Trust, Reassurance, and Cooperation

Andrew Kydd

Mistrust and fear play a crucial role in many explanations of international conflict. Some structural realists argue that there is "little room for trust among states" because intentions are difficult to discern and hence fear "can never be reduced to a trivial level."[1] Defensive realists see recurrent security dilemmas in international affairs, which are said to force states to arm against each other even though they would both prefer mutual cooperation.[2] At the heart of the security dilemma is mistrust, a fear that the other side is malevolently inclined and bound to exploit one's cooperation rather than reciprocate it. The Cold War, in particular, is often blamed on mistrust between the United States and the Soviet Union.[3] Such explanations of conflict have a tragic character to them. States are held to be fundamentally willing to live in peace with each other if the other side is also willing, but, out of a false conviction that the other side is not, they take offensive measures and end up in conflict.[4]

The possibility that conflict may result from exaggerated perceptions of hostility makes the issue of reassurance important. The fact that war is extremely costly generates strong incentives to avoid it if possible.[5] If the force driving two states into conflict is a set of false beliefs, it would be very beneficial for both sides to dispel these beliefs through strategies of reassurance. Indeed, any understanding of the security dilemma is incomplete without an understanding of the potentialities and limitations of reassurance, and vice versa; reassurance is the flip side of the security dilemma coin. Given that security is an inherently defensive goal that all can enjoy

Earlier versions of this article were presented at workshops at the Hoover Institution and the University of California at Riverside. I would like to thank Juliann Emmons Allison, Shaun Bowler, Kevin Esterling, James Fearon, Charles Glaser, Henk Goemans, Joanne Gowa, Ashley Leeds, Will Moore, James Morrow, Jonathan Nagler, Dan Reiter, Monica Toft, and Barbara Walter for helpful comments on earlier drafts.

1. Mearsheimer 1994, 11.
2. See Jervis 1978; Van Evera 1999; and Glaser 1994 and 1997.
3. See Larson 1997, 5, 235–39; and Lebow and Stein 1994, 4, 366.
4. Spirtas 1996, 387–88.
5. Fearon 1995.

International Organization 54, 2, Spring 2000, pp. 325–357
© 2000 by The IO Foundation and the Massachusetts Institute of Technology

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

simultaneously, for security seekers to end up in conflict with each other there must be some form of misperception about motivations that should be susceptible, at least in principle, to strategies of reassurance.[6] The importance of developing a good theoretical and empirical understanding of reassurance in international affairs is therefore manifest.

In this article I develop what can be called the "costly signaling" theory of reassurance, because it focuses on the sending and interpretation of costly signals.[7] Costly signals, in this context, are signals designed to persuade the other side that one is trustworthy by virtue of the fact that they are so costly that one would hesitate to send them if one were untrustworthy.[8] I examine a model in which players are either trustworthy and prefer to reciprocate cooperation or untrustworthy and would exploit cooperation. Trust is conceived of as a belief that the other side is likely to be trustworthy and will therefore want to reciprocate cooperation rather than exploit it. Costly signals serve to separate the trustworthy types from the untrustworthy types; trustworthy types will send them, untrustworthy types will find them too risky to send.

Two central implications of the costly signaling theory of reassurance are as follows. First, to reassure, a signal must be adequately costly; that is, gestures with little risk attached will be dismissed as feints by the other side and will not change beliefs. For instance, withdrawing a handful of troops from a heavily fortified border will be unpersuasive to the other side if the remaining forces retain the same basic offensive capabilities as before. Such small gestures are "cheap talk" and fail to reassure. Second, in order to be willing to send a signal that is so costly that the untrustworthy type finds it too risky to mimic, the trustworthy type must be willing to take greater risks for peace than the untrustworthy type. Only if the trustworthy type will take greater risks for peace than the untrustworthy type can a signal be found that the former will send but the latter will not. This will be the case if the trustworthy type places a higher value on mutual cooperation or the sucker's payoff for unilateral cooperation than the untrustworthy type. For instance, in the case explored here, Mikhail Gorbachev and the Soviet "new thinkers" had a greater valuation for mutual cooperation than previous Brezhnevite leaders because they wanted a tranquil international environment in which to conduct domestic reforms. This difference in preferences between them and previous leaders enabled them to send signals that previous Soviet leaders would have been unwilling to send, and hence these signals were persuasive to Western publics and leaders.

The broader implications of the theory are encouraging for those concerned with the problem of mistrust-induced conflict but challenge conventional accounts of such conflict. Although mistrust can indeed cause conflict, reassurance through costly signals can also reduce mistrust, leading to full cooperation. Put another way, cooperation is ultimately possible between actors who have significant mistrust of each

6. See Schweller 1996; and Kydd 1997. The only exception is preventive situations, discussed in Kydd 1997, 147–52.

7. For the origin of this literature, see Spence 1973.

8. The normal context of costly signals in international relations is the crisis bargaining literature; see Fearon 1993, chap. 3.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

other, such that they are initially unwilling to cooperate over important issues. Reassurance, therefore, is a rational response to problems of mistrust in potential conflict situations. This fact challenges adherents of structural realist and security dilemma–based explanations of international conflict to refine their accounts. If reassurance is rational in a broad array of situations, it becomes necessary to specify clearly why reassurance is blocked if a mistrust-based explanation of conflict is to be maintained in a particular case. If mistrust between the United States and the Soviet Union is to blame for the Cold War, why were the two sides unable to reassure each other, despite several attempts, until the late 1980s? Of course, many answers to this question are possible. I do not claim that mistrust-based explanations of conflict are ruled out by the model presented here. I would only advocate that a closer degree of attention be paid to the issue of reassurance by those who advance such explanations, especially those who proceed from a rational choice perspective.

The article has five parts. First, I briefly discuss previous work on the issue of trust and reassurance in international relations. Second, I present a basic game representing the mistrust problem in which there is uncertainty over the other player's payoffs and examine how the beliefs and payoffs interact to make cooperation more or less difficult to achieve. Mistrust is shown to hinder cooperation, whereas trust facilitates it. Third, I modify the game to allow actors to move first and cooperate "preemptively" with costly signals. Changing the game in this way enables it to accommodate strategies of reassurance. Analysis of this game shows when reassurance is possible and when it is not. Fourth, I apply the model to the end of the Cold War. I argue that the central events that brought the Cold War to an end can be interpreted as costly signals designed to achieve reassurance, as conceptualized in the model.

Soviet "new thinkers," most importantly Mikhail Gorbachev himself, used a strategy of reassuring concessions such as the 1987 Intermediate-range Nuclear Forces (INF) treaty, the 1988 withdrawal from Afghanistan, and the 1989 noninterference in the Eastern European revolutions to radically change Western perceptions of Soviet motivations and build trust. This understanding of the process of ending the Cold War complements explanations presented by scholars focusing on the ideas underlying the new thinking.[9] Although it is important to trace, as these scholars have, the origin and institutionalization of the new, more benign security worldview adopted by Soviet leaders in the second half of the 1980s, it is equally important to understand how the rest of the world was persuaded that this new thinking was serious and not a ploy designed to split the alliance or weaken NATO resolve. Soviet leaders had to demonstrate a *credible commitment* to the new worldview in order to build trust with the West, and this is what the strategy of costly signaling allowed them to do.[10] Finally, I extend the argument to the related issue of ethnic conflict and discuss how

9. See Risse-Kappen 1994; Mendelson 1993; and Checkel 1993.

10. This use of the phrase *credible commitment* differs slightly from the usage in the crisis bargaining literature where it refers to a commitment to take or refrain from taking certain actions. However, the notion is similar, in that Soviet leaders, by demonstrating the seriousness of their commitment to the new thinking, also demonstrated their commitment not to act in an aggressive fashion in the manner of previous Soviet leaders.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

the model challenges security dilemma–based explanations of conflict offered here as well as in the international arena.

Mistrust and Reassurance

The concept of reassurance has long been salient in the literature on international conflict and cooperation.[11] Two basic strands of thought can be discerned. The first roots itself in psychological theory and critiques what are perceived to be excessively belligerent policies associated with deterrence theory. The second, more rationalist though nonformal, is associated with the security dilemma literature and explores the possibility of using the distinction between offense and defense for purposes of reassurance. A small formal literature on trust and reassurance can also be identified.

The first, more psychological, strand originates in two pioneering studies by Charles Osgood and Amitai Etzioni.[12] Both authors identified the U.S.–Soviet arms race as a product of mistrust and suggested strategies for overcoming this fear through gradual, unilateral cooperative gestures. Osgood was especially critical of the conventional deterrence mindset, calling it the "Neanderthal mentality" and characterizing it as a product of fear and bias.[13] Some scholars found support for the efficacy of Osgood's GRIT (Graduated Reciprocation In Tension-reduction) strategy in the experimental psychological literature on cooperation, as well as in the Arab-Israeli peace process and statistical studies of U.S.–Soviet–Chinese relations, whereas others have expressed skepticism.[14] Deborah Larson has found GRIT in action in U.S.–Soviet relations in a case study of the 1955 Austrian State Treaty and more recently has analyzed superpower relations with a framework influenced by rational choice that stresses mistrust as a central impediment to superpower cooperation.[15] Robert Jervis's examination of perception and misperception in international relations can be said to fall in this group, especially his spiral model of conflict, which focuses on how perceptions of hostility can become exaggerated and lead to war.[16] More recent advocates of reassurance, such as Janice Gross Stein, continue the tradition of criticizing deterrence theory as excessively simplistic, belligerent, and blind to the conflicting pressures affecting decision makers, while arguing for the usefulness of cooperative gestures in fostering less conflictual relationships.[17]

The second, more rationalist, strand originates in Jervis's examination of the security dilemma and the offense/defense distinction. Jervis argued that if offensive and

11. The concept of trust has enjoyed a resurgence of interest in other fields as well, including political philosophy and political economy. See Hollis 1998; Fukuyama 1995; Lane and Bachmann 1998; Landa 1994; Braithewaite and Levi 1998; and Gambetta 1988.

12. See Osgood 1962; and Etzioni 1962.

13. Osgood 1962, 18–36.

14. See Lindskold 1978; Collins 1998; Kelman 1985; Goldstein and Freeman 1990, 130–44; and for the skeptic, Bitzinger 1994.

