{2-L+k}$. These are the same constraints as in the pure strategy trust-building equilibrium.

Lemma A.20 *If*

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

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, $\omega_B p_B(1+1) + (1-\omega_B p_B)(\beta_A - (1-\beta_A)k) = \omega_B p_B L + (1-\omega_B p_B)\beta_A L$, 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 $\omega_B p_B H + (1-\omega_B p_B)H\beta_A > \omega_B p_B(1+H) + (1-\omega_B p_B)(\beta_A - (1-\beta_A k))$. Rearranging to $\frac{H\beta_A + k - \beta_A(1+k)}{p_B(2+H\beta_A+k-\beta_A(1+k)} > \omega_i$. Plugging in $\omega_B = \omega_A^z$, this is always satisfied if $\omega_A^z \in [0,1]$. There are no off-path beliefs. This completes the proof.

Remark This equilibrium is Pareto dominated by the mixed strategy trust building equilibria characterized in Lemma A.17, and is contained within its parameter ranges.

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: $p_B(1+H) + (1-p_B)(\omega_B^g + (1-\omega_B^g)(\beta_A + (1-\beta_A)k) = p_B H + (1-p_B)(H\omega_B^g + (1-\omega_B^g)\beta_A H$. This simplifies to, $p_B - (1-p_B)(\beta_A(H-1-k) - k) = \omega(1-p_B)(H-1-k)(1-\beta_A)$. Note this cannot hold for $\omega < 1$. 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**. As a reminder, we have solved for the unique, on path second-period mixing strategy:

$$\omega_i^* = pr(i_2 = c) = \frac{k}{\sigma_2^j(1-L+k)}$$

This always imposes the equilibrium condition: $\sigma_2^i > \frac{k}{1-L+k}$.

A critical feature to note is that if i_s mixes in the second period, then no matter posterior beliefs σ_2^i second period expected utilities are:

$$U_2^A|A_s = \frac{Lk}{1-L+k} \qquad U_2^A|A_g = \frac{Hk}{1-L+k}$$

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 p_i.

$$\omega_i^\gamma = \frac{k-\beta_j(1-L+k)}{p_i(\frac{Lk}{1-L+k}+(1-\beta_A)(1-L+k))}$$

The requirement $\omega_i^\gamma \in (0,1)$ imposes a minimum bound on $p_i > \frac{k-\beta_j(1-L+k)}{(\frac{Lk}{1-L+k}+(1-\beta_A)(1-L+k))}$, $\frac{k}{1-L+k} > \beta_j$.

Lemma A.21 *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 $\sigma_2^i > \frac{1}{1-L+k}$. 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, $\omega_B p_B(1+\frac{Lk}{1-L+k}) + (1-\omega_B p_B)(\beta_A(1+k)-k) = \omega_B p_B L + (1-\omega_B p_B)\beta_A L$, which gives us $\omega_B = \omega_B^\gamma$, as desired.

The greedy type cannot profitably deviate from $a_1 = d \to c$ if: $p_B\omega_B H + (1-p_B\omega_B)\beta_A H > p_B\omega_B(1+\frac{Hk}{1-L+k}) + (1-p_B\omega_B)(\beta_A(k+1)-k)$. This solves for $p_B\omega_B((H-k-1)(1-\beta_A) - \frac{Hk}{1-L+k}) > -k-\beta_A(H-1-k)$, always true. This completes the proof.

Finally, recall the mixing probability,

$$\omega_i^x = \frac{k-\beta_j(1+k-L)}{p_i(1+k-L)(1-\beta_j)}$$

Lemma A.22 *If $\omega_i^x \in (0,1)$, and*

$$p_i > \frac{2k-(1+k-L)(k^2+\beta_j)}{(1+k-L)(1-\beta_j)} > \frac{k}{1+k-L}$$

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

The analyses above are enough to show that i_g cannot profitably deviate given that i_s will mix regardless. Thus, we analyze i_s strategy. We've shown in the second period, i_s cannot profitably deviate if $\sigma_2^i > \frac{k}{1+k-L}$. A's posterior belief is $\sigma_2^A|b_1 = d = \frac{p_B(1-\omega_B^x)}{1-\omega_B^x p_B}$. Plugging in ω_B^x, A_s is only willing to cooperate even after observing defection if $p_B > \frac{2k-(1+k-L)(k^2+\beta_A)}{(1+k-L)(1-\beta_A)}$, as written. We

emphasize that $\frac{2k-(1+k-L)(k^2+\beta_j)}{(1+k-L)(1-\beta_j)} > \frac{k}{1+k-L}$ 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^x. i_s is held indifferent given ω_i^x. Focusing on A_s, $\omega_B p_B(1+\frac{Lk}{1-L+k}) + (1-\omega_B p_B)(\beta_A(1+k) - k + \frac{Lk}{1-L+k}) = \omega_B p_B(L + \frac{Lk}{1-L+k}) + (1-\omega_B p_B)(L\beta_A + \frac{Lk}{1-L+k})$, which gives us $\omega_B = \omega_B^x$, as desired.

