Signaling Under the Security Dilemma: An Experimental Analysis

Brandon Yoder* & Kyle Haynes†

September 30, 2020

Abstract

One of the most intractable debates in IR revolves around the severity and frequency of the security dilemma. Offensive realists argue that states are compelled to make worst-case assumptions about each other's intentions, which yields inexorable competition and conflict even between mutually-benign actors. Yet others have argued that rational benign states should always be able to find cooperative signals that are costly enough to be credible, but not too costly to risk sending. This should alleviate the security dilemma and facilitate cooperation, even under high initial distrust. However, there is little empirical work on interstate reassurance and the conditions under which mutually-benign actors can build trust. We advance this debate using laboratory experiments to test Andrew Kydd's canonical model of the security dilemma. We find strong support for the directional effects of the hypothesized signaling mechanisms. However, the frequency of cooperation is significantly lower than the model predicts, and the feasibility of reassurance is highly sensitive to the degree of prior trust. This implies that although reassurance can mitigate the security dilemma, offensive realism may still capture important psychological mechanisms that impede interstate cooperation.

Forthcoming at the Journal of Conflict Resolution

*The authors share equal credit for this article. This is one of several joint projects by the authors, and the ordering of names follows a rule of alternation. This research was funded by the Centre for Asia and Globalization in the Lee Kuan Yew School of Public Policy at the National University of Singapore. The authors would like to thank Mark Fey, Charles Holt, Josh Kertzer, Andrew Kydd, Julia Yujung Lee, Yon Lupu, Kai Quek, Mike Schreck, Rick Wilson, and two anonymous referees for their helpful feedback on earlier versions of this paper. The authors would also like to thank Tim Cason for allowing us to use the Vernon Smith Experimental Economics Laboratory to carry out our experimental trials.

*School of Politics and International Relations, Australian National University. brandon.k.yoder@gmail.com

†Department of Political Science, Purdue University. kylehaynes@purdue.edu

The security dilemma, a phenomenon in which states with mutually-benign intentions find themselves in conflict due to misplaced distrust, is a foundational concept in international relations (IR) scholarship. As a result, the prevalence and severity of security dilemmas have been the subject of longstanding debate. Whereas offensive realists claim that security dilemmas are a pervasive and unavoidable consequence of international anarchy, more optimistic scholars argue that the security dilemma can often be overcome.

This debate hinges on disagreement regarding the degree to which states can ameliorate uncertainty about each other's intentions. Optimists have suggested numerous signaling mechanisms that would allow states with benign intentions to identify each other with relatively high confidence. Yet offensive realists have categorically rejected these mechanisms. They maintain that states can virtually never form confident beliefs about others' intentions, and must instead act in accordance with worst-case assumptions to secure themselves against potential exploitation.

Surprisingly, despite the centrality of this debate in the IR literature and mountains of theoretical work attempting to resolve it, the mechanisms posited by signaling optimists have been subjected to few systematic empirical tests.[1] Addressing this critical gap, we employ a laboratory experiment to evaluate Andrew Kydd's (2005) model of reassurance under the security dilemma. Kydd shows formally that even with low initial trust, rational benign states should *always* be able to credibly signal their intentions through a process of iterated reassurance, building trust under stakes that are low – but not too low – before moving on to cooperation on issues of greater importance. Although Kydd's findings are devastating to the logic of offensive realism, it is unclear if real-world decisionmakers, who are subject to myriad cognitive limitations and psychological biases, behave as his model predicts or according to offensive realist expectations.

Our experimental findings are mixed. On one hand, we find strong evidence that Kydd's signaling mechanism does indeed operate. As his model predicts, benign players can best identify one another and establish mutual cooperation under intermediate stakes. However, the magnitude of those effects is far smaller than expected. Even under the very best conditions, cooperative signals still leave actors with considerable uncertainty, and inefficient competition among benign dyads remains common. Furthermore, contrary to Kydd's claim that credible reassurance is feasible irrespective of actors' initial trust, we find that lower initial trust substantially inhibits cooperation among mutually-benign players. Thus, our data suggest that although Kydd's signaling mechanism does *mitigate* the security dilemma, it does not eliminate it, as his theoretical model suggests it should.

[1] For a rare exception, see Quek (2017*b*).

Instead, the behavior and beliefs of our experimental subjects appear in some ways consistent with the pessimistic claims of offensive realism.

1 Signaling and the Security Dilemma

In a security dilemma, states with benign intentions compete for power to increase their own security, but in doing so reduce the security of others, unleashing escalating spirals of military and economic competition (Jervis 1978; Glaser 1997; Montgomery 2006; Tang 2009). This dynamic is predicated upon uncertainty. If states were completely informed of each other's intentions, benign actors could engage in mutually-beneficial cooperation without risk of being exploited, and competition for relative gains would occur only between hostile states with incompatible goals (Schweller 1996). But under uncertainty, benign states might be too fearful of others' intentions to risk taking cooperative actions that a hostile counterpart could exploit. This can engender costly competition even between benign actors with no true conflict of interests.

How pervasive and consequential is uncertainty about states' intentions? This question has been intensely debated. Offensive realists argue that uncertainty about intentions is intractable and consistently high (Layne 1993; Mearsheimer 2001; Rosato 2015). They claim that this uncertainty, combined with the self-help nature of the anarchic international system, compels rational states to make worst-case assumptions about each other's intentions. In effect, states must behave *as if* they were certain of others' hostility to protect themselves against any possibility of exploitation, no matter how unlikely (Copeland 2000). With all states operating under this assumption, competition for power is pervasive, making war an inevitable feature of international politics. Cooperation, when it occurs, is instrumental and transient, never based on genuine trust, convergent identities, or compatible preferences.

Scholars from other theoretical perspectives – defensive realists, liberals and constructivists – have drawn very different conclusions about the severity of the security dilemma. These scholars – hereafter termed "optimists" – argue that under many conditions it may be less risky for a highly uncertain state to gamble on cooperation and hope that it will be reciprocated than to take non-cooperative actions that ensure a hostility spiral with a counterpart that might in fact be benign (Glaser 1994, 1997, 2010; Kydd 1997, 2005; Mitzen 2006; Wendt 1999).