15. Larson 1987 and 1997, 1–34.

16. Jervis 1976, 62–76.

17. Stein 1991.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

defensive weapons or postures were distinguishable from each other, then this distinction could be used for signaling purposes. Security-seeking states could invest more heavily in defensive weapons or configure their forces in more defensive ways, thereby distinguishing themselves from states with more aggressive motivations, who would presumably have to invest in offense in order to have any hope of achieving their aggressive goals. This distinction would signal their motivation to the other side, reassuring them that the state is indeed defensively motivated and not interested in nonsecurity-related expansion.[18] Subsequent offense/defense theorists, such as Charles Glaser and Stephen Van Evera, argue that when the defense is relatively strong and identifiable, the security dilemma will be much lessened and security-seeking states will be able to reassure each other and avoid conflict.[19] When the offense is dominant, however, this avenue for signaling may be foreclosed because status quo states will be forced to buy offensive weapons and hence will look just as threatening as aggressive states. Reassurance through investment in defense will then be viewed as too risky a strategy to pursue. Glaser introduces the costly signaling concept in the reassurance context, although without developing a game-theoretic model or going deeply into the formal logic.[20]

Some scholars have analyzed game-theoretic models of international mistrust and reassurance. In a series of papers, S. Plous argues that the U.S.–Soviet arms race is better conceived of as a "perceptual dilemma" with three characteristics: each side most prefers mutual disarmament, each side least prefers unilateral disarmament, and each side thinks the other actually prefers superiority to mutual disarmament (incorrectly, as the first assumption makes clear).[21] Plous adduces polling data from the U.S. Senate, Israeli Knesset, and Australian Parliament in support of this thesis and presents an experiment in which unilateral cooperative gestures initiate cooperation in these perceptual dilemmas. Hugh Ward presents a quasi-game-theoretic analysis of a situation in which actors are uncertain about each other's preferences as well.[22] Each actor has assurance (Stag Hunt) preferences, while believing it possible that the other actor has less cooperative preferences. Unfortunately, Ward departs from standard game-theoretical assumptions of full strategic rationality and common prior beliefs, so it is unclear if his results are consistent with the standard rational choice approach or artifacts of the specific assumptions he makes about behavior.[23] George Downs and David Rocke present the only fully formal model and simulation of reassurance in international relations that I am aware of in the context of their work on arms races. They argue that reassurance is by and large a risky strategy that will seldom succeed because of the danger that concessions will be taken advantage of by the other side if they are uninterested in cooperation.[24]

18. Jervis 1978, 199–201.
19. See Glaser 1994; and Van Evera 1999.
20. Glaser 1994, 68.
21. Plous 1985, 1987, 1988, and 1993.
22. Ward 1989.
23. Ibid., 282–83.
24. Downs and Rocke 1990, 107–45. For a very sophisticated formal treatment of reassurance in economics, see Watson 1999.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

Although these accounts of reassurance are helpful, I would argue that our understanding of reassurance is handicapped by a lack of a general rational choice model of the phenomenon. The psychological literature presents a welter of hypotheses and conjectures, some of which are rationalist whereas others are not, and has often focused on prescription at the expense of analysis. Osgood's GRIT consists of fifteen rules with no clear prioritization or underlying theory. Rule *c,* for instance, states that "unilateral initiatives must be graduated in risk according to the degree of reciprocation obtained from opponents," which seems compatible with, though not identical to, the costly signaling theory that will be developed here. Rule *e,* however, states that "prior to announcement, unilateral initiatives must be unpredictable by an opponent as to their sphere, locus, and time of execution."[25] This rule is totally unnecessary from a costly signaling perspective and might even be counterproductive.[26] Overall, the reader is left with the impression that leaders usually fail to pursue reassurance strategies because they are irrational or cognitively limited, but this judgment is impossible to sustain unless we know when it is rational to pursue reassurance strategies in the first place. Conversely, the more rationalist security dilemma strand of the literature highlights one possible avenue for reassurance, the offense/defense distinction, and introduces the costly signaling concept as appropriate in this context but fails to pursue the latter insight in depth. We are left with questions such as, when is it rational for a state to attempt to reassure another with costly signals, when will such signals be believed, and how do the prior beliefs and payoffs of the actors affect the signaling process? To answer this kind of question, a full fledged formal analysis is required.[27] In what follows I present a general model of reassurance that will begin to address such questions. First, however, I start with a simple model of trust and conflict.

The Trust Game

One of the best and most widely influential rational choice analyses of trust is provided by the sociologist James Coleman.[28] Coleman posits a game with two actors, a trustor and a trustee. Coleman uses an example of one farmer deciding whether to help another farmer bale hay. The first farmer has not received a favor from the second and has no positive assurance that the second will return the favor currently being contemplated. Later on, the first farmer will ask the second farmer for a favor and the second farmer will then decide whether to reciprocate the first farmer's earlier cooperation or to exploit it by not cooperating. The crucial dilemma for the first

25. Osgood 1962, 94–96.

26. The word *unpredictable* may be somewhat ambiguous here, but completely unexpected actions take one off the equilibrium path where it is difficult to determine what rational behavior entails.

27. Although Downs and Rocke provide such a model, their results are derived from a numerical simulation of their model and thus are less general than might be desirable.

28. Coleman 1990, 91–116.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

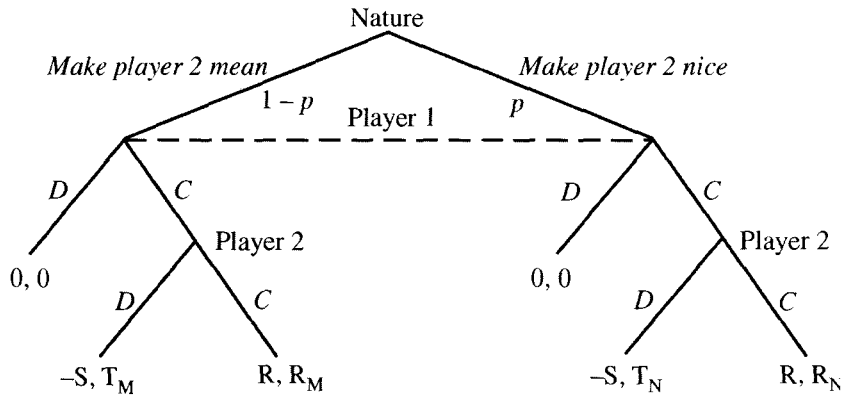


FIGURE 1. *The trust game*

farmer is that he does not know if the second farmer is trustworthy or not, that is, whether the second farmer is inclined to return favors when needed.

This scenario can be formalized in a simple incomplete information game, as shown in Figure 1. Nature first decides whether to make player 2 trustworthy or untrustworthy. I will call the trustworthy player who returns favors "nice" and the untrustworthy player who does not "mean." The crucial difference between the types is that the nice type prefers to reciprocate cooperation, whereas the mean type does not; that is, the nice type's payoff for cooperation, R_N, (for "reward" for cooperation) is higher than his payoff for exploiting the other side's cooperation, T_N (the temptation to defect).[29] For the mean type, the temptation to defect, T_M, beats the reward for mutual cooperation, R_M.[30]

Nature makes player 2 nice with probability p, choosing the right branch of the game tree, and mean with probability $1 - p$, on the left branch of the tree. As indicated by the information set linking player 1's decision nodes, player 1 is not informed of what type of player is being faced. However, the probability that player 2 is nice, p, is known to player 1 and can be thought of as player 1's level of trust. The greater p is, the more certain player 1 is that player 2 is trustworthy, and so the greater is player 1's level of trust for player 2. This prior level of trust can be a product of several things. If one state has had previous interactions with another state in which the state has acted in a hostile fashion, this could result in a lower level of p. For instance, France's distrust of Germany in the immediate post–World War II era, which led it to oppose German rearmament, was a natural result of the French experience of being invaded by Germany three times in the past seventy years. The prior level of trust could also be a result of generalized experience with many other states. If a state has found in the course of its interactions that most states in its neighbor-

29. This payoff notation was made familiar in Robert Axelrod's analysis of the Prisoners' Dilemma, but its origins may lie further back. See Axelrod 1984, 8.

30. In terms of the usual 2 × 2 normal form games, nice types have Stag Hunt preferences, whereas mean types have Prisoners' Dilemma payoffs.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

hood honor their trade commitments, then it may have a generally trusting prior belief, or high p, when it commences a trade relationship with a new, more distant trading partner. Finally, the prior could be influenced by theories or hypotheses about state behavior as they apply to specific cases. For instance, a democracy that believes that democracies are generally more peaceful, at least toward other democracies, may approach a relationship with a new democracy with a greater level of trust than if the other state were authoritarian. For instance, although most new revolutionary regimes are viewed with fear and suspicion, which can in some cases lead to war, the new regimes in Eastern Europe in 1989 were welcomed by the Western democracies because their revolutionary transformations moved them in the direction of Western norms.[31]

Following Nature's choice, player 1 then has the first move and may cooperate or defect. If player 1 decides to cooperate, player 2 has the option to reciprocate that cooperation or exploit it by defecting. If player 1 does not decide to cooperate, the game ends and the payoffs are zero for both players.[32] If player 1 cooperates but player 2 does not, then player 1 gets the sucker's payoff, $-S$, and player 2 gets the temptation to defect, T_N if nice or T_M if mean. The game then ends.[33]

What are the equilibria of the trust game? Backward induction indicates that on the right branch where player 2 is nice, player 2 will cooperate, and on the left branch, where player 2 is mean, player 2 will defect. Given that player 1 values the reward for mutual cooperation over mutual noncooperation, $R > 0$, player 1 has an incentive to cooperate if player 1 thinks player 2 is likely enough to be nice. The payoff for cooperation for player 1 is $pR + (1 - p)(-S)$, and this will be greater than zero, the payoff for defecting, if p (the level of trust player 1 has for player 2) exceeds a critical threshold p^* defined in the following equation:

$$p^* = \frac{S}{R + S}$$

If $p > p^*$, player 1 is trusting enough of player 2 to cooperate. If $p < p^*$, then player 1 is too mistrustful to cooperate and will defect.[34] Thus p^* is a critical value for player 1 and defines player 1's attitude toward the risk of being exploited in a trust relationship. The greater p^* is, the more trusting player 1 has to be, to be willing to cooperate in the trust game. The lower p^* is, the greater risk of being exploited player 1 is willing to run in order to try to get mutual cooperation.

The values of the payoffs affect this critical value of trust in a straightforward way. The worse the sucker's payoff S is, the higher the critical value will be. That is, as the payoff for being exploited gets worse for player 1, the more trusting player 1 must be in order to be willing to cooperate. Increasing the reward of mutual cooperation, R,

31. Walt 1996.

32. The Axelrodian equivalent would be P, the punishment for mutual defection. This is normalized here to zero for convenience and to avoid confusion with p, the probability that player 2 is nice.

33. For other versions and extensions of this game, see Lahno 1995a,b; and Güth and Kliemt 1994 and 1998.

34. Coleman calls S the loss (L), R the gain (G), and solves for $p/(1 - p)$, which must exceed L/G for player 1 to cooperate. The result is equivalent.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

has the opposite effect: it lowers the threshold of cooperation. The greater the rewards for mutual cooperation, the greater risk player 1 will be willing to run in an effort to secure them. Thus states that place a higher value on the rewards of mutual cooperation or are less injured if their cooperation goes unrequited will have a lower p^* and will be willing to cooperate at lower levels of trust. A lower value on mutual cooperation or a greater sensitivity to the pain of being exploited will make for a higher p^*, producing cooperation only for higher levels of trust, and defection otherwise.

The trust game is a convenient way to formulate the problem of mistrust rigorously, and it confirms some intuitions about how the interaction between the level of trust and the payoffs affect the possibility of cooperation. Conflict arises because mistrust becomes too high, and there is a critical threshold of trust, p^*, above which the rational thing to do is cooperate, and below which the rational thing to do is defect. The model seems to validate the traditional structural realist gloom about mistrust and cooperation in international relations.

The Reassurance Game

The trust game, however, does not model the idea of reassurance, or building trust between actors. Player 1 has to cooperate fully without learning anything about player 2's type beyond the prior p. Yet, as was detailed earlier, it has often been argued that trust can be established and fostered by small, unilateral cooperative gestures that initiate chains of mutually rewarding behavior. These gestures often involve some vulnerability on the part of the side that makes them, and confer some advantage to the recipient, and can be considered a type of costly signal. In the trust game neither side has an opportunity to make such a gesture. Thus the model does not directly tackle the reassurance question.

To encompass the possibility of trust-building gestures, I modify the trust game by giving the players the opportunity to divide the game into two rounds, a lesser initial round, followed by a final, more important round.[35] This structure enables the actors to predicate their choice in the second round on what happened in the first. If trust can be built through behavior in the first round, then nice types can cooperate in the second round without fear. This modification will make cooperation possible for levels of trust far lower than in the simple trust game. I also include two-sided incomplete information, rather than the one-sided uncertainty of the trust game. This enables an analysis of situations in which each side is fearful of the other, which is more realistic in the context of international relations. With two-sided incomplete information, the model can represent situations in which one side, which does not fully trust its adversary, nonetheless attempts to foster cooperation by reassuring the adversary about its own intentions.