Remark Both of these equilibria are Pareto dominated by the mixed strategy trust building equilibria characterized in Lemma A.17, and is contained within its parameter ranges.

A.3.7 Pareto efficient equilibria

To begin, we summarize the total expected utilities of candidate equilibria. That is, all equilibria we have not yet shown are Pareto dominated in the parameter ranges where we can sustain them. When strategies are symmetric we denote all total expected utilities from A's perspective. When they are asymmetric, we denote utilities for both players. Finally, we add descriptors when there are different variants of each equilibria. When there is only a pure strategy symmetric equilibrium to consider, we do not add a descriptor.

Remark The first period total expected utilities from on path play in all candidate PBE are:

1. Symmetric, trust-building (prop A.6): $EU_1^A|A_s = 2p_B + (1-p_B)(\beta_A - (1-\beta_A)k)$, $EU_1^A|A_g = H(p_B + (1-p_B)\beta_A)$.
2. Asymmetric, trust-building (prop A.12) $EU_1^A|A_s = (\beta_A + p_B)(1+k) - 2k$, $EU_1^A|A_g = \beta_A H$, $EU_1^B|B_s = p_A[1 + L(1-\beta_B)] + \beta_B L$, $EU_1^B|B_g = H[p_A(2-\beta_B) + \beta_B]$.
3. Symmetric mixed strategy trust building (prop A.17) $EU_1^A|A_s = p(1 + L\frac{Lk}{1+k-L}) + (1-p)(\beta - (1-\beta)k)$, $EU_1^A|A_g = pH + (1-p)\beta H$,
4. Suckers (prop A.8): $EU^A|A_s = 1 - k + p(1+k)$, $EU^A|A_g = 1 + p_B H$.
5. No learning (prop A.9): $EU^A|A_s = \beta_A L - k + p(1+k)$, $EU^A|A_g = \beta_A H + p_B H$
6. Tragic (prop A.7): $EU_1^A|A_s = \beta_A L$, $EU_1^A|A_g = \beta_A H$
7. Symmetric, semi-tragic (prop A.11)) $EU_1^A|A_s = p_B + (1-p_B)(\beta_A - (1-\beta_A)k)$,$EU_1^A|A_g = p_B H + (1-p_B)\beta_A H$
8. Asymmetric, semi-tragic (prop A.14) $EU_1^A|A_s = p_B - (1-p_B)(k - \beta_A(1+k))$, $EU_1^A|A_g = \beta_A H$, $EU_1^B|B_s = \beta_B L + p_A L(1-\beta_B)$, $EU_1^B|B_g = \beta_B H + p_A H(1-\beta_B)$.

Proposition A.23 *Pure strategy equilibria are Pareto efficient with the following additional parameter constraints.*

1. *The symmetric trust-building equilibrium (prop A.6) is dominated by the suckers equilibrium if $\beta_i < \frac{1}{(1+k)}$ holds. Thus, when this condition holds, the symmetric trust building survives refinement with the additional restriction, $p_i < 1 - \frac{1}{H} \sim \frac{k}{1+k-L}$.*

2. *The asymmetric trust-building equilibrium (prop A.12) is not dominated by any other equilibrium and survives refinement under stated conditions.*

3. *The symmetric, mixed strategy trust-building equilibrium (prop A.12) is Pareto dominated by the pure strategy trust-building equilibrium and the suckers equilibrium. Thus, it survives refinement with the additional restriction,* $\frac{\beta_j(H-1-k)+1}{\beta_j(H-1-k)+1+k} < p_i < 1 - \frac{1}{H} \sim \frac{k}{1+k-L}$. *Note it cannot survive refinement when* $\beta_i = 0$.

4. *The suckers equilibria (prop A.8) is not dominated by any other equilibrium and survives refinement under stated conditions.*

5. *The no learning equilibrium (prop A.9) is dominated by the suckers. Thus, it survives refinement with the additional parameter restriction* $p_i < 1 - \frac{1}{H}$.