Furthermore, optimists have identified several sources of information that states have about each other's intentions. Most prominently, states can credibly signal benign

intentions by taking actions that would be prohibitively costly for hostile types.[2] For example, a state can signal benign intentions by reducing military investments, forgoing low-cost opportunities for expansion, or investing in primarily defensive technologies (Jervis 1978; Kydd 1997; Glaser 1994, 1997, 2010). These actions are highly costly to a hostile, expansionist state because they limit its capacity for offensive military action, but less so to a benign state that would not benefit from expansion. Generalizing beyond the military realm, states can signal benign intentions by joining institutions that constrain their behavior (Ikenberry 2001; Martin 2017), bearing costs to support existing international regimes (Johnston 2003), or publicly espousing a benign ideology to domestic audiences that would punish deviations from it (Kydd 1997; Weiss 2013). From these sources of information, optimists conclude that states can often develop high degrees of confidence about each other's intentions.

Offensive realists have critiqued these signaling mechanisms on the grounds that the security dilemma itself precludes credible reassurance. By exercising restraint and not competing for power, states not only forgo opportunities to exploit others, but also increase their own vulnerability to exploitation. Cooperative behavior is therefore costly to benign states as well as to hostile ones. Offensive realists claim that this engenders a "Goldilocks problem," wherein conditions that support credible signals rarely, if ever, obtain in reality (Montgomery 2006; Rosato 2015). They argue that when actors are insecure or distrustful, cooperative signals would be prohibitively costly even to benign types. Conversely, when exploitation is especially difficult or unlikely, hostile states are likely to behave cooperatively as well, rendering cooperative signals non-credible. Doing so allows them to conceal their malign intentions until more auspicious opportunities for expansion emerge, and dupe benign states into adopting cooperative behaviors that will make them more exploitable. Thus, for offensive realists, cooperative signals should either be noncredible or nonexistent.

To resolve this debate, several studies have developed formal models examining the credibility of cooperative signals under various conditions. The most prominent of these are Andrew Kydd's (2000; 2005, ch 7) models of the security dilemma as a two-stage interaction in which the relative value of the issues at stake in each stage can vary. Kydd shows that rational states can credibly signal benign intentions through a process of iterated reassurance. No matter how distrustful and insecure the actors initially are, there is always a range of first-round stakes valuable enough that hostile types are unwilling to forgo the chance at immediate exploitation, but not so valuable that benign types are

[2]See Fearon (1997) on the basic logic of costly signaling in international relations.

unwilling to risk cooperation.[3] In other words, offensive realists' Goldilocks problem is readily overcome, such that benign types can always identify each other and avoid costly competition. Kydd (2005, 201) states this conclusion emphatically: "cooperation is possible between [benign states] *no matter how mistrustful they are to begin with*...if they are genuine security seekers, [states] can find an appropriate set of costly signals that will enable them to reassure each other and cooperate completely over time" (emphasis in original). Recent work has extended Kydd's insights to conditions of shifting power, which offensive realists have rightly pointed out pose additional barriers to credible signals (Haynes 2019; Yoder 2019*a*,*b*; Haynes and Yoder 2020). This research demonstrates that reassurance remains theoretically feasible even under the most difficult conditions for credible signaling, as identified by offensive realists.

Kydd's argument follows from the premise that both benign and hostile types value their security, and thus lose equally from their cooperation being exploited. The risks of cooperation are the same for each type. But only hostile types gain from exploiting others, and so they necessarily bear opportunity costs for cooperation that benign types do not. Thus, even if benign types are fearful, they are always *more* willing to cooperate than hostile types are. This implies that there must be some range of stakes under which benign types are willing to risk cooperation but hostile types are not, and it is in this range that cooperative signals are credible. The next section characterizes Kydd's model and its results, which are the basis of our experimental protocol and hypotheses.

2 The Security Dilemma Game

Kydd (2000; 2005, ch 7) models the uncertainty inherent to the security dilemma as incomplete information about whether the players are in a prisoners' dilemma, a stag hunt, or a hybrid of the two. Benign types prefer to reciprocate their counterpart's cooperation rather than defecting unilaterally. This corresponds to the ordinal preferences of a player in the "stag hunt" game (Table 1a). In contrast, hostile types prefer unilateral defection to mutual cooperation – they prefer to exploit the cooperation of others, and therefore have a dominant strategy to defect. This corresponds to a "prisoners' dilemma" preference ordering (Table 1b).

[3]Kydd shows that, theoretically, higher initial distrust requires simply requires smaller initial stakes to support effective trust building.

Table 1: Type Combinations in the SD Game

(a) Stag Hunt

		Benign Player 2 C	**Benign Player 2** D
Benign Player 1	C	2, 2	-1, 1
	D	1, -1	0, 0

(b) Prisoners' Dilemma

		Hostile Player 2 C	**Hostile Player 2** D
Hostile Player 1	C	1, 1	-1, 2
	D	2, -1	0, 0

(c) Mixed Stag Hunt/Prisoners' Dilemma

		Hostile Player 2 C	**Hostile Player 2** D
Benign Player 1	C	2, 1	-1, 2
	D	1, -1	0, 0

Kydd's model – which we call the SD game – has two-sided incomplete information such that each actor knows its own type, but does not know the other's. Thus, from the perspective of a benign type, the game may be a pure stag hunt if the other player is also benign, or it could be a mixed stag hunt/prisoner's dilemma (Table 1c) if the other player is hostile. In a stag hunt, two benign states can readily coordinate on mutual cooperation in equilibrium. However, with a hostile counterpart, the only equilibrium is mutual defection. Whether the benign player cooperates or defects therefore depends on its probabilistic belief that its counterpart is benign, which Kydd defines as "trust." From the perspective of a benign type, the level of trust is the probability that it is playing the game in Table 1a, rather than Table 1c.

The SD game has two rounds, with the first-round payoffs proportional to the second-round payoffs, but multiplied by a weighting factor, α. For convenience, we assume symmetrical levels of prior trust, denoted t_0.[4] In the first round, both actors simultaneously choose whether to cooperate or defect, then observe the other player's move and update their beliefs to form a posterior level of trust, t'. The game ends with the second-round moves, with both actors again choosing whether to cooperate or defect.

[4]Kydd (2005 demonstrates that relaxing this assumption has no effect on the model's substantive results.

2.1 Equilibria

The SD game yields four equilibria.[5] When initial trust is relatively high and first-round stakes relatively low, a cooperative pooling equilibrium (PE) occurs in which both types cooperate in the first round but only benign types cooperate in the second. In this case, benign types have sufficiently optimistic priors that they cooperate in the second round even in response to completely uninformative cooperative signals. This gives hostile types an incentive to misrepresent in the first round in order to dupe a benign counterpart into cooperating under higher stakes in the second round.