35. Depending on the parameter values, the first round may not always be smaller than the second in equilibrium, but it will be for high levels of mutual distrust, the most interesting case.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

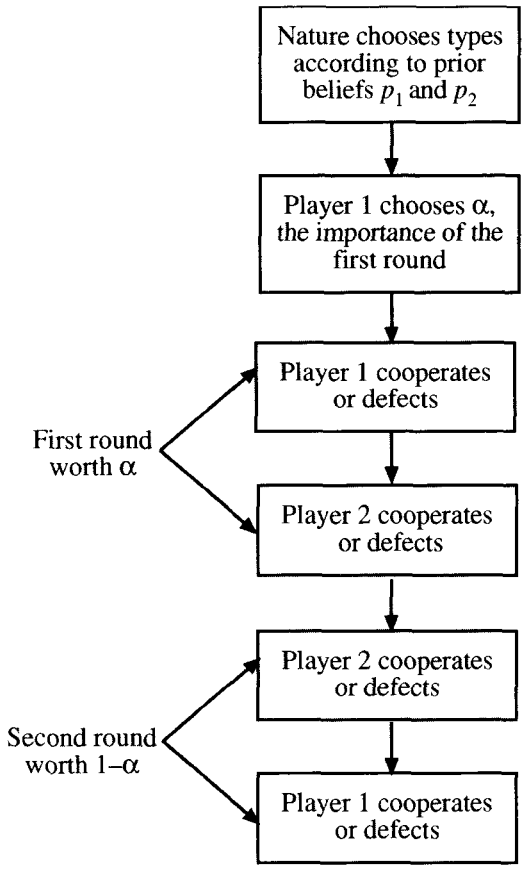


FIGURE 2. *The reassurance game (timeline representation)*

The Structure of the Game

Consider the "reassurance" game presented in timeline form in Figure 2. As before, Nature starts the game by choosing each player's type, with probability p_i that player i is nice. The actors then play two rounds of the trust game.

Before the players get to the trust games, however, player 1 gets to decide the "weight" or importance of the first round as opposed to the second. This choice determines the strength of the signal that will be conveyed by cooperating in the first round. The more important the first round is, the riskier it will be for player 1 to make the first cooperative gesture. The weight of the first round will be denoted α, and that of the second round 1 – α. The greater α is, the more risky it will be to cooperate in the first round; the smaller α is, the less dangerous it will be. This parameter α can be interpreted in several ways. Typically there are a number of issues facing states at any one time. When one state wants to reassure another, it can cooperate on some

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

subset of these issues, hoping that its cooperation will be reciprocated and that the other state will then offer to cooperate on other issues. In 1987, for instance, Gorbachev selected the intermediate-range nuclear forces issue and made several concessions that amounted to an acceptance of the U.S. position. Although significant, this concession did not amount to cooperation on all of the important issues facing the United States and the Soviet Union. The more important strategic nuclear weapons issue, for instance, was unresolved. In terms of the model, Gorbachev cooperated in the INF issue area, worth α, in hopes of sparking cooperation from the United States on the rest of the strategic agenda, worth $1 - \alpha$. Another example is phased agreements. Many international agreements are implemented in phases. In the first phase a certain number of troops are withdrawn, a certain number of tanks destroyed, or a certain reduction in ozone-destroying chemicals is attained. Then in the second phase, the final targets of the agreement are reached; the rest of the troops are withdrawn, the rest of the tanks destroyed, the final acceptable level of emission achieved. In terms of the model, the first phase is worth α to the participants, and the second phase is worth $1 - \alpha$. Note that in either case the more important the first round is (the greater α is), the riskier it will be to cooperate in it; that is, the greater α is, the more costly will be the signal sent by cooperating in the first round.

In the first round, player 1 cooperates or defects, followed by player 2. The payoffs for this round are multiplied by the weight of the first round, α. For instance, if both players are nice and cooperate, they will receive αR_{1N} and αR_{2N}. If player 2 had defected, this player would have received αT_{2N} and player 1 would have received $-\alpha S_{1N}$. Note that simply reducing the scale of the game by multiplying the payoffs by α does not in and of itself make cooperation easier. Going back to the trust game, multiplying all the payoffs by α leaves p^* unchanged; cooperation is no easier or harder than before.

Then the second round of the game, worth $1 - \alpha$, begins. Player 1 having made the first gesture, the onus now shifts to player 2 to move first. I have player 2 move first in the second round so that the players exchange the role of trustor and trustee. If player 1 moved first in both rounds, then player 1 would always be the trustor and player 2 would always be the trustee. In that case, since player 2 would always have the luxury of moving last, it would not matter whether player 2 trusted player 1 or not. Thus player 1's initial cooperative gesture in the first round could not be conceived of as a gesture designed to reassure player 2 and build trust, because player 2's level of trust would be strategically irrelevant. In order to model player 1's initial cooperative gesture as a signal designed in a meaningful way to gain player 2's trust, player 2 must have an opportunity to display that trust in the second round by being the trustor instead of the trustee, that is, by having to move first and assume the risk of defection by player 1. Player 2 therefore starts the second round by cooperating or defecting, and player 1 then ends the game by cooperating or defecting in response. If a nice player 2 cooperated and a mean player 1 defected, the payoffs would be $(1 - \alpha)T_{1M}, -(1 - \alpha)S_{2N}$. Here all payoffs are multiplied by the weight of the second round, $1 - \alpha$.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

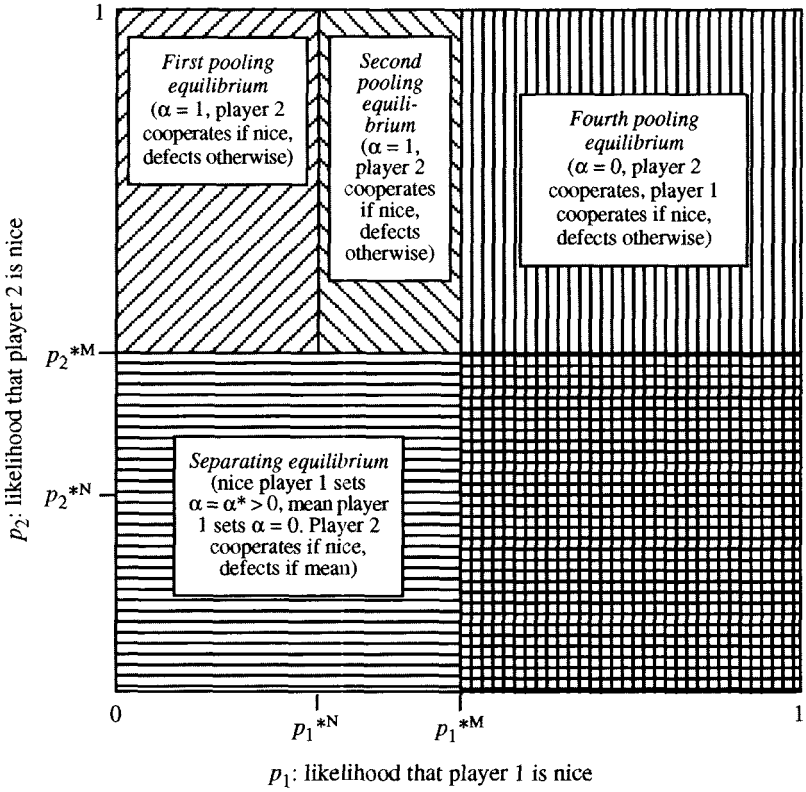


FIGURE 3. *Equilibria in the reassurance game*

Equilibria in the Reassurance Game

The solution concept I employ is the perfect Bayesian equilibrium.[36] This equilibrium concept requires that beliefs be updated on the equilibrium path according to Bayes' rule and does not restrict them off the equilibrium path.[37] There are five classes of perfect Bayesian equilibria in the game, one separating equilibrium and four different types of pooling equilibrium. To get an overview of the different types of equilibria, consider Figure 3. The payoffs in this example are the classic 4, 3, 2, 1 values typically used to illustrate normal form games, normalized around $P = 0$, so that $R_N = T_M = 2$, $T_N = R_M = 1$, $P_N = P_M = 0$, and $S_N = S_M = -1$.

Four equilibria are possible here, three of the pooling equilibria and one separating equilibrium. On the right side of Figure 3 is the fourth pooling equilibrium in which both types of player 1 set α equal to zero, thus making no initial cooperative gesture at all. The second round therefore is the only round, and in this equilibrium player 2

36. Morrow 1994, 170.
37. I discuss off-equilibrium-path beliefs in the appendix.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

cooperates and player 1 responds by cooperating if nice and by defecting if mean. Here, player 2 is known to be trusting enough to cooperate in the second round even without any reassurance about player 1's trustworthiness, so player 1 essentially passes the buck, saddling player 2 with the risk of cooperating first. Note this equilibrium is possible when p_1, the likelihood that player 1 is nice, is relatively high. In the upper-middle portion is the second pooling equilibrium, in which both types of player 1 cooperate fully by setting α to 1. Here, the nice type of player 2 cooperates in response, and the mean type defects. In the upper-left portion is the first pooling equilibrium. Here, player 1 sets α equal to 1 and cooperates fully in the first round. This is again possible because p_2 is relatively high, so player 1 is relatively trusting. Player 2 will then reciprocate by cooperating, if nice.[38] In all three pooling equilibria, nice types cooperate fully with each other in equilibrium. These represent the cases in which one or the other player is relatively trusting and so is willing to assume the risk of cooperation in the full game.

In the bottom half of the figure is the separating equilibrium. Here the nice type of player 1 sets α equal to α^*, some value greater than zero but less than one, and cooperates in the first round. This sends a positive signal designed to reassure player 2. The mean type of player 1 sets α equal to zero, thus failing to send the reassuring signal. Player 2 therefore knows what type of player 1 is being faced when the time comes for player 2 to move. Player 2 cooperates in the first round if nice and defects if mean. In the second round, if player 1 cooperated at the appropriate level in the first round, player 2 will then cooperate in turn, and player 1, if nice, will reciprocate. If player 1 failed to cooperate at the right level, then player 2 will defect. In the separating equilibrium, nice types also end up cooperating fully with each other. The nice player 1 sends the appropriate signal, and the nice player 2 responds by cooperating in the second round, which player 1 reciprocates.

The Separating Equilibrium

Because the separating equilibrium is where reassurance takes place, it is worth considering it in more detail. First I consider how it works and then analyze when it is possible.

The separating equilibrium works essentially because the costliness of the signal, α, has been set at a level too high for the mean type but tolerable for the nice type. For player 1 to cooperate in a portion of the game worth α is to incur a risk that player 2 will defect, leaving player 1 with the sucker's payoff. The greater α is, the greater the risk. The advantage of cooperating, however, is that player 1 persuades player 2 that player 1 is nice, convincing player 2, if also nice, to cooperate in the second round. If α were very small, therefore, the mean type would be tempted to cooperate in the first round as well. Player 1 would undergo a small risk that player 2 would

38. The first and second pooling equilibria differ only off the equilibrium path. Solutions are discussed in detail in a formal supplement to this article entitled Trust, Reassurance, and Cooperation: Formal Supplement, available on request from the author or at <http://wizard.ucr.edu/~akydd>.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

defect in the first round, but, given the small level of α, this would be negligible in comparison with the benefit that player 1 would accrue from potentially fooling a nice player 2 into cooperating in the second round, where player 2 could be exploited. Thus the mean type's preferences set a lower bound on α; α cannot go too low or the mean type will cooperate in the first round in an effort to fool player 2 into cooperating in the second round. Signals with too little cost are "cheap talk," something that both types would engage in, and hence are uninformative.[39] However, the signal size cannot be made arbitrarily large either. The greater α is, the greater the risk of exploitation if player 2 turns out to be mean. Make α too big and the signal will be too risky for the nice type to send, and the equilibrium will not work. Thus the preferences of the nice types put an upper bound on the possible level of α. The secret to making the separating equilibrium work is finding a signal that is adequately costly to deter the mean types from sending it but not so costly that the nice type is afraid to send it.