6. *The tragic equilibrium is Pareto dominated by all other candidate equilibria in the conditions we can sustain it. Thus, it only survives refinement when no other candidate equilibria survive.*

7. *The symmetric semi-tragic equilibria is Pareto dominated by all symmetric equilibria except the tragic equilibrium in the conditions we can sustain it. Thus, it only survives refinement when no other candidate equilibria survive.*

8. *The asymmetric semi-tragic equilibria, is Pareto dominated by the asymmetric trust-building equilibria. Thus, it only survives refinement if either of the additional parameter restrictions are met,* $\frac{p_B H-k}{H-1-k} > \beta_A$ *or* $\beta_B > \frac{k-p_A(1-L+k)}{(1-p_A)(1-L+k)}$.

The results follow from a simple comparison of the total expected utilities of security seekers and greedy types. Some notable points. First, the suckers equilibrium dominates even trust building when β_i is low. The reason is that both equilibria run the risk of being exploited, but in the trust building equilibria mutual cooperation only follows in the condition that players don't cheat each other. By contrast, mutual cooperation is assured in the suckers equilibrium. As β_i increases, it is increasingly attractive to run the risk of exploitation in the first period, and this can offset the loss of conditional cooperation relative to assured cooperation. Second, the asymmetric and symmetric equilibria do not dominate each other in their respective parameter ranges. The reason is that asymmetric equilibria always have a quality where one state knows they will be cheated by their rival in the first period, and the other is certain they will get to cheat their rival. Thus, we can always find at least one type that prefers the symmetric over asymmetric and vice versa.

Third, when $\beta_A, \beta_B > \frac{k}{1+k-L}$, then we can sustain the minmax in the semi-tragic equilibrium given any levels of trust. When $\beta_A, \beta_B < \frac{k}{1+k-L}$, then we can sustain the minmax in the tragic equilibrium given any levels of trust. Since all other equilibria leave at least one player with more than their minmax, (and trivially, no player can do worse), these are always dominated.

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.

Proposition A.24 *If*

$$\theta > \frac{k - [\beta_j + p_i(1-\beta_j)](1+k-L)}{p_i} \tag{6}$$

and

$$\theta < \frac{k + (H-k-1)(\beta_j + p_i(1-\beta_j)}{p_i H} \tag{7}$$

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:

$$p_B(1+\theta) + (1-p_B)(\beta_A - (1-\beta_A)k) > p_B L + (1-p_B)\beta_A L$$

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:

$$p_B H + (1-p_B)\beta_A H > p_B(1+\theta H) + (1-p_B)(\beta_A(1+k) - k)$$

This re-arranges to condition 7, as desired.

There are no off-path beliefs. This completes the proof.

A.4.1 Result 2: Implications of salience

In the manuscript, we make the claim that if independence is sufficiently large, then there is no Goldilocks problem, and grand gestures are useful tools for trust-building. We support this with two remarks.

Remark Condition 6 is certainly satisfied if the dependence threshold ($\beta_j > \frac{k}{1+k-L}$) is met.

This assures the LHS is negative. As a result, there is no upper bound on the relative importance of a domestic choice.

Remark For any β_A, there always exists a θ sufficiently large to violate Condition 7.

This suggests that domestic choices must be at least non-trivial. If they are very unimportant relative to foreign rivalries, then the equilibrium degenerates.

Remark At full dependence, we can satisfy condition 7 if $\theta < \frac{H-1}{p_B H}$, which must be satisfied if $\theta \to 0$

This in combination with the other remarks assures that if we reach a certain level of independence, then we can always find a domestic action that is sufficiently important to sustain the trust building equilibrium.

To be clear, this does not always mean that increasing independence assures trust building arises under greater conditions:

Remark If $H > k+1$ Condition 7 is increasingly easier to satisfy as β_A increases. If $H < k+1$ Condition 7 is increasingly harder to satisfy as β_A increases.