The three other equilibria occur when initial trust is relatively low. Combined with relatively high first-round stakes, a competitive PE occurs in which both types defect in both rounds. This occurs straightforwardly because the first-round stakes are too high, given the degree of initial trust, for benign types to risk cooperating and incurring the sucker's payoff. Thus, the competitive PE captures the pessimistic predictions of offensive realism.

Conversely, when first-round stakes and initial trust are both sufficiently low, a mixed-strategy equilibrium (MSE) occurs. Here, benign types are willing to risk first-round cooperation, but so are hostile types if it means inducing benign types to cooperate under higher second-round stakes. Thus, hostile types misrepresent in the first round by behaving cooperatively with some probability. These cooperative signals have limited credibility, but they still induce benign types to probabilistically cooperate in the second round. Thus, hostile types have some chance of duping a benign type into incurring the sucker's payoff under high second-round stakes.[6]

Finally, a separating equilibrium (SE) in which benign types cooperate and hostile types defect in the first round occurs under low initial trust and intermediate first-round stakes. Here, cooperation is a completely credible signal that the sender's intentions are benign. Moreover, a range of α that supports the SE exists under all levels of prior trust. Benign types – which inherently prefer mutual cooperation and thus do not face

[5] This discussion brackets a fifth equilibrium, in which all players defect in both rounds and which is supported across the entire parameter space. Like Kydd, we ignore this equilibrium in our analysis, on the grounds that it is unrealistic to think that sufficiently trusting stag hunt actors would choose to coordinate on defection when the more profitable coordination on cooperation is possible. See Kydd (2005, 37, 191). Moreover, even though cheap-talk communication is not possible in our experiment, the first-round cooperative signal doubles as a coordinating device for the second round, ensuring that trusting stag hunt actors can achieve mutual second-round cooperation. This, in turn, also incentivizes first-round cooperation among sufficiently trusting benign actors.

[6] Pure-strategy pooling on cooperation is out of equilibrium because, given low initial trust, cooperative signals would not be sufficiently credible to induce second-round cooperation by benign types.

opportunity costs of forgoing unilateral defection – are necessarily more willing to cooperate than are hostile types. As such, there is always some range of stakes under which benign types are willing to risk cooperation but hostile types are not, no matter how low the initial level of trust. With lower levels of prior trust, the actors simply require smaller first-round stakes in order to achieve successful reassurance.

This separating equilibrium is the key to Kydd's theory of reassurance. It occurs only under intermediate values of α that balance the two obstacles to reassurance highlighted by offensive realists. The first-round interaction must not be so important that benign types are unwilling to cooperate for fear of being suckered, but must be important enough that cooperation serves as a costly signal of the sender's type. Thus, uniquely under intermediate stakes, cooperative signals should have high credibility and be sent with high frequency, yielding second-round cooperation among mutually-benign dyads.

2.2 Theoretical Implications

The SD game demonstrates that offensive realists' conclusions do not follow from their assumptions. Even with high initial distrust, credible signals of benign intentions should always be available if the stakes of the initial interaction are calibrated properly. As has been noted repeatedly elsewhere, rational actors – which offensive realists assume states to be – cannot ignore the information from these signals, and therefore cannot make worst-case assumptions about each other's intentions (Brooks 1997; Glaser 2010; Kydd 2005). As a rationalist theory, offensive realism is demonstrably incoherent.[7]

Nevertheless, offensive realism might survive as a *behavioral* theory that accurately describes the decisions of boundedly-rational real-world actors. Formal models of reassurance have established only the conditions under which cooperative signals are *objectively* credible, i.e., more likely to be sent by benign types than hostile ones. However, these models cannot establish the *subjective* credibility of cooperative signals, i.e., the degree to which policymakers actually update their beliefs in response. Indeed, the logical contradictions in offensive realism might accurately reflect cognitive biases that are systematically present in human decisionmaking, and which could cause real-world leaders to discount objectively credible cooperative signals.

As Stephen Brooks (1997) pointed out long ago, whether this is the case is ultimately

[7] Acharya and Ramsay (2013) present a formal model that they claim supports the logic of offensive realism. Importantly, however, their model assumes that only costless "cheap talk" communication is possible. It thus does not refute Kydd's logic of reassurance through costly signaling, and does not rescue offensive realism as a rationalist theory.

an empirical question. Yet there have been remarkably few empirical tests of rational security dilemma models.[8] Moreover, existing empirical work on signaling in other contexts yields considerable evidence that real-world actors are often far from rational in their belief formation (McDermott 2001; Stein 2013; Yarhi-Milo 2014; Quek 2017*a*; Hafner-Burton et al. 2017). As such, we cannot reject offensive realism as a behavioral theory by simply assuming that rationalist signaling mechanisms apply to real-world human decisionmaking.

The next section presents a laboratory experiment that reproduces the incentives of the SD game as closely as possible. We find considerable support for the directional effects of the model's signaling mechanism, but the magnitude of those effects is substantially smaller than would be expected even based on a conservative interpretation of the theoretical results. Reassurance is indeed most effective under intermediate stakes, when cooperative signals are neither too cheap to be meaningful nor too costly for benign actors to risk sending. Credible signals do therefore mitigate the security dilemma and promote cooperation among benign actors under the conditions the SD game predicts. However, we also find that the information subjects derive from these signals is far from complete. Even the most credible signals leave actors with considerable uncertainty, and costly competition remains quite common among benign dyads. Furthermore, contrary to Kydd's contention that prior beliefs are irrelevant, we find that in practice reassurance is significantly more difficult under high initial distrust. Thus, although Kydd's signaling mechanism is supported, it may not be a decisive cure for the security dilemma, and the behavior and beliefs of many human decisionmakers appear to be more consistent with the pessimistic expectations of offensive realism.

3 Research Design

Laboratory experiments have previously been used to evaluate signaling and credibility in bargaining contexts (McDermott, Cowden and Koopman 2002; Dickson 2009; Tingley 2011; Tingley and Walter 2011; Quek 2016, 2017*a*; Kertzer 2017).[9] We adopt a similar approach to examine the dynamics of reassurance signaling.[10] Our experimental design is what Morton (1999, 111) calls a "theory test," which tightly couples the experimental incentives to the structure of the theoretical model in order to determine whether human subjects behave and form beliefs as the model predicts under ideal conditions. This type of

[8] Kydd's (2005, ch 8) case study is, to our knowledge, the lone exception.

[9] For a review of this literature see Hyde (2015).