When can such a signal be found? The key consideration is player 1's level of trust, p_2. An upper bound on player 1's level of trust separates this equilibrium from the pooling equilibria discussed earlier. If p_2 rises above this level, the mean type of player 1 becomes too trusting to sustain the equilibrium strategy of not cooperating at all in the first round. That is, if player 1 is too trusting, the mean type will prefer to cooperate in the first round, regardless of how large α^* is. This upper bound for p_2 is simply the critical value p^* identified earlier in the one-round trust game but in this case using the payoffs for the mean type of player 1, which I denote $p_2{}^{*M}$:

$$p_2{}^{*M} \equiv \frac{S_{1M}}{R_{1M} + S_{1M}}$$

If p_2 exceeds $p_2{}^{*M}$, the mean player 1 would be willing to cooperate regardless of how important the first round is. The separating equilibrium is not possible if the mean player 1 would be willing to cooperate because this player would then deviate to the equilibrium signal for the nice type.

A lower bound for p_2 can also be derived by comparing the upper and lower bounds on the feasible equilibrium signal α^* and solving for when they are compatible.[40] The lower bound is denoted $\underline{p}_2$ and is defined as follows:

$$\underline{p}_2 \equiv \frac{S_{1N}/R_{1N} - S_{1M}/T_{1M}}{(R_{1N} + S_{1N})/R_{1N} - (R_{1M} + S_{1M})/T_{1M}}$$

This lower bound can be lower than p^* from the simple trust game. In fact, it can be zero, or even negative, and therefore pose no constraint at all. In Figure 3 it is not illustrated because with the payoffs used in the example, $\underline{p}_2$ is zero. In this case, therefore, cooperation is possible *regardless* of how fearful the players are of each

39. Farrell and Rabin 1996.
40. Shown in the appendix.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

other in the beginning, because nice types will always be able to reassure each other and build mutual trust. In the one-round trust game, in contrast, cooperation would be possible only if p were greater than p^*, which with these payoffs equals 0.33. The reassurance game, therefore, makes cooperation possible for low levels of trust that would prevent cooperation in the absence of an ability to send costly signals.

The signaling equilibrium will be possible if the upper bound $p_2{}^{*M}$ is greater than the lower bound $\underline{p}_2$. Some algebra demonstrates that this will be the case when the critical value of trust for the mean type of player 1, $p_2{}^{*M}$, exceeds that for the nice type, $p_2{}^{*N}$, defined as follows:

$$p_2{}^{*N} \equiv \frac{S_{1N}}{R_{1N} + S_{1N}}$$

This means that if the nice types are willing to run greater risks in pursuit of cooperation in the simple trust game, so that $p_2{}^{*N} < p_2{}^{*M}$, then the separating equilibrium will be possible for some levels of trust p_2 lower than $p_2{}^{*M}$. Some signal α^* can be found such that the nice types will be willing to cooperate in a first round with that level of importance whereas the mean type will not. On the other hand, if the nice type's critical risk is too high, that is, if $p_2{}^{*N} > p_2{}^{*M}$, the separating equilibrium is impossible because then the lower bound on p_2 will be greater than the upper bound. In general, the higher $p_2{}^{*N}$ *is*, the more risky cooperation is for the nice type and the harder reassurance is to achieve. Conversely, the higher $p_2{}^{*M}$ *is*, the riskier cooperation is for the mean type and the easier reassurance is. This result is summarized in the following proposition.

> *Proposition:* If the nice type is willing to run a greater risk to establish cooperation in a trust game than the mean type is willing to run, such that $p^{*N} < p^{*M}$, then a separating equilibrium in which reassurance takes place is possible in the reassurance game.

By examining the effects of the payoffs on these critical values, we can establish comparative statics that tell us how the payoffs affect the possibility of reassurance. First consider the nice type and $p_2{}^{*N}$. This critical value will fall, making reassurance easier, as the reward for cooperation for the nice type R_{1N} increases and the nice type's penalty for being the sucker, S_{1N}, shrinks. These comparative statics make sense, since the more the nice types value cooperation, the greater risk they should be willing to take to secure it and the easier reassurance should be. On the other hand, raising the mean type's reward for cooperation, R_{1M}, will lower $p_2{}^{*M}$, making reassurance more difficult. This increases the mean type's incentive to mimic the nice type, in this case by increasing the first round payoff from cooperating. Decreasing the cost of being the sucker for the mean type S_{1M} will lower the critical value, also making reassurance more difficult. This is because decreasing S_{1M} increases the mean type's incentive in general to gamble on cooperation in the face of a risk of being exploited.

These last comparative statics, on the mean type's payoffs, highlight an important advantage of using game theory to analyze the reassurance problem. Typically the

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

question of whether reassurance is possible is analyzed in terms of the incentives facing a trustworthy actor who wants to reassure or be reassured by the other side. The incentives facing the nice types are important, of course, but as the preceding analysis makes clear, the incentives facing the mean types are equally important. Reassurance can be made possible or impossible by shifts in the incentives for the mean players alone, even though the mean players, in equilibrium, never end up reassuring anyone. Game theory forces us to take the incentives of the mean types fully into account and to provide an objective analysis incorporating the perspectives of all players and all types.

A final critical feature of the separating equilibrium concerns the equilibrium size of the signal or importance of the first round. As shown in the appendix, the equilibrium weight of the first round, α^*, is an increasing function of the level of trust, p_2. This means that as player 1 grows more trusting of player 2, the importance of the first round must increase in equilibrium. Conversely, the more fearful player 1 is of player 2, the smaller the equilibrium signal must be. This lends support to the intuition, formulated by Osgood in the GRIT strategy, that in situations of extreme mistrust it is necessary to begin with small cooperative gestures and then gradually proceed to more important issues as the level of trust is built up over time. The trajectory of the Israeli-Palestinian peace process, the end of the Cold War, and many other cases attest to the empirical relevance of this idea.

Summation of the Reassurance Game

Reassurance is accomplished by costly signaling. In order to work, the signals must be costly, but not too costly. Make them too easy and they become cheap talk, which untrustworthy types will send as well. Hitler's claim that he had no more territorial demands to make in Europe is a classic example. Such claims are unpersuasive because there is nothing to prevent untrustworthy types from making them. However, the signals cannot be made too costly either, or the nice types will be too fearful to send them. A signal that meets these two constraints can be found if the nice type is willing to take greater risks to establish mutual cooperation than the mean type. That is, if the nice type is willing to run a greater risk to secure the nice type's most preferred payoff, R_{1N}, than the mean type is to secure the mean type's second best payoff, R_{1M}, then the reassurance equilibrium is possible. In this case, cooperation may be possible for levels of trust that are quite low—low enough to preclude cooperation in the simple one-round trust game. The reassurance game, therefore, powerfully reaffirms the rationality of reassurance. Giving player 1 the chance to move first with a costly signal makes cooperation possible even in the context of a much lower level of trust.

Reassurance and the End of the Cold War

To provide some empirical grounding for the reassurance game, I trace the evolution of U.S.–Soviet relations in the late 1980s, with special attention to Soviet gestures

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

designed to reassure the United States and to their impact on U.S. perceptions of Soviet motivations. I argue that the decisive events that ended the Cold War can be interpreted as costly signals in the framework of the reassurance game. Several events stand out in the period from 1985 to 1990 as particularly clear examples of signals by the Soviets that their motivations had changed. These include the 1987 INF treaty, the 1988 withdrawal from Afghanistan and announcement of conventional force reductions, and the 1989 revolutions in Eastern Europe.

The end of the Cold War is a story that has often been told; the events belong to the recent and familiar past and there is no lack of conflicting interpretations.[41] My goal here is not to bring new facts to light or to present a completely new interpretation, since the reassurance/GRIT aspect of the tale is itself twice told. Richard Bitzinger argues that Gorbachev employed a classic GRIT strategy but that GRIT was irrelevant to the end of the Cold War, which was a simple product of Soviet decline.[42] Alan Collins counters that GRIT was instrumental in ending the Cold War.[43] The aim here is to see what additional purchase on the case the costly signaling reassurance theory can provide while using the case to illustrate and render plausible the main theoretical implications of that theory.

Toward that end I examine the initiatives by Gorbachev during the latter half of the 1980s and how these affected Western beliefs about the Soviet Union. In doing so, we are immediately confronted by the fact that reality, as always, is more complicated than the model. Two complications are especially salient. First, there were many different actors in the West who held different opinions of Soviet motivations, and these opinions evolved at different rates. Second, instead of one round of reassurance that resolves all doubts, there were several rounds of reassurance over a period of several years. Although this departs from a literal reading of the simple bilateral model offered earlier, I think the interpretive leap is not too great. Rather than focus on a single actor, such as President Reagan, I try to reflect a broader sample of opinion, culled from historical accounts, memoirs of important officials, and the opinions of prominent foreign policy commentators and newspapers such as the *New York Times* and the *Economist*. The main trend of opinion, across a remarkably broad spectrum of opinion from the liberal left to conservatives such as Reagan and Secretary of State George Schultz, is in accord with the model's prediction of increasing trust in response to the costly signals. Looking more closely, we can see that more conservative observers were slower to change their beliefs and sometimes took several rounds of reassurance before they were fully convinced. This may be accounted for by differing prior levels of trust. Conservatives, starting with very low prior levels of trust, had more "distance" to cover before they could come to believe that Gorbachev was sincere in his desire for a better relationship, and hence it took several rounds of reassurance before their posterior levels of trust reached the levels that more liberal observers reached earlier in the process. Thus I think the model, simple

41. For collections of essays, see Allan and Goldman 1992; Hogan 1992; and Lebow and Risse-Kappen 1995.

42. Bitzinger 1994.

43. Collins 1998.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

as it is, does account for the main effects of what transpired and is helpful in addressing some of the complications arising from multiple audiences and multiple stages.

The model distinguishes between a mean, or untrustworthy, type of actor and the nice, or trustworthy, type. The dilemma is that each side fears that the other may be mean and so is afraid to initiate cooperation. In the context of the Soviet political spectrum of the 1980s, the mean type corresponds to the old-style, Brezhnevite Soviet leader who was interested in exploiting the West and viewed U.S.–Soviet competition as a zero-sum game in which the correlation of forces would eventually favor the Soviet Union. The nice type corresponds to the new breed of Soviet leader such as Gorbachev and Foreign Minister Eduard Shevardnadze, imbued with "new thinking," who favored a more cooperative relationship. The origins, adoption, and eventual triumph of this new thinking is one of the most important aspects of the end of the Cold War. Thomas Risse-Kappen argues that a set of ideas under the rubric "common security" arose in the Western European peace research institutes and Social Democratic parties that emphasized the unwinnability of war in the nuclear age and the severity of the security dilemma, in which no state could achieve security through unilateral means.[44] These ideas were transmitted to Russian scholars though international conferences and to Russian reform-minded politicians as well. When Gorbachev came to power, these individuals came increasingly to positions of prominence, securing the dominance of the new thinking. Sarah Mendelson interprets the Soviet decision to withdraw from Afghanistan with reference to an epistemic community of these new thinkers that triumphed in the political process.[45] In the context of the theory presented here, the crucial new thinker/old thinker dichotomy corresponds to the nice/mean player typology of the model.