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 4 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 p_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.24. 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 former

Proposition A.25 *When*

$$p > \frac{k(1-\alpha)}{\alpha(1-L)} \tag{8}$$

$$p > \frac{\alpha\left[k-\beta(1+k-L)\right]+(1-\alpha)\left[\theta k-1+L\right]}{\alpha\left[\theta+(1+k-L)(1-\beta)\right]+(1-\alpha)\left[\theta k-(1+k-L)(1-\beta)\right]} \tag{9}$$

and either

$$p < \frac{\alpha\left[k+\beta(H-k-1)\right]+(1-\alpha)\left[\theta k+H-1\right]}{\alpha\left[\theta-(H-k-1)(1-\beta)\right]+(1-\alpha)\left[\theta k-(H-k-1)(1-\beta)\right]} \tag{10}$$

or the denominator of 10 is negative, and

$$p < \frac{\alpha\left[k-\beta(H-1-k)\right]+(1-\alpha)\left[H-1\right]}{\alpha\left[\theta-(H-k-1)(1-\beta)\right]+(1-\alpha)(H-k-1)(1-\beta)} \tag{11}$$

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:

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

Setting $\sigma_2^A > \frac{k}{1+k-L}$ this solves for equilibrium condition 12.

We now turn to first-period strategies. On path, A_s cooperates in the $\alpha = 1$ condition (i.e, the good type cooperates if she has good preferences). A^s cannot profitably deviate to defect in this condition iff:

$$p\alpha(1+\theta)+p(1-\alpha)(\beta(1+k)-k)+(1-p)\alpha(\beta(1+k)-k)+(1-p)(1-\alpha)(1-\theta k) > p\alpha L+p(1-\alpha)\beta L+(1-p)\alpha\beta L+(1-p)(1-\alpha)L$$

This solves for equilibrium condition 9. We deliberately disaggregated the denominator and numerator on $\alpha, 1-\alpha$. Note that when $\alpha = 1, \theta = 1$, the inequality converges to our main trust-building result.

On path, A_s defects in the $\alpha = 0$ condition (i.e, the good type defects if she has bad preferences). A^s cannot profitably deviate to cooperate in this condition iff:

$$p\alpha(1+\theta)+p(1-\alpha)(\beta(1+k)-k)+(1-p)\alpha(\beta(1+k)-k)+(1-p)(1-\alpha)(1-\theta k) < p\alpha H+p(1-\alpha)\beta H+(1-p)\alpha\beta H+(1-p)(1-\alpha)H$$

This solves for equilibrium 10. Note this inequality has no analog in the baseline model because it assumes good types defect.

Finally, we turn to A_g first period incentive. The binding constraint is the $\alpha = 1$ case.[30] We conjecture that A_g defects (the bad type defects if she has bad preferences). A_g cannot profitably deviate if:

$$p\alpha(1+\theta H)+p(1-\alpha)(\beta(1+k)-k)+(1-p)\alpha(\beta(1+k)-k)+(1-p)(1-\alpha) < p\alpha H+p(1-\alpha)\beta H+(1-p)\alpha\beta H+(1-p)(1-\alpha)H$$

This solves for condition 11. Note when $\alpha = 1, \theta = 1$, this condition converges to the baseline.

There are no off-path beliefs. This completes the proof.

We now solve the latter.

Proposition A.26 *When*

[30] In the other case, the bad type achieves her maximum possible payoff because she strictly prefers first period cooperation to exploiting the other side, and doing so gives the opportunity to exploit in the second period.

$$p_B > \frac{k(1-\alpha_B)}{1-L+k(1-\alpha_B)} \tag{12}$$

and

$$\frac{\alpha(k+\beta(H-k-1)}{H\theta-H+1+\alpha(k+\beta(H-k-1))} > p > \frac{\theta k+H-1-\alpha[(H-k-1)(1-\beta)+\theta k]}{\theta+k-\alpha[(H-k-1)(1-\beta)+\theta k]} \tag{13}$$

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:

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

Setting $\sigma_2^A > \frac{k}{1+k-L}$ this solves for equilibrium condition 12, 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:

$$p_B(1+\theta)+(1-p_B)\alpha_B(\beta_A(1+k)-k)+(1-\alpha_B)(1-p_B)(1-\theta k) > p_BH+(1-p_B)\alpha_B\beta_AH+(1-\alpha_B)(1-p_B)H$$

This solves for the RHS of the equilibrium condition 13.

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 a strictly higher payoff from cooperation in the first. Focusing on the α case, A_g prefers first-period defection iff:

$$(pH+(1-p)\alpha\beta H+(1-p_B)(1-\alpha_B)H > p_B(1+H\theta)+(1-p_B)\alpha_B(\beta_B(1+k)-k))+(1-p_B)(1-\alpha_B)H$$

This solves for the LHS of the equilibrium condition 13. There are no off-path beliefs.

B Supporting Empirical Tables

U.S.-Soviet Mapping

Table B.1 presents our mapping of the model to symmetric trust-building at the end of the Cold War.