[10] Our study also differs from the few existing experimental papers on reassurance. Kertzer, Rathbun, and Rathbun (Forthcoming), present a survey experiment on psychological bias in reassurance and Quek (2017*b*) conducts a laboratory experiment on audience costs as reassurance signals.

experiment therefore scrutinizes the model's *behavioral* assumptions regarding subjects' cognitive processes and preference functions, rather than its *structural* assumptions about the actors' external environment (Davis and Holt 1993, 22). This design is particularly useful as a "first cut" empirical examination, as it constitutes a relatively easy test that evaluates the model on its own terms. As Davis and Holt (1993, 23) observe, empirical assessment of a theory:

> should ideally begin, not in the domain of the complex natural world, where numerous confounding events may impinge on variables of interest, but strictly on the domain of the theory, where all structural assumptions can be implemented...Of course, observation of the theory 'working' in the laboratory does not imply that it explains behavior in the natural world. But the failure of a theory under the 'best shot' circumstances of the laboratory suggest that the theory is not a good explainer of behavior. It is perhaps in this role of theory rejection that experimentation is most useful.

Hypotheses that pass this initial test can then be subjected to additional experimental and observational studies that interrogate them under more realistic conditions (Davis and Holt 1993, 31, Morton 1999, 111-116, 179-181). Our experiments should therefore be viewed as a baseline test of the SD game, with results derived under highly stylized conditions laying the groundwork for theoretical refinements and additional empirical analyses.

Our experiments were conducted at Vernon Smith Experimental Economics Laboratory at Purdue University using z-tree software.[11] We collected a total of 8,046 observations from 321 subjects over 18 experimental sessions, with each session consisting of between 24 and 43 iterations of the game. Participants were almost entirely undergraduate students. Below (section 5.1), we discuss the implications of this sample for generalizing our results to the behavior of policymakers.[12]

The experimental incentives corresponded directly to the SD game. In each iteration of the protocol, subjects were randomly paired and each assigned a type, with "benign" players having stag hunt incentives and "hostile" players having prisoners' dilemma incentives. Each pairing was randomly assigned a level of prior trust, t_0. This was either low, moderate, or high, manifested as a $\frac{1}{3}$, $\frac{1}{2}$, and $\frac{2}{3}$ chance, respectively, of each player

[11] Software and supporting documentation available at: `https://www.ztree.uzh.ch/en.html`. For a more detailed description of our experimental protocol, subject pool, recruitment procedures, etc., see Appendix B.

[12] The literature on experimental methods in political science has provided strong arguments that the use of undergraduate student samples does not necessarily threaten external validity (Morton and Williams 2008; Druckman and Kam 2011). We argue below that this is largely true of our study.

being benign. The players' types were assigned randomly and independently according to these probabilities, with players informed of their own assigned type, but only the probability that their counterpart was each type.

Each iteration of the game proceeded over two rounds. The second-round payoffs, presented in Table 2, were constant across all iterations of the game, but the first-round payoffs varied randomly across iterations.[13] First-round payoffs were proportional to second-round payoffs, but multiplied by weighting factor $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1, 2\}$. After learning the various parameter values and their own type, both players simultaneously chose whether to cooperate or defect in the first round. They were then informed of their counterpart's first-round action and their first-round payoff. Players then guessed their counterpart's type, receiving \$0.50 if correct and \$0 otherwise. They also assigned a probability that this guess was correct, ranging from 50% (complete uncertainty) to 100% (complete certainty). Finally, the players simultaneously chose their second-round strategies and then observed their final payoffs, ending that iteration of the game.

Table 2: Experimental Payoffs: 2nd Round

(a) Benign Types

		Own Strategy	
		Cooperate	Defect
Counterpart's Strategy	Cooperate	\$1	\$0.67
	Defect	\$0	\$0.33

(b) Hostile Types

		Own Strategy	
		Cooperate	Defect
Counterpart's Strategy	Cooperate	\$0.67	\$1
	Defect	\$0	\$0.33

4 Hypotheses and Results

4.1 Experimental Equilibria

As shown in Table 3, our experimental parameter values yield the full range of equilibrium predictions in the SD game. However, the parameters are clearly skewed in favor of cooperation. According to the model, benign players should cooperate in the first round under 17 of the 18 possible parameter combinations, and benign pairs should achieve 100% mutual second-round cooperation under 14 of the 18 possible conditions.

[13] We recognize the imperfect correspondence between monetary payments and players' subjective utilities. Nevertheless, they are easily quantifiable and readily understood by subjects. As such, our analyses below treat monetary payoffs as equivalent to the players' utilities.

Table 3: Theoretical Point Predictions

	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1$	$\alpha = 2$
$t_0 = \frac{1}{3}$	MSE	MSE	MSE	SE	SE	Competitive PE
$t_0 = \frac{1}{2}$	Cooperative PE	Cooperative PE	Cooperative PE	Cooperative PE	SE	SE
$t_0 = \frac{2}{3}$	Cooperative PE	Cooperative PE	Cooperative PE	Cooperative PE	Cooperative PE	SE

This bias toward cooperation was imposed deliberately, for two reasons. First, as discussed above, experimentation is most useful in the role of theory rejection (Davis and Holt 1993). This sort of first-cut "theory test" is most effective under best-case conditions for the theory to succeed, to determine if further empirical evaluation is warranted. Secondly, we conducted preliminary trials with only our low trust condition, $t_0 = \frac{1}{3}$, but (as discussed below) this yielded absolute rates of cooperation and successful reassurance far below theoretical predictions, with successful reassurance rates peaking at lower values of α than hypothesized. These results implied that in practice, for many players, the range of the separating equilibrium was skewed toward higher levels of prior trust than indicated by the model. In order to examine the effects of variation in initial trust, with each value of t_0 generating the full range of variation in the equilibria across α values in practice, we therefore used initial trust levels that were higher than the model implies would be appropriate. Table 2 of Appendix A, which presents predicted equilibrium behavior *given* the observed rates of cooperation from other experimental subjects, confirms that our parameter values did indeed achieve this aim of producing the full range of equilibrium behaviors for all three values of our prior trust variable.