In the reassurance game, reassurance takes place in the separating equilibrium. For the separating equilibrium to be possible, it must be the case that the critical value for trust for the nice type must be lower than that for the mean type, that is $p^{*N} < p^{*M}$. Translating this to the Cold War, the new-thinking Soviet leader must be willing to cooperate at a level of trust low enough to cause an old-thinking communist true believer to defect. This could be the case if the new thinker's payoff for achieving mutual cooperation is higher than the old thinker's, or if the sucker's payoff is not as bad (relative to the noncooperation outcome, here normalized to zero.) An argument can be made that both of these conditions held in the case of Gorbachev and the new thinkers. Perhaps the most important factor driving Gorbachev was a desire to place a greater priority on the domestic economy and satisfy the wants of the Soviet people. Indeed, some analysts claim that Gorbachev's beliefs on foreign policy were relatively unformed in comparison to his analysis of what needed to be done domestically when he took office.[46] As a result, international cooperation was seen as crucial in order to provide a conducive environment for facilitating domestic reform. In terms of the model, the rewards for mutual cooperation, R, were perceived to be

44. Risse-Kappen 1994.

45. Mendelson 1993. For related arguments about ideas and institutions in Gorbachev's Russia, see Checkel 1993.

46. Stein 1994.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

much more desirable by the new administration than by previous Soviet leaders. Also, the danger of being exploited by the West was probably perceived as of lesser concern by the new administration than by previous ones, because the international system was no longer viewed in harsh, zero-sum terms. The essence of the core concept of the new thinking, common security, after all, was that security would be achieved mutually or not at all. The old emphasis on strategic nuclear and conventional superiority was being replaced by notions of "reasonable sufficiency," essentially a defensive philosophy.[47] The Soviet Union could therefore survive a few unrequited cooperative gestures, given the robustness of deterrence and the overall strength of the country's military position. Thus, I would argue that the new Soviet leadership viewed the gains of mutual cooperation as greater and the cost of exploitation as lesser than previous Soviet leaderships, which suffices to make them willing to run greater risks to secure cooperation and, in turn, makes the separating equilibrium possible.

In March 1985, Gorbachev assumed power in the Soviet Union with a commitment to reform. In the early months of his tenure, however, reform meant "acceleration," a quickening of economic growth without fundamental change. Andropovian efforts to reinstill discipline and cutback on drunkenness by reducing the sale of alcohol were among the practical steps taken.[48] In the international sphere, Gorbachev made two arms control gestures: a moratorium on nuclear testing and one on SS20 intermediate-range ballistic missile (IRBM) deployments. These moves were ineffective in changing opinions in the West. Bitzinger argues that this was a failure for GRIT because the United States failed to cooperate in response.[49] Collins and Joshua Goldstein and John Freeman counter that although there was no direct response, these gestures did "change the climate" in a more subtle way and set the stage for later, more important gestures.[50] In light of the theory here, these gestures can be usefully conceived of as signals that fall short of the crucial level, α^*, that the nice type must send to reassure the other side. In other words, these signals were "cheap talk" rather than "costly signals." As the model demonstrates, signals that are too low fail to reassure and hence do not elicit cooperation from the other side. The Soviet moves in 1985 were interpreted in just these terms. The Soviets were believed to have completed a nuclear testing cycle and deployed most of their IRBMs, so the moratoriums seemed hollow. An IRBM freeze would have locked in a huge Soviet numerical advantage and was viewed as an attempt to retard the process of deployment in the West. Thus these signals failed, largely because they were regarded as moves that did not really hurt the Soviets; that is, they were signals with little or no cost.[51]

The first U.S.–Soviet summit of the Reagan administration was also held in 1985. President Reagan's attitudes toward the Soviet Union had already evolved somewhat since his first inauguration. By late 1983 he had come to the conclusion that Soviet

47. Gartoff 1990.
48. Thom 1989, 31.
49. Bitzinger 1994, 75.
50. See Collins 1998, 205; and Goldstein and Freeman 1990, 116.
51. Garthoff 1994, 213–14.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

leaders were genuinely afraid of the United States and that therefore a personal meeting in which he could reassure them would be desirable.[52] As the summit approached he began to speak of "misunderstandings" that could be cleared up, rather than of the irrevocably evil nature of the Soviets.[53] The summit appears to have improved Reagan's image of the Soviets somewhat, yet progress was limited. Agreement was reached "in principle" on 50 percent reductions in strategic nuclear weapons and on an interim INF agreement, details unspecified. The Soviets also seemed to soften their position on Afghanistan. But in the end, fundamental attitudes appeared to remain unchanged, though surface tensions were lessened.

The Reykjavik summit of October 1986 was far more dramatic, yet ultimately no more successful. Gorbachev made dramatic concessions in the negotiations on strategic and intermediate-range nuclear forces. In a process of one-upmanship the two leaders moved from agreements to eliminate IRBMs in Europe and cut strategic forces by 50 percent to a rather visionary program to eliminate all nuclear weapons in ten years. For Gorbachev, however, the linchpin was adhering to the Anti-Ballistic Missile (ABM) treaty and limiting the Strategic Defense Initiative (SDI) to the laboratory. Reagan balked at this final element of the deal. The summit ended without agreement and with considerable anger on Reagan's part.[54] Though Reykjavik was an intense encounter, it would appear to have had little impact on U.S. evaluations of Soviet intentions.[55] In January 1987 a White House policy paper still postulated that, "Moscow seeks to alter the existing international system and establish Soviet global hegemony."[56] What little change in Western opinion occurred stemmed more from the increasing openness of Soviet domestic affairs, as indicated by such events as the release from internal exile of the prominent dissident Andrei Sakharov.[57]

In 1987 the Soviets made their first important costly signal in the form of the INF treaty. On 28 February Gorbachev picked out the INF portion of the Reykjavik deal and announced that he was willing to go ahead with it by itself, with no reference to SDI. This essentially represented acceptance of the "zero option" proposed by Reagan in 1981 and hence was a substantial victory for the United States. Negotiations continued through the year on the details of the agreement and concluded in November. The Soviets agreed to destroy far more missiles than the United States did (1,846, as opposed to 848 by the United States).[58] Even more of a departure from Soviet practice were the intrusive verification procedures agreed upon. Past Soviet objections to verification by any means other than "national technical" (spy satellites) were abandoned and on-site inspections were allowed.

Critics of the reassurance perspective argue that the INF treaty was a product of simple tough bargaining by the West rather than a reassurance strategy by Gor-

52. See Reagan 1990, 588; Oberdorfer 1991, 15; and Schultz 1993, 164.
53. Garthoff 1994, 235.
54. See Reagan 1990, 667; Oberdorfer 1991, chap. 5; and Schultz 1993, 757.
55. Garthoff 1994, 289.
56. Ibid., 308.
57. Kaiser 1992, 146–49.
58. Garthoff 1994, 327.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

bachev.[59] However, as Risse-Kappen argues, it is extremely unlikely that the INF treaty would have emerged in anything like its actual form without the regime change in the Soviet Union and the consequent shift to new thinking.[60] Although the Soviets had clearly failed to prevent NATO from deploying its own intermediate-range forces, there was nothing preventing it from living with the status quo of both sides having them, the defect-defect outcome as it were. Instead, Gorbachev seized on the INF issue as a vehicle for reassurance. Particularly notable is that the signal was of an intermediate level of importance, much like the missiles it concerned. Getting rid of an entire class of missiles served as an undeniably costly signal, without yet addressing the entire range of issues confronting the East–West relationship, including the peak arms control issue of strategic nuclear weapons. Thus the signal size, α^*, was positive but considerably less than 1 in this case.

The INF treaty and the Washington summit that followed began to shift Western opinion toward the Soviets. Some evidence suggests that Reagan still viewed Gorbachev as simply making the best of a bad situation, that is, not fundamentally less aggressive in motivation but just more realistic about Soviet limitations and the failures of communism.[61] On the other hand, Secretary of State Schultz claims that Reagan realized that Gorbachev "represented a powerful drive for a different Soviet Union in its foreign policy."[62] Schultz's own attitude seems to have been decisively affected by late 1987. On 6 November, in a meeting with CIA analysts on the Soviet Union, Robert Gates described Gorbachev as a Leninist who was merely trying to secure breathing space for a future round of conflict. Schultz disagreed and later wrote, "I felt that a profound, historic shift was underway; the Soviet Union was, willingly or unwillingly, consciously or not, turning a corner; they were not just resting for round two of the Cold War."[63] He pointed to the Soviet's desire to leave Afghanistan and their diminishing role in other regional trouble spots as well as the INF agreement. Thus by late 1987 Schultz would seem to have come to the conclusion that Soviet motivations were changing.

The treaty was signed at the Washington summit in December 1987, an event with considerable public relations impact of its own. "Gorbymania" had taken hold of the public at large; Gorbachev had a 65 percent approval rating from the American people and was more popular in some European countries than Reagan.[64] Gorbachev's personal style, his ability to mingle with crowds, and his fresh approach to international security issues won him popular approval and began to change attitudes toward the Soviet Union. A debate also began in the media in which some began to consider the possibility of serious Soviet change. Some commentators, even Richard Pipes, pointed to the intrusive verification measures as indicating a fundamental change of Soviet

59. Bitzinger 1994, 77.
60. Risse-Kappen 1991.
61. See Reagan 1990, 703; and Garthoff 1994, 332.
62. Schultz 1993, 1015.
63. Ibid., 1003.
64. The Gorbachev Effect, *The Economist,* 27 February 1988, 38. For an analysis of public opinion in reaction to these events, see Peffley and Hurwitz 1992.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

policy for the better.[65] However, other voices in the press were more skeptical. On 2 December the *Washington Post* carried an editorial acknowledging some concessions on Gorbachev's part but expressing skepticism about his motives. David Broder in the same issue warned that Gorbachev was good at public relations but not to be trusted. Charles Krauthammer the next day equated Gorbachev's values with those of Brezhnev. The *New York Times* was similarly circumspect, generally positive toward the treaty yet skeptical of Gorbachev.[66] Some critics of the treaty, including former president Richard Nixon and many conservative senators, worried about the large Soviet conventional advantage that would remain after the missiles were gone.[67] Reagan himself began to be the object of vitriolic assault from the right, and he struck back by accusing his attackers of believing that war with the Soviets was inevitable.[68]

These suspicions were further undermined in 1988. Three events are of special importance: the beginning of the Soviet withdrawal from Afghanistan, the Moscow summit in May and June of 1988, and Gorbachev's announcement of conventional force reductions in a December speech to the UN. In February Gorbachev announced that the Soviet Union would withdraw from Afghanistan upon the conclusion of international negotiations, which were duly completed in April 1988. This development, like the INF treaty, can be interpreted as a costly signal in accordance with the reassurance model. To the extent that Afghanistan was a pawn in the superpower struggle, letting it go served both to reduce the Soviet threat to the West and to demonstrate a lack of territorial ambitions. As Mendelson argues, "After the summit at Reykjavik, Gorbachev and his advisers came to the conclusion that the United States would not entertain seriously the idea of new political thinking until a Soviet withdrawal from Afghanistan was complete."[69] Coming on the heels of the INF treaty, the withdrawal began to build credibility for the notion that the Soviet new thinking on security was for real. On 19 February Robert Manning argued in an op-ed piece in the *New York Times* that the withdrawal from Afghanistan will "begin to render credible Moscow's 'new thinking' about the Soviet role in the world."[70] Robert Kaiser underlined the importance of the move: "Nothing he could have done made a stronger statement than that decision—it proved that he meant what he said about 'new thinking.'"[71] Even conservative commentator Jeanne Kirkpatrick argued that it constituted "dramatic evidence of Gorbachev's determination to de-emphasize military action and also to change the Soviet image" while explicitly denying that it

65. Gary Lee, INF Treaty Shows Impact of Gorbachev, *Washington Post*, 29 November 1987, A1.
66. Gorbachev, The Movie and the Reality, *New York Times*, 2 December 1987, A34.
67. Schultz 1993, 1007.
68. Ibid. However, Reagan's endorsement of the treaty seems to have defanged much conservative public opinion; see Sigelman 1990.
69. Mendelson 1993, 356.
70. Robert A. Manning, Exorcising Brezhnev's Foreign Policy, *New York Times*, 19 February 1988, A35.
71. Robert G. Kaiser, Finally the New Soviet Man Appears, *Washington Post*, 24 May 1988, A23; and Kaiser 1992, 259–60.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

was economically motivated.[72] Once more, however, the interpretation that the Soviets acted from weakness had some adherents. The *Economist* claimed that, "His whole programme—the economic reforms, the diplomatic boldness—is intended to make his country a more formidable adversary for the West, not a partner with it."[73] The *Washington Post* weighed in with a skeptical editorial on 9 February that contained no praise for Gorbachev.