Table B.1: Parameter Values in US-Soviet Case

Item	Description
Equilibrium	Symmetric trust-building
p_{USSR}	Low: US believed Soviets were immoral, bent on global domination, cheated on agreements.
p_{US}	Low: Soviets feared U.S. nuclear first strike and broader "anti-Soviet crusade"
β_{USSR}	Moderate-to-High: Benefits of political, economic, and emigration reforms (vs. status quo) mostly did not depend on what the US did.
β_{US}	Moderate: Encouraging Gorbachev's initiatives (vs. dismissing them) would still bring important benefits even if Soviet reform proved minimal.
θ_{USSR}	High: Glasnost, perestroika, and emigration were highly salient, transformative policies.
θ_{US}	Moderate: Stance of encouragement and support to Soviet Union was a key foreign policy choice with domestic political implications for Reagan.

Table B.2 presents the game matrix for a stylized "period 1" in the Cold War case, with the cells showing the choices corresponding to cooperation and defection for either side. Table B.3 presents the game matrix for a stylized "period 2" in the Cold War case. Bold cells indicate the choices taken.

Table B.2: Period 1: US-Soviet Symmetric Trust-Building

		USA: Cooperate	USA: Defect
USSR	Cooperate	**USSR: Liberalizing reform; US: Encourage Gorbachev**	USSR: Liberalizing reform; US: Dismiss Gorbachev
	Defect	USSR: ¬ Liberalizing reform; US: Encourage Gorbachev	USSR: ¬Liberalizing reform; US: Dismiss Gorbachev

Table B.3: Period 2: US-Soviet International Cooperation

		USA: Cooperate	USA: Defect
USSR	Cooperate	**USSR: Implement INF Treaty, troop reductions in Europe, etc. ; US: Implement INF Treaty, economic aid to USSR, etc.**	USSR: Implement INF Treaty, troop reductions in Europe, etc.; US: Cheat on INF Treaty, no aid to Soviets, etc.
	Defect	USSR: Cheat on INF Treaty, keep troops in Europe, etc.; US: Implement INF Treaty, economic aid to USSR, etc.	USSR: Cheat on INF Treaty, keep troops in Europe, etc.; US: Cheat on INF Treaty, no aid to Soviets, etc.

U.S.-South Korea Mapping

Table B.4 presents our mapping of the model to asymmetric trust-building in the U.S.-South Korea case.

Item	Description
Equilibrium	Asymmetric trust-building
p_{US}	High: South Korea confident in U.S. foreign policy goal of containing Communism internationally
p_{ROK}	Low: U.S. feared Park was a Communist.
β_{ROK}	High: Benefits of economic modernization, anti-Communism (vs. not) mostly depended on Park regime's intrinsic values, vision for ROK.
β_{US}	Low: Benefits of recognizing and legitimizing Park (vs. not) depended greatly on whether Park was anti- or pro-Communist.
θ_{ROK}	High: Park's reforms and anti-Communism represented important choices for future of South Korea.
θ_{US}	Moderate: As a key ally and client, U.S. ties to South Korea's government were important for the U.S.

Table B.4: Parameter Values in US-ROK Case

Table B.5 presents the game matrix for a stylized "period 1" in the U.S.-ROK case, with the cells showing the choices corresponding to cooperation and defection for either side. Table B.6 presents the game matrix for a stylized "period 2" in the U.S.-ROK case case. Bold cells indicate the choices taken.

Table B.5: Period 1: US-ROK Asymmetric Trust-Building

		USA: Cooperate	USA: Defect
ROK	Cooperate	ROK: Modernizing reforms, domestic anti-Communism ; US: Recognize Park regime	**ROK: Modernizing reforms, domestic anti-Communism; US: Withhold recognition of Park regime**
	Defect	ROK: ¬Modernizing reforms, lax on domestic Communism; US: Recognize Park regime	ROK: ¬Modernizing reforms, lax on domestic Communism; US: Withhold recognition of Park regime

Table B.6: Period 2: US-ROK International Cooperation

		USA: Cooperate	USA: Defect
ROK	Cooperate	**ROK: Support U.S. foreign policy (e.g., send forces to Vietnam); US: Extended deterrence, provide aid**	ROK: Support U.S. foreign policy (e.g., send forces to Vietnam); US: Abandon ROK as ally, end aid
	Defect	ROK: ¬Support U.S. foreign policy goals; US: Extended deterrence, provide aid	ROK: ¬Support U.S. foreign policy goals; US: Abandon ROK as ally, end aid