4.2 Experimental Findings

This subsection presents the results of our experiments, focusing on the directional effects of our treatment variables – first-round stakes and prior trust – on our outcomes of interest, posterior beliefs and cooperation rates. The next section discusses additional implications of our results regarding the model's "point predictions," i.e., its precise equilibrium cooperation rates and degrees of trust. We recognize that it is unrealistic to expect our data to strictly support these point predictions (e.g., that cooperative signals are ever 100% credible or that mutually-benign dyads ever achieve 100% cooperation rates), and that evaluating our results relative to these point predictions therefore entails a large element of interpretation. However, it is essential that we evaluate the SD game's point predictions because, as detailed above, the substantive theoretical debate between signaling theorists and offensive realists hinges on the magnitude of our causal effects, not

just their direction.[14]

The SD game's top-line theoretical result is that cooperative signals should have high credibility and be sent with high frequency under intermediate stakes, yielding second round cooperation among mutually-benign dyads. In contrast, cooperative signals are frequently sent by hostile types under low first-round stakes, reducing their credibility, and are too risky for benign types when first-round stakes are excessively high. Under high and low values of α, therefore, reassurance is impeded, and second-round cooperation rates among benign dyads should be low.

Hypothesis 1: The frequency of mutual second-round cooperation among benign dyads is an inverse-U shaped function of α.

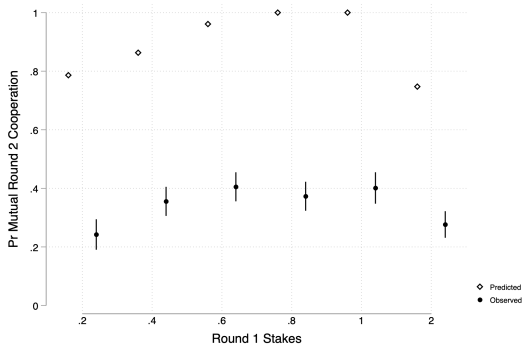
The dependent variable for $H1$ is a binary measure that takes a value of 1 if a pair of players both cooperated in round 2, and 0 if either or both players defected. We examine $H1$ looking only at observations in which both players were benign, as these are the only cases in which the SD game would predict mutual second-round cooperation. Figure 1 presents results with the observations pooled across all levels of prior trust. Figure 1a shows the mean likelihood of mutual cooperation, with 95% confidence intervals, for each value of α, with the SD game's equilibrium predictions plotted as hollow diamonds. Figure 1b then shows the predicted probability of mutual second-round cooperation based on an ordinary least squares (OLS) regression that included α, α^2, t_0, and a battery of demographic controls as right-hand side variables.[15] The plots in Figures 1c, 1d, and 1e mirror Figure 1a, depicting the mean likelihood of mutual cooperation alongside the equilibrium predictions, but with observations disaggregated by prior trust.

[14]This kind of subjective evaluation of point predictions is common in experimental research and often produces invaluable insights (Morton 1999, 165-169). Experiments on public goods provision and voting behavior, for example, show that the magnitude of the free-rider problem is far smaller than pure theory predicts (Ledyard 1995, Morton 1999, 179-181).

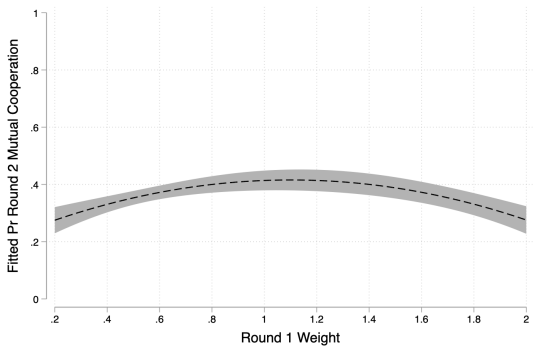
[15]The results were consistent across OLS and logit models, with and without demographic controls. Results for all model specifications are presented in Appendix B.

Figure 1: Probability of Round 2 Mutual Cooperation Among Benign Dyads

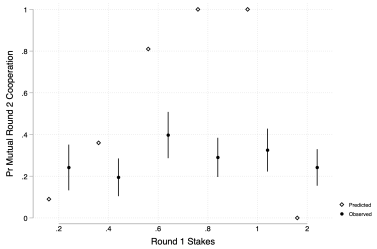
(a) Pooled Means



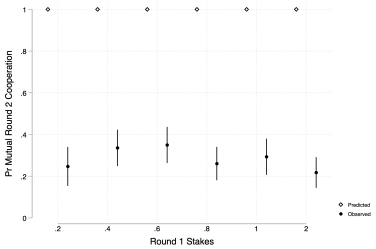
(b) Pooled Regression



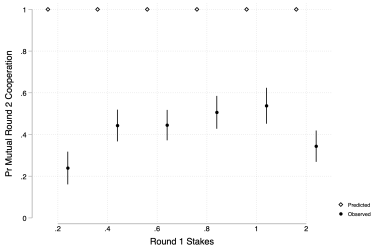
(c) Low Prior Trust



(d) Moderate Prior Trust



(e) High Prior Trust



The results strongly support $H1$. Figure 1 shows that in the pooled data, mutual cooperation between benign types is markedly less likely at the extrema of α, 0.2 and 2, compared to the four intermediate values. The results disaggregated across prior trust in Figures 1c, 1d, and 1e largely confirm this finding. The inverse-U shaped relationship is most clearly visible under high prior trust, but still apparent under moderate and low trust. This quadratic effect is statistically significant under both high and low prior trust (see Appendix B). Overall, mutual second-round cooperation between benign dyads is clearly most likely under intermediate first-round stakes.

This result is contrary to the SD game's prediction that mutually-benign dyads should achieve second-round cooperation across *all* α values under both moderate and high trust, as shown in Figures 1d and 1e. Instead, the model's core signaling mechanism still appeared to operate even under high levels of prior trust where the model predicts that hostile and benign types should pool on cooperation. This resulted in an observed frequency of mutual second-round cooperation far below the model's expectations. In effect, although (and in part because) absolute rates of cooperation were lower than

expected, the conditions supporting the separating equilibrium were evident across an even wider range of parameters than expected.

$H1$ is underpinned by the expected differences in the first-round behavior of benign and hostile types across α values. According to the SD game, benign types should cooperate under both low and intermediate values of α, where the MSE and SE occur, and defect only under high α, in the competitive PE. Conversely, hostile types should cooperate only under low α in the MSE (probabilistically) and CPE. Hostile types should defect under intermediate and high α. This implies that first-round cooperation rates for each type should be non-linear. Benign players' cooperation rates should be high under low and intermediate stakes and then curve downward under high stakes, whereas hostile cooperation rates should be relatively high under low stakes but decrease rapidly and remain quite low under both intermediate and high stakes. Thus, under low stakes cooperative signals should be sent with high frequency but lack credibility, and under high stakes they should be highly credible but sent infrequently. Only under intermediate stakes should cooperative signals be both credible and common.