The second event of importance was the Moscow summit in May and June. For the first time, Reagan had a broad and direct exposure to the Soviet Union. He met with a wide swath of Soviet society, among whom were Soviet officials, dissidents, Orthodox clergy, and university students. In the process of attempting to project a benign image of the United States, he appears to have developed a more benign image of the Soviets. In a famous interchange, Reagan was asked if he still held to the idea that the Soviet Union was an evil empire, and he responded, "No, I was talking about another time, another era."[74] This marked increasing movement toward the view that Soviet intentions had fundamentally altered; after all, at this point the Soviet Union was still an empire, if a less evil one. In his memoir Reagan notes how he realized that "the world was changing."[75]

The third major event of 1988 was Gorbachev's December speech at the UN in which he promised substantial force reductions in Eastern Europe. Traditional Soviet strategy mandated that should war break out for whatever reason, the Red Army would launch an overwhelming offensive with the object of pushing NATO forces off the continent. Throughout much of the Cold War the Soviets had a strong numerical advantage in several categories of conventional weaponry and were generally thought to pose a severe threat to NATO.[76] Gorbachev decided, relatively early in his administration, to abandon this long-standing pillar of Soviet strategy and adopt a policy of "defensive defense."[77] This approach mandated attempting to hold the line in Central Europe if attacked while mobilizing additional forces and attempting to negotiate a speedy conclusion to the war. Such a strategy had two advantages. It was significantly less costly, since it required far fewer conventional forces, and it was much less threatening to potential adversaries. However, the strategy change challenged vested military interests in the traditional offensive doctrine. As several scholars have argued, military bureaucracies have an affinity for offensive doctrine in part because of the larger resources needed to carry them out.[78] Thus, for a long time this policy could not be fully implemented because of opposition from the military. Only

72. Kirkpatrick 1990, 188. In a column from November 1989 Kirkpatrick described the withdrawal as very much in the costly signal vein, and though she was somewhat moved by it, she still inferred a Soviet desire for a defenseless Western Europe. She closed by calling for more signals. Jeanne Kirkpatrick, Change and Chutzpah, *Washington Post*, 1 November 1989, A25.

73. As Russia Retreats, *Economist*, 16 April 1988, 13.

74. Garthoff 1994, 352

75. Reagan 1990, 711.

76. Some academics disputed this conventional wisdom, arguing that NATO would be able to hold the line or even prevail in a conventional war; see Cohen 1988; Mearsheimer 1982; and Posen 1984a.

77. Garthoff 1990.

78. See Posen 1984b; Snyder 1984; and Sagan 1994.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

in the spring of 1987, in the wake of Matthias Rust's embarrassing landing in Red Square, did Gorbachev begin to consolidate his hold on the military through large-scale personnel changes.

In December 1988 Gorbachev felt able to move ahead. In a speech to the UN he announced a unilateral troop reduction of 500,000 men, including six tank divisions, in central Europe. The significance of the reduction lay in four aspects. First, it was sizable and could not readily be dismissed as a propaganda device. The conventional force reduction could not be easily explained as a feint or ploy, as the 1985 nuclear testing moratorium was—α was definitely greater than zero. Second, it lent credence to the notion that the Soviets were indeed shifting to a strategy of "defensive defense" and thus posed much less of a threat to NATO. Third, by removing troops from Eastern Europe it began to undermine the fragile Eastern European communist regimes and signaled a Soviet disengagement from the region. Finally, it answered the Western critics of INF who were concerned about Soviet conventional superiority in a less nuclear Europe.

Although some negative press reaction remained, opinion was definitely shifting in a positive direction. On the one hand, the *Washington Post* opined that Gorbachev was still motivated by economic factors and had not shown himself a "reliable partner." Conservatives blasted the move in op-ed pieces; for example, George Will railed against an "epidemic of complacency," and the *Economist* asserted once again that economic factors were the prime motivator and that if the Soviet Union were to recover under Gorbachev it would be a greater threat than ever.[79] However, these voices were becoming atypical. Admiral William T. Crowe, chairman of the Joint Chiefs of Staff, asserted that the move "lent credence" to Soviet moves toward a defensive strategy. Other military experts declared that the move would make a surprise attack impossible, and the *New York Times* called it an "invitation the West can't ignore."[80] As Bill Keller reported in the *New York Times*, "As the Soviet leader prepares for his second trip to America, Western curiosity and skepticism seem to have shifted from his intentions to his prospects: not 'Does he mean it?' but 'Can he pull it off?'"[81] Margaret Thatcher had declared the Cold War over on the eve of the UN speech. Public opinion polls showed that fear of the Soviet Union plummeted in 1988.[82]

The process of reassurance was essentially completed in 1989. After an initial "pause" in U.S.–Soviet relations when the Bush administration took office, the relationship developed quickly. NSD-23, drafted in March 1989, coined the phrase "beyond containment" to summarize the new U.S. policy. The document suggested that it might be possible to shift to a strategy that "actively promotes the integration of the Soviet Union into the international system."[83] The United States proposed cut-

79. See Unfortunately, 1917 Did Happen, *Economist*, 17 December 1988, 12–13; and George Will, Epidemic of Complacency, *Washington Post*, 8 December 1988, A27.

80. See Jonathan Dean, On Arms: We Need the Details, *Washington Post*, 11 December 1988, C7; and An Invitation We Can't Ignore, *New York Times*, 11 December 1988, E24.

81. Bill Keller, Gorbachev's Grand Plan: Is It Real or a Pipe Dream? *New York Times*, 5 December 1988, A1.

82. Richman 1991.

83. Beschloss and Talbott 1993, 69.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

ting U.S. troops stationed in Europe by 20 percent and limiting troops from each side to 275,000.[84] In a very significant indicator of the new level of trust, the Bush administration began to press for an accord on chemical weapons without the usual emphasis on verification. When General Scowcroft, President Bush's national security advisor, objected, Bush responded, "My gut just tells me that the danger of proliferation is more important than the risk of Soviet cheating." The Soviets themselves realized the importance of the development. Sergei Tarasenko told Eduard Shevardnadze, the Soviet foreign minister, "This is really something new and important. The Americans are no longer quite so obsessed with our cheating."[85] This shift is crucial because the concern for verification is a relatively good indicator of the level of trust between two states. The less each side trusts the other, the more extensive will verification procedures have to be to enable cooperation to take place. The more trust that exists, the more relaxed verification standards can be because the other side is assumed to have less desire to cheat and hence less willingness to expend resources to deceive the verification regime.

The rapprochement accelerated tremendously in the summer and autumn as the Eastern European regimes crumbled one by one and the Soviets permitted them to fall. Gorbachev had begun to reformulate policy toward Eastern Europe in a speech celebrating the seventieth anniversary of the Bolshevik Revolution in November 1987.[86] He spoke of the need to reexamine the Soviet invasion of Czechoslovakia in 1968 and asserted that the Soviet Union no longer viewed itself as a model that had to be followed by all socialist states. The issue lay dormant, however, though Gorbachev did occasionally press the Eastern European leaderships to introduce glasnost and reform themselves. The ball began to roll when Solidarity returned to an active role in late 1988 and early 1989. In January 1989 the Soviet press commented favorably on the beginning of multiparty politics in Poland. Finally, in April Gorbachev explicitly rejected the use of force in Eastern Europe. This set the stage for the cataclysm in the fall of 1989. As regime after regime fell to peaceful protest, Western observers were stunned. When the Berlin Wall fell on 9 November, Bush acknowledged, with legendary restraint, that the Soviets were "more serious than I realized."[87] The fall of the Berlin Wall also affected press attitudes. As late as October 1989, conservative columnists were asserting that "the Russians are still coming" and that Gorbachev would most likely become a hard-line dictator.[88] After the wall fell, few questioned Gorbachev's desire to end the Cold War, and some explicitly acknowledged that he had abandoned the old Soviet goal of hegemony.[89] Bush was openly chided for his mild response to the breach of the wall, however justified it

84. Ibid., 77.
85. Ibid., 120.
86. Chafetz 1993, 73.
87. Beschloss and Talbott 1993, 132.
88. See Richard Pipes, The Russians Are Still Coming, *New York Times*, 9 October 1989, A17; and William Safire, Goodbye to Glasnost, *New York Times*, 19 October 1989, A17.
89. See David Broder, Our Great Mission in Europe, *Washington Post*, 15 November 1989, A21; East Germany's Great Awakening, *New York Times*, 10 November 1989, A36; and Hobart Rowen, Transitions East and West, *Washington Post*, 16 November 1989, A27.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

might have been by a desire to avoid undermining Gorbachev by gloating over the collapse of communism.[90]

In looking at the end of the Cold War, then, one can observe a series of costly signals leading to mutual trust between former adversaries. The attitudes of Western leaders, press, and publics toward the Soviet Union all underwent a substantial transformation. Soviet military and geopolitical concessions, particularly the INF treaty, the withdrawal from Afghanistan, the December 1988 conventional arms initiative, and the withdrawal from Eastern Europe were decisive in changing overall Western opinion about the Soviet Union. By 1990 most observers viewed the Soviet Union as a state that had abandoned its hegemonic ambitions and could be trusted to abide by reasonably verified arms control agreements and play a constructive role in world politics. The Cold War was over, not simply in remission pending a Soviet economic recovery, well before the breakup of the Soviet Union in December 1991.

Although reassurance is an essential aspect of the end of the Cold War, I do not mean to claim that it is the only factor. Soviet concessions were inspired partly by weakness, as some scholars have argued.[91] A perception of fundamental weakness was undoubtedly an essential goad to Gorbachev in his reevaluation of Soviet priorities and goals and an essential lever in his internal struggles with those who opposed his policy of accommodation with the West. The fact that *some* form of retrenchment was obviously unavoidable enabled Gorbachev to gain acquiescence for his radical reorientation of Soviet policy. I would argue, however, that economic weakness alone did not dictate the depth of rapprochement with the West achieved by Gorbachev.[92] Military spending could have been reduced while keeping a tight grip on Eastern Europe and the non-Russian republics of the Soviet Union. A type of authoritarian capitalist reform could have been attempted, much as China has implemented, that promised economic growth and political stability. Such a path would have presented much more of an enduring threat to NATO than the path Gorbachev chose. China's relations with its East Asian neighbors and the United States are a reminder that a transition path from communism exists that does not provide a great deal of reassurance to the outside world. In short, although some retrenchment was mandated by economic constraints, the *nature* and *extent* of Gorbachev's reforms are not predicted by economic factors alone. Gorbachev's apparent final goal for the Soviet Union—a socialist, democratic, multinational state participating fully in the global economic system—was not dictated by economic decline (nor, as it turned out, was it feasible).