Hypothesis 2a: First-round cooperation rates for benign types should monotonically decrease in α, gradually at low α and more rapidly at high α.

Hypothesis 2b: First-round cooperation rates for hostile types should monotonically decrease in α, rapidly at low α and more gradually at high α.

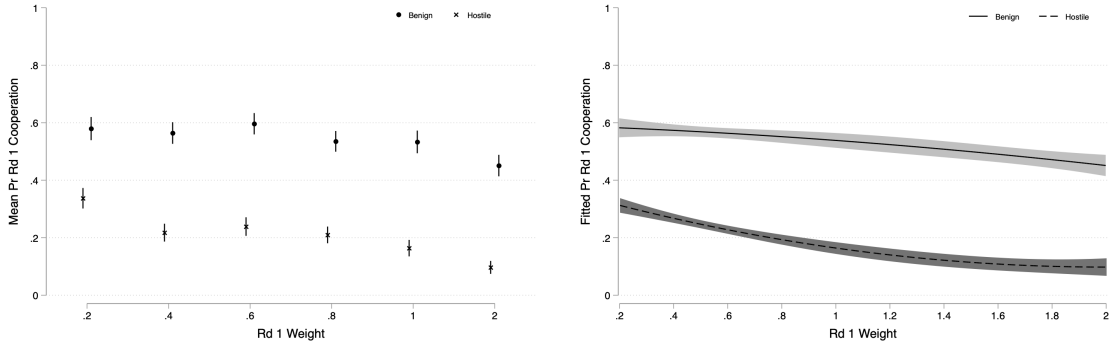
The dependent variable for $H2_a$ and $H2_b$ is a binary measure that equals 1 if a player cooperated in round 1, and 0 if the player defected in round 1. The results support both $H2_a$ and $H2_b$. Figure 2a shows the mean likelihood of first-round cooperation for both benign and hostile types across the range of α values.[16] Figure 2b shows the predicted probability of first-round cooperation for both benign and hostile types, based on OLS models with demographic controls.[17] These figures clearly show that first-round cooperation rates for both benign and hostile types decrease in α.

[16]For clarity of presentation, this figure does not show the model's equilibrium expectations. The theoretical predictions are included in the disaggregated figures (3 and 4) below.

[17]We regressed the dependent variable on α, α^2, t_0, and demographic controls. We included α^2 because $H2_a$ and $H2_b$ each predict a non-linear relationship between α and first-round cooperation.

Figure 2: Probability of Round 1 Cooperation

(a) Means **(b) Predicted Probabilities**



Moreover, there is some support for the predicted non-linear effects. Figure 2a shows that for benign types, cooperation rates are roughly constant when α equals 0.2, 0.4, and 0.6, then decrease significantly at higher values.[18] For hostile types, the probability of cooperation decreases sharply as α increases from 0.2 to 0.4, with the negative relationship less pronounced at higher α. Together, the panels in Figure 2 show that the absolute difference in cooperation rates between benign and hostile types is largest at intermediate α values, as the SD game predicts. Figures 3 and 4 show that these results are also supported for each level of prior trust, but that the observed frequency of cooperation is generally well below theoretical expectations.[19]

[18] In the fitted model (Figure 2b), however, a linear regression that drops α^2 actually provides better fit than the non-linear model according to Aikake's Information Criterion.

[19] A partial exception occurs for benign types under high prior trust (Figure 3c), where there appears to be an inverse-U relationship. Still, the results show the predicted sharp decline in cooperation as α increases from 1 to 2.

Figure 3: Disaggregated Probability of Round 1 Cooperation, Benign Types

(a) Low Prior Trust **(b) Moderate Prior Trust** **(c) High Prior Trust**

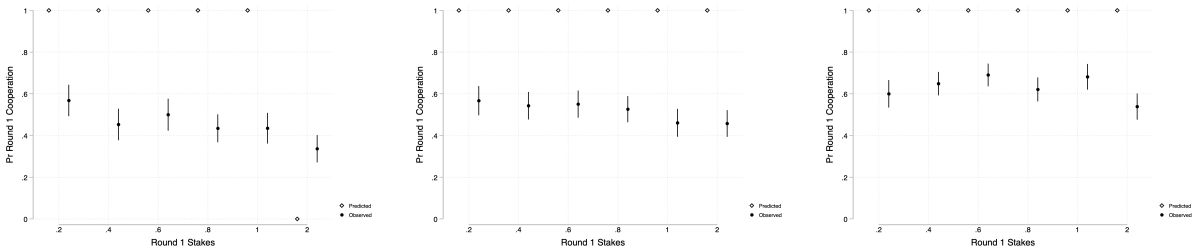
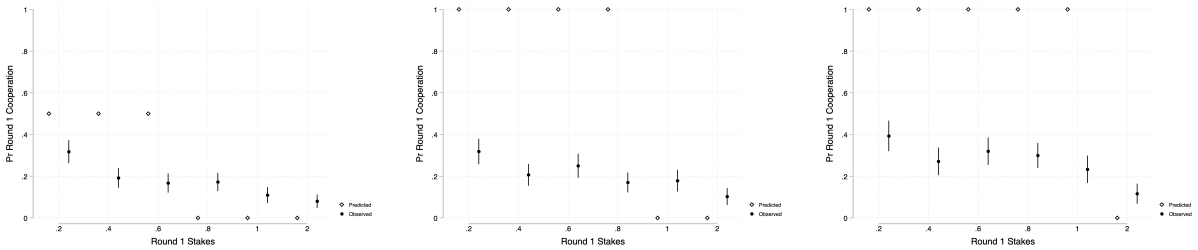


Figure 4: Disaggregated Probability of First Round Cooperation, Hostile Types

(a) Low Prior Trust **(b) Moderate Prior Trust** **(c) High Prior Trust**

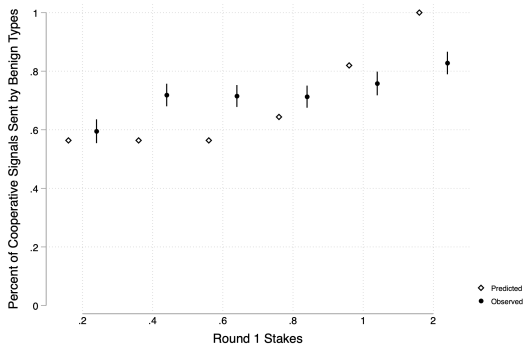


These cooperation rates allow us to establish the objective credibility of cooperative signals – i.e., the percentage of cooperative signals sent by benign types – which should increase with α. This expectation follows from the basic logic of costly signaling: as the risks of first-round cooperation increase, cooperative signals become more credible. This logic holds even in the competitive pooling equilibrium at high values of α, where the SD game predicts that cooperative signals should be too costly even for benign types to send. If cooperative signals *are* sent in contradiction of equilibrium predictions, the costs entailed should make them highly credible, since hostile types would be even less inclined to cooperate under such high stakes.

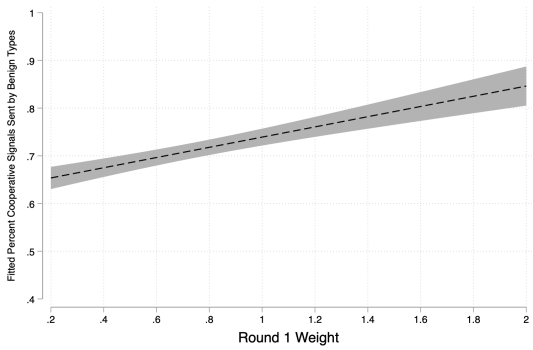
Hypothesis 3: The proportion of all first-round cooperative signals sent by benign types should increase in α.

Figure 5: Percentage of Cooperative Signals Sent by Benign Types

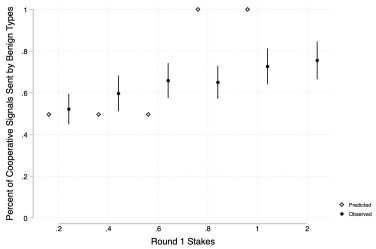
(a) Pooled Means



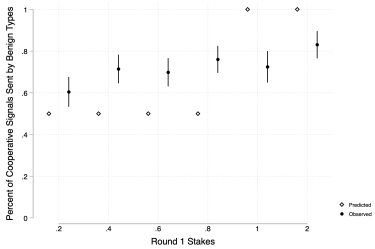
(b) Pooled Regression



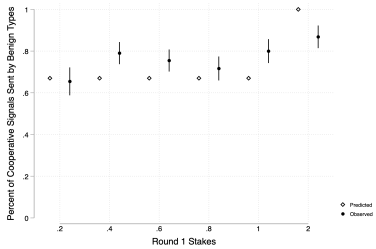
(c) Low Prior Trust



(d) Moderate Prior Trust



(e) High Prior Trust



The dependent variable for $H3$ captures the proportion of first-round cooperators that had benign preferences, and was generated for all observations in which a player cooperated in round 1. It is a binary measure that takes a value of 1 if the player is benign, and 0 if the player is hostile. The results strongly support $H3$. Figure 5a shows a clear positive relationship between objective credibility and α in the pooled data, with benign types sending less than 60% of cooperative signals when $\alpha = 0.2$, but nearly 85% when $\alpha = 2$. These values correspond well to the theoretical predictions. This finding is corroborated by the predicted probabilities presented in Figure 5b. Figures 5c, 5d, and 5e indicate that this effect again operates across all levels of prior trust.[20]

Finally, since the objective credibility of cooperative signals increases with α, we also expect players to more positively update their subjective beliefs in response to first-round cooperation under higher first-round stakes.

[20]There is some apparent non-monotonicity under high prior trust (Figure 5e), but the overall trend is clearly positive.

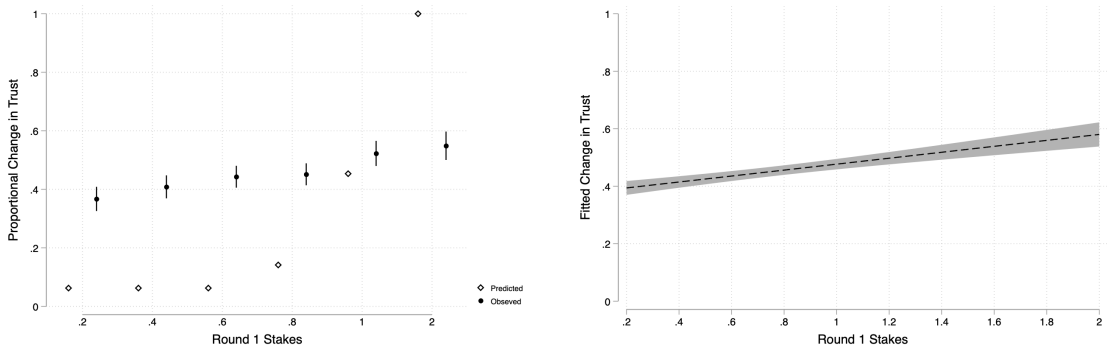
Hypothesis 4: The proportional increase in subjects' confidence that their counterpart is benign (i.e. $\frac{t'-t_0}{1-t_0}$) after observing a cooperative signal should increase in α.

The dependent variable for $H4$ captures the difference between posterior and prior trust $(t' - t_0)$ as a proportion of initial mistrust $(1 - t_0)$. If a player guessed their counterpart was benign, t' simply equals their stated confidence in this guess. If the player guessed their counterpart was hostile, t' equals one minus their stated confidence. We then subtract t_0 to capture change in trust, and divide this by $1 - t_0$ to account for the reduced size of the potential shift in beliefs that is possible as initial trust increases.

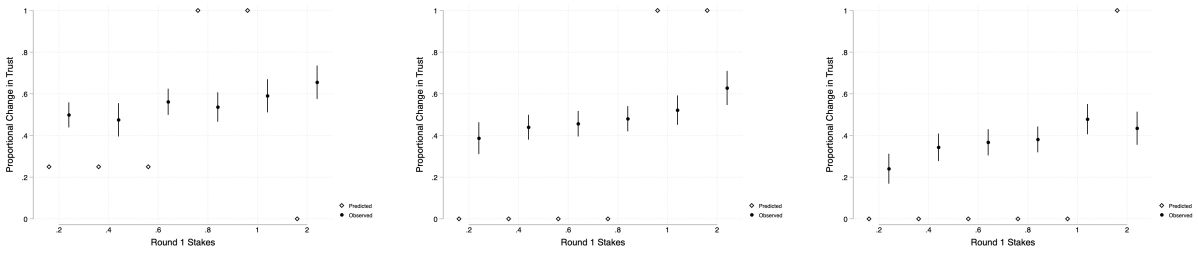
The results strongly support $H4$. Figure 6 shows that, in both the pooled and disaggregated data, the increase in reported trust following a cooperative signal increases with the first-round stakes. Thus, costlier cooperative signals are indeed more credible. This is consistent with the comparative statics of the SD game, which predicts that cooperation should be more informative under higher α values, yielding greater increases in trust.