Finally, I would also like to emphasize the value added by the reassurance/signaling approach beyond this discussion of "new thinkers" and the epistemic community approach to which it is related. Writers such as Risse-Kappen, Mendelson, and Jeff Checkel emphasize the role of ideas in explaining the end of the Cold War. They focus on the origins of these ideas about common security, the way they were introduced to the Soviet political system, and the way they infiltrated the institutions

90. Beschloss and Talbott 1993, 135.
91. See, for example, Wohlforth 1995; Blacker 1993.
92. Risse-Kappen 1994, 190.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

of power in the Soviet Union and eventually became predominant. As Risse-Kappen observes, ideas do not float freely, they must be adopted by actors inside appropriate institutions that have the power to turn them into policy. Although ideas do not float freely, it is also important to point out that they do not interact strategically either. The possession of a more benign security worldview on the part of Soviet leaders does not assure that other states will believe mere assertions to this effect and resolve all disputes. Risse-Kappen points out that the Soviet adoption of common security rhetoric was eagerly received in Western Europe, the home of these ideas, but was greeted with skepticism in the United States, which had no particular allegiance to them, at least as phrased in the European lexicon. Indeed, the European origin of these ideas only raised the suspicion that the Soviets were up to a more clever version of their old game, "split the alliance." The result was that a significant signaling process, involving costly signals and not cheap talk, was necessary before the United States would credit the Soviet Union with having actually turned a corner and become less threatening. Thus although the origins, dissemination, and institutionalization of common security ideas are essential parts of the end of the Cold War story, so too is the *credible commitment* to these ideas the demonstrable proof that one actually has adopted them and can therefore be trusted.

An Extension: The Case of Ethnic Conflict

Before I conclude, I would like to briefly discuss the implications of this analysis for the case of ethnic conflict. Students of ethnic conflict have also focused on the concepts of fear and mistrust to explain the origin and continuation of the civil wars that have plagued the post–Cold War world.[93] In a widely influential article, Barry Posen applied the security dilemma framework to ethnic conflict, arguing that ethnic conflict was the result of the vulnerability of ethnic groups in times of anarchy following the failure of the state.[94] Many scholars have taken up this idea and the related spiral model of conflict and developed them in the context of ethnic war.[95] Barry Weingast developed a model of mistrust and argues that it explains two puzzles about ethnic conflict.[96] One, which he labels the economic puzzle, is that ethnic conflict happens in spite of the fact that it is very costly for both sides. The second he calls the political puzzle, which has to do with the timing of the onset of ethnic conflict or why ethnic groups go from relative quiescence to extreme conflict in such short order.

These arguments, like their counterparts in the international arena, have largely ignored the issue of reassurance, even though it would seem to be central. The model developed here shows that reassurance is quite rational in equilibrium; costly signals that demonstrate one's lack of aggressive motivations should often serve to diffuse

93. See Walter 1997; and Lake and Rothchild 1996.
94. Posen 1993.
95. See Van Evera 1994; Kaufman 1996; Walter 1997; and Lake and Rothchild 1998.
96. Weingast 1998.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

conflict. Thus explanations of ethnic conflict that focus on mistrust must also be refined to take reassurance into account. Weingast's model, for instance, is much like the trust game presented here in which reassurance is impossible to consider because it is ruled out by the structure of the game. The rationality of reassurance suggests that the ethnic conflicts that occur may actually be attributable to nonrational behavior on the part of the various actors or to the presence of genuinely aggressive and hateful ethnic groups. After all, in many cases the fear of ethnic persecution is not mistaken but well justified, as in the case of the Tutsi victims of the genocide in Rwanda. Nonrational behavior and beliefs also play a role in ethnic conflict. The cultivation of a "victim mentality," notably by the Serbs in the context of the ethnic wars in the former Yugoslavia, often justifies aggressive action against others who pose no real threat. Another reason why reassurance may be difficult to achieve in the ethnic conflict case is the presence of leaders who wish to stymie the process for their own political ends. Rui de Figueiredo and Weingast address this question in a model of ethnic conflict where leaders and followers have different incentives.[97] Leaders, wanting to retain office, have an interest in manipulating mass-level beliefs, making the masses more fearful. Thus false beliefs among the masses are not resolved, because a leader, who does not suffer the costs of the unnecessary war, does not wish them to be resolved.

In the ethnic conflict case, then, as well as with international conflict, mistrust-based explanations of conflict need to be refined. The simple idea that mistrust causes conflict in a straightforward way needs to be jettisoned in favor of a more nuanced approach that starts from a realization that the rational response to mistrust between two actors is often reassurance, not conflict. Conflict is more likely when the mistrust is actually justified, and hence impossible to dispel, or if there are psychological or other barriers to dispelling it.

Conclusion

The overall implication of the model and case study presented here is that reassurance can rationally overcome mistrust and lead to cooperation. When the trust game is modified to incorporate the possibility of reassurance, the scope of cooperation increases greatly, in that actors who would be too fearful to cooperate in the trust game find it rational to reassure each other and cooperate in the reassurance game. This process can be seen at work at the end of the Cold War, in the signals made by Gorbachev, and in the way Western perceptions of the Soviet Union changed in response. By sending costly signals to the other side, trust can be built.

This is an optimistic message for those interested in conflict resolution. As James Fearon and David Laitin point out for ethnic conflict, the proportion of potential ethnic conflicts that become actual is quite small.[98] Much the same can be said for

97. de Figueiredo and Weingast 1999.
98. Fearon and Laitin 1996.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

international conflict. Even states with ongoing disputes are rarely at war. The model is in accord with these facts; most of the time states should be able to reassure each other and allay unfounded fears, and only rarely will this process fail to work. Tremendous concern should obviously be focused on those cases where it fails, but the failures should not obscure our view of the day-to-day successes that sustain normal cooperative relationships.

The model also provides a theoretical underpinning for a developing body of work within security studies on cooperative security, confidence-building measures, and security communities. In the wake of the Cold War's end, renewed attention has been focused on security strategies based more on reassurance than on deterrence. Much empirical work has been done on reassurance tasks undertaken by the UN, OSCE, NATO's Partnership for Peace, and on confidence-building measures more generally.[99] However, with the exception of the advent of the offense/defense literature in the 1970s,[100] the theoretical basis for this approach has been largely undeveloped since Osgood and Etzioni advanced their basic reassurance hypotheses in the early 1960s. This lacuna is especially apparent in comparison with the high level of theoretical sophistication achieved by the related literatures on deterrence and crisis bargaining. The reassurance game presented here is a first step in providing a rational choice foundation for this literature. Many issues remain to be explored, such as problems introduced by the existence of multiple actors and mass publics, asymmetries between the actors, and bounded rationality. But as Elinor Ostrom has recently argued, trust is one of the core concepts in human relations and it needs to be more fully integrated into our theoretical understanding of political action.[101]

Appendix

The assumption about off-equilibrium-path beliefs underlying the depiction of the equilibrium boundaries in Figure 3 is that signals larger than the nice type is expected to send convince the other side that the sender is nice, whereas signals smaller than the nice type is expected to send convince the other side that the sender is mean. This assumption is substantively motivated. In general, large cooperative gestures are typically more reassuring than small cooperative gestures. Because they pose a greater risk for the sender, they seem to indicate a greater desire to cooperate. For instance, if you expect your opponent, if they are sincere about peace, to withdraw one division of troops from a disputed border region and instead they withdraw three, you might be even more willing to believe they are sincere in their desire for peace. A natural way to reflect this is to posit that if a signal is sent that is even larger than the equilibrium signal that the nice type is expected to send, then player 2 should become convinced that player 1 is nice for sure. Conversely, low signals are less reassuring, if they are reassuring at all. If we expect the mean type to send no signal at all, low signals can even be seen as feints by mean players to deceive others into cooperating. If the other side, rather than leaving its troops fully deployed as we expect if they are mean or withdrawing a full division as we expect if

99. See, for example, Ganguly and Greenwood 1996; Lindley 1998; and Adler and Barnett 1998.
100. And continuing on to the present, Van Evera 1998; and Glaser and Kaufmann 1998.
101. Ostrom 1998.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

they are nice, instead withdraws a few small units, such a signal would be viewed with suspicion. A simple way to capture this "small signals are not reassuring" feeling is to posit that if the signal sent is lower than the equilibrium signal for the nice type, player 2 remains convinced that player 1 is mean for sure. Thus the signal that the nice type is expected to send forms a boundary between two regions. Greater signals convince the other side that you are nice, lesser signals convince them that you are mean.

In the formal supplement to the article I also consider an alternative assumption in which posterior beliefs off the equilibrium path simply remain the same as prior beliefs. Thus if player 1 sends a signal that has no likelihood of being sent by either the nice or the mean type, player 2's beliefs will remain what they were before. All of the results mentioned in the article hold regardless of which of these assumptions is made.

In the signaling equilibrium, the mean type of player 1 will not wish to deviate to the equilibrium signal for the nice type if

$$\alpha^*(p_2 R_{1M} + (1 - p_2)(-S_{1M})) + (1 - \alpha^*)T_{1M} < 0$$

which produces a lower bound for the nice type's signal:

$$\alpha^* > \frac{T_{1M}}{T_{1M} - p_2 R_{1M} + (1 - p_2)S_{1M}}$$

The nice type will not wish to deviate to the mean type's signal (zero) if

$$\alpha^*(p_2 R_{1N} + (1 - p_2)(-S_{1N})) + (1 - \alpha^*)R_{1N} > 0$$

which produces an upper bound on the nice type's signal:

$$\alpha^* < \frac{R_{1N}}{(1 - p_2)(R_{1N} + S_{1N})}$$

Note that both of these bounds are increasing in p_2; that is, as the parties grow more trusting, the size of the signal, or importance of the first round, increases; as they grow more fearful, the first round must diminish in importance. The upper bound exceeds the lower bound, making an equilibrium signal possible, if $p_2 > p_2$ as defined in the text.

References

Adler, Emanuel, and Michael Barnett, eds. 1998. *Security Communities*. Cambridge: Cambridge University Press.

Allan, Pierre, and Kjell Goldmann, eds. 1992. *The End of the Cold War*. Dordrecht: Martinus Nijhoff Publishers.

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

Beschloss, Michael R., and Strobe Talbott. 1993. *At the Highest Levels: The Inside Story of the End of the Cold War*. Boston: Little, Brown.

Bitzinger, Richard A. 1994. Gorbachev and GRIT, 1985–1989: Did Arms Control Succeed Because of Unilateral Actions or in Spite of Them? *Contemporary Security Policy* 15 (1):68–79.

Blacker, Coit D. 1993. *Hostage to Revolution: Gorbachev and Soviet Security Policy, 1985–1991*. New York: Council on Foreign Relations Press.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

Braithwaite, Valerie, and Margaret Levi, eds. 1998. *Trust and Governance.* New York: Russel Sage Foundation.

Chafetz, Glenn R. 1993. *Gorbachev, Reform, and the Brezhnev Doctrine: Soviet Policy Toward Eastern Europe, 1985–1990.* Westport, Conn.: Praeger.

Checkel, Jeff. 1993. Ideas, Institutions, and the Gorbachev Foreign Policy Revolution. *World Politics* 45 (2): 271–300.

Cohen, Eliot A. 1988. Toward Better Net Assessment: Rethinking the European Conventional Balance. *International Security* 13 (1):50–89.

Coleman, James S. 1990. *Foundations of Social Theory.* Cambridge, Mass.: Belknap Press.

Collins, Alan R. 1998. GRIT, Gorbachev, and the End of the Cold War. *Review of International Studies* 24 (2):201–19.

de Figueiredo Jr., Rui J. P., and Barry R. Weingast. 1999. The Rationality of Fear: Political Opportunism and Ethnic Conflict. In *Civil Wars, Insecurity, and Intervention,* edited by Barbara F. Walter and Jack Snyder, 261–302. New York: Columbia University Press.

Downs, George W., and David M. Rocke. 1990. *Tacit Bargaining, Arms Races, and Arms Control.* Ann Arbor: University of Michigan Press.

Etzioni, Amitai. 1962. *The Hard Way to Peace: A New Strategy.* New York: Crowell-Collier Press.

Farrell, Joseph, and Matthew Rabin. 1996. Cheap Talk. *Journal of Economic Perspectives* 10 (3):103–18.

Fearon, James D. 1993. Threats to Use Force: Costly Signals and Bargaining in International Crises. Ph.D. diss., University of California, Berkeley.

———. 1995. Rationalist Explanations for War. *International Organization* 49 (3):379–414.

Fearon, James D., and David D. Laitin. 1996. Explaining Interethnic Cooperation. *American Political Science Review* 90 (4):715–35.

Fukuyama, Francis. 1995. *Trust: The Social Virtues and the Creation of Prosperity.* New York: Free Press.

Gambetta, Diego, ed. 1988. *Trust: Making and Breaking Cooperative Relations.* New York: Basil Blackwell.

Ganguly, Sumit, and Ted Greenwood, eds. 1996. *Mending Fences: Confidence and Security Building Measures in South Asia.* Boulder, Colo.: Westview Press.

Garthoff, Raymond L. 1990. *Deterrence and the Revolution in Soviet Military Doctrine.* Washington, D.C.: Brookings Institution.

———. 1994. *The Great Transition: American-Soviet Relations and the End of the Cold War.* Washington, D.C.: Brookings Institution.

Glaser, Charles L. 1992. Political Consequences of Military Strategy: Expanding and Refining the Spiral and Deterrence Models. *World Politics* 44 (4):497–538.

———. 1994. Realists as Optimists: Cooperation as Self-Help. *International Security* 19 (3):50–90.

———. 1997. The Security Dilemma Revisited. *World Politics* 50 (1):171–202.

Glaser, Charles L., and Chaim Kaufmann. 1998. What Is the Offense-Defense Balance and Can We Measure It? *International Security* 22 (4):44–82.

Goldstein, Joshua S., and John R. Freeman. 1990. *Three-Way Street: Strategic Reciprocity in World Politics.* Chicago: University of Chicago Press.

Güth, Werner, and Hartmut Kliemt. 1994. Competition or Co-Operation: On the Evolutionary Economics of Trust, Exploitation, and Moral Attitudes. *Metroeconomica* 45 (2):155–87.

———. 1998. The Indirect Evolutionary Approach: Bridging the Gap Between Rationality and Adaptation. *Rationality and Society* 10 (3):377–99.

Hogan, Michael J., ed. 1992. *The End of the Cold War: Its Meaning and Implications.* Cambridge: Cambridge University Press.

Hollis, Martin. 1998. *Trust Within Reason.* Cambridge: Cambridge University Press.

Jervis, Robert. 1976. *Perception and Misperception in International Politics.* Princeton, N.J.: Princeton University Press.

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

Kaiser, Robert G. 1992. *Why Gorbachev Happened: His Triumphs, His Failure, and His Fall.* New York: Simon and Schuster.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

Kaufman, Stuart J. 1996. Spiraling to Ethnic War: Elites, Masses, and Moscow in Moldova's Civil War. *International Security* 21 (2):108–38.

Kelman, Herbert C. 1985. Overcoming the Psychological Barrier: An Analysis of the Egyptian-Israeli Peace Process. *Negotiation Journal* 1 (3):213–34.

Kirkpatrick, Jeanne J. 1990. *The Withering Away of the Totalitarian State . . . and Other Surprises*. Washington, D.C.: The American Enterprise Institute Press.

Kydd, Andrew. 1997. Sheep in Sheep's Clothing: Why Security Seekers Do Not Fight Each Other. *Security Studies* 7 (1):114–55.

Lahno, Bernd. 1995a. Trust and Strategic Rationality. *Rationality and Society* 7 (4):442–64.

———. 1995b. Trust, Reputation, and Exit in Exchange Relationships. *Journal of Conflict Resolution* 39 (3):495–510.

Lake, David A., and Donald Rothchild. 1996. Containing Fear: The Origins and Management of Ethnic Conflict. *International Security* 21 (2):41–75.

Lake, David A., and Donald Rothchild, eds. 1998. *The International Spread of Ethnic Conflict: Fear, Diffusion, and Escalation*. Princeton, N.J.: Princeton University Press.

Landa, Janet Tai. 1994. *Trust, Ethnicity, and Identity: Beyond the New Institutional Economics of Ethnic Trading Networks, Contract Law, and Gift Exchange*. Ann Arbor: University of Michigan Press.

Lane, Christel, and Reinhard Bachmann, eds. 1998. *Trust Within and Between Organizations: Conceptual Issues and Empirical Applications*. New York: Oxford University Press.

Larson, Deborah W. 1987. Crisis Prevention and the Austrian State Treaty. *International Organization* 41(1):27–60.

———. 1997. *Anatomy of Mistrust: U.S.–Soviet Relations During the Cold War*. Ithaca, N.Y.: Cornell University Press.

Lebow, Richard Ned, and Thomas Risse-Kappen, eds. 1995. *International Relations Theory and the End of the Cold War*. New York: Columbia University Press.

Lebow, Richard Ned, and Janice Gross Stein. 1994. *We All Lost the Cold War*. Princeton, N.J.: Princeton University Press.

Lindley, Daniel A. 1998. Transparency and the Effectiveness of Security Regimes: A Study of the Concert of Europe Crisis Management and United Nations Peacekeeping. Ph.D. diss., Massachusetts Institute of Technology.

Lindskold, Svenn. 1978. Trust Development, the GRIT Proposal, and the Effects of Conciliatory Acts on Conflict and Cooperation. *Psychological Bulletin* 85 (4):772–93.

Mearsheimer, John J. 1982. Why the Soviets Can't Win Quickly in Central Europe. *International Security* 7 (1):3–39.

———. 1994. The False Promise of International Institutions. *International Security* 19 (3):5–49.

Mendelson, Sarah E. 1993. Internal Battles and External Wars: Politics, Learning, and the Soviet Withdrawal from Afghanistan. *World Politics* 45 (3):327–60.

Morrow, James D. 1994. *Game Theory for Political Scientists*. Princeton, N.J.: Princeton University Press.

Oberdorfer, Don. 1991. *The Turn: From the Cold War to a New Era: United States and the Soviet Union, 1983–1990*. New York: Poseidon Press.

Osgood, Charles Egerton. 1962. *An Alternative to War or Surrender*. Urbana: University of Illinois Press.

Ostrom, Elinor. 1998. A Behavioral Approach to the Rational Choice Theory of Collective Action. *American Political Science Review* 92 (1):1–22.

Peffley, Mark, and Jon Hurwitz. 1992. International Events and Foreign Policy Beliefs: Public Response to Changing Soviet–U.S. Relations. *American Journal of Political Science* 36 (2):431–61.

Plous, S. 1985. Perceptual Illusions and Military Realities. *Journal of Conflict Resolution* 29 (3):363–89.

———. 1987. Perceptual Illusions and Military Realities: Results from a Computer-simulated Arms Race. *Journal of Conflict Resolution* 31 (1):5–33.

———. 1988. Modeling the Nuclear Arms Race as a Perceptual Dilemma. *Philosophy and Public Affairs* 17 (1):44–53.

———. 1993. The Nuclear Arms Race: Prisoner's Dilemma or Perceptual Dilemma? *Journal of Peace Research* 30 (2):163–79.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press

Posen, Barry R. 1984a. Measuring the European Conventional Balance: Coping with Complexity in Threat Assessment. *International Security* 9 (3):47–88.
———. 1984b. *The Sources of Military Doctrine: France, Britain, and Germany Between the World Wars*. Ithaca, N.Y.: Cornell University Press.
———. 1993. The Security Dilemma and Ethnic Conflict. *Survival* 35 (1):27–47.
Reagan, Ronald. 1990. *An American Life*. New York: Simon and Schuster.
Richman, Alvin. 1991. Changing American Attitudes Toward the Soviet Union. *Public Opinion Quarterly* 55 (1):135–48.
Risse-Kappen, Thomas. 1991. Did "Peace Through Strength" End the Cold War?: Lessons from INF. *International Security* 16 (1):162–88.
———. 1994. Ideas Do Not Float Freely: Transnational Coalitions, Domestic Structures, and the End of the Cold War. *International Organization* 48 (2):185–214.
Sagan, Scott D. 1994. The Perils of Proliferation: Organization Theory, Deterrence Theory, and the Spread of Nuclear Weapons. *International Security* 18 (4):66–107.
Schultz, George. 1993. *Turmoil and Triumph: My Years as Secretary of State*. New York: Charles Scribner's Sons.
Schweller, Randall L. 1996. Neorealism's Status Quo Bias: What Security Dilemma? *Security Studies* 5 (3):90–121.
Sigelman, Lee. 1990. Disarming the Opposition: The President, the Public, and the INF Treaty. *Public Opinion Quarterly* 54 (1):37–47.
Snyder, Jack. 1984. *The Ideology of the Offensive: Military Decision Making and the Disasters of 1914*. Ithaca, N.Y.: Cornell University Press.
Spence, Michael. 1973. Job Market Signaling. *Quarterly Journal of Economics* 87 (3):355–74.
Spirtas, Michael. 1996. A House Divided: Tragedy and Evil in Realist Theory. *Security Studies* 5 (3):385–423.
Stein, Janice Gross. 1991. Reassurance in International Conflict Management. *Political Science Quarterly* 106 (3):431–51.
———. 1994. Political Learning by Doing: Gorbachev as Uncommitted Thinker and Motivated Learner. *International Organization* 48 (2):155–83.
Thom, Françoise. 1989. *The Gorbachev Phenomenon*. London: Pinter Publishers.
Van Evera, Stephen. 1994. Hypotheses on Nationalism and War. *International Security* 18 (4):5–39.
———. 1998. Offense, Defense, and the Causes of War. *International Security* 22 (4):5–43.
Walt, Stephen M. 1996. *Revolution and War*. Ithaca, N.Y.: Cornell University Press.
Walter, Barbara F. 1997. The Critical Barrier to Civil War Settlement. *International Organization* 51 (3):335–64.
Ward, Hugh. 1989. Testing the Waters: Taking Risks to Gain Reassurance in Public Goods Games. *Journal of Conflict Resolution* 33 (2):274–308.
Watson, Joel. 1999. Starting Small and Renegotiation. *Journal of Economic Theory* 85 (1):52–90.
Weingast, Barry R. 1998. Constructing Trust: The Political and Economic Roots of Ethnic and Regional Conflict. In *Institutions and Social Order*, edited by Karol Soltan, Eric Uslaner, and Virginia Haufler, 163–200. Ann Arbor: University of Michigan Press.
Wohlforth, William C. 1995. Realism and the End of the Cold War. *International Security* 19 (3):91–129.

https://doi.org/10.1162/002081800551190 Published online by Cambridge University Press