Figure 6: Change in Trust Following Cooperative Signal

(a) Pooled Means **(b) Pooled Regression**



(c) Low Prior Trust **(d) Moderate Prior Trust** **(e) High Prior Trust**



This change in players' beliefs is also reflected in their second-round behavior. Appendix B shows that the probability of benign players cooperating in round 2 following mutual first-round cooperation significantly increased with α. Higher first-round stakes therefore induced players to be more responsive to cooperative signals, both in terms of their reported beliefs and their willingness to cooperate in round 2.

5 Discussion

These experimental results have important and nuanced implications for the longstanding debate surrounding the security dilemma. Most importantly, we find strong evidence that Kydd's hypothesized signaling mechanism does indeed operate in practice. Benign dyads' rates of mutual second-round cooperation are highest under intermediate first-round stakes ($H1$). Furthermore, this top-line result appears to be driven by the causal logic of the SD game. First, benign players were considerably less likely to send cooperative signals under high first-round stakes, which implies that they largely viewed the risks of cooperation under high stakes as prohibitive ($H2_a$). Secondly, however, the probability that hostile types cooperated in round 1 decreased even *more* quickly and precipitously as the stakes increased ($H2_b$). Thus, the objective credibility of cooperation as a signal of benign preferences increased with the first-round stakes ($H3$). Correspondingly, cooperative signals prompted participants to positively update their subjective beliefs more strongly under high stakes than under low stakes ($H4$).[21]

In combination, these results yielded the highest second-round cooperation rates among benign dyads under intermediate first-round stakes. Under high first-round stakes, cooperative signals were highly credible but benign types were generally unwilling to risk sending them. Conversely, under low first-round stakes, cooperative signals were sent relatively frequently by both benign and hostile types, and participants rightly viewed them as having low credibility. It was only under intermediate stakes where cooperative signals were both relatively credible *and* relatively frequent, yielding the highest rate of second-round cooperation among benign types. This is precisely the mechanism underpinning Kydd's theoretical results.

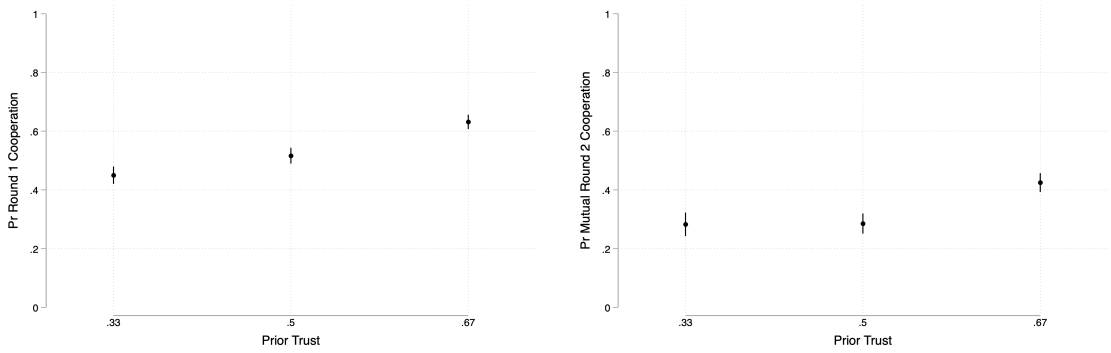
Although they support Kydd's general logic, our experimental results also raise questions about some of the model's more optimistic implications regarding the feasibility of cooperation under anarchy. First, the SD game implies – and Kydd is explicit – that

[21] As reported in Appendix B, this also affected second-round *actions*, as benign players were more likely to cooperate in round 2 after mutual round 1 cooperation when first-round stakes were higher.

initial mistrust should be inconsequential for the feasibility of reassurance. Our results do not support this claim. Figure 2 and Figure 7a clearly demonstrate that benign types were less likely to send cooperative signals as prior trust decreased. Furthermore, rates of second-round cooperation among benign dyads were significantly higher under high prior trust than under low and moderate prior trust (Figure 7b).[22] Indeed, under low initial trust, tragic conflict between benign players emerged 60-80% of the time even under intermediate stakes (α between 0.4 and 1), where the model predicts at least one value should have produced a separating equilibrium with highly credible cooperative signals and high second-round cooperation rates among benign types. Thus, initial distrust appears to hinder credible reassurance and cooperation among benign actors, as offensive realism suggests (Edelstein 2002; Montgomery 2006; Rosato 2015).[23]

Figure 7: Cooperation by Prior Trust

(a) Rd 1 Cooperation, Benign Players (b) Rd 2 Cooperation, Benign Dyads



Second, we observed rates of cooperation and posterior trust far below the equilibrium point predictions of the SD game. Of course, we recognize that it is unrealistic to expect experimental subjects to act with perfect rationality and adhere precisely to the model's equilibrium expectations. However, the magnitude of these causal effects, not just their direction, is substantively important for the theoretical debate we are engaging.

[22] These figures show the data pooled across all α values.

[23] It is important to note that although we reject Kydd's substantive conclusion that credible reassurance is equally possible regardless of prior trust, Kydd's model does find that the range of α values supporting the separating equilibrium shrinks as prior trust decreases. It is therefore unsurprising that, in practice, successful reassurance is less likely under lower prior trust, and this result does not constitute a refutation of the theoretical logic. Moreover, to reemphasize, the model's core comparative statics predictions were generally borne out across all levels of prior trust. Thus, we find directional support for Kydd's reassurance mechanism irrespective of prior trust, even though lower initial trust does indeed hinder credible reassurance and cooperation among benign actors.

Observed cooperation rates and levels of posterior trust that fall well below what even a generous interpretation of the SD game might reasonably expect constitute an important anomaly for the theory. The larger this disparity, the greater our confidence that offensive realism accurately describes the behavior of at least some real-world decisionmakers.[24]

Indeed, our experimental results show that subjective signal credibility and cooperation rates are far lower than expected, even under the intermediate stakes where reassurance is most effective. In the data pooled across all levels of prior trust (Figure 1a), the highest frequency of second-round cooperation among mutually-benign dyads was just over 40%, when $\alpha = 0.6$. Even under high initial trust, second-round cooperation among benign dyads peaked at only 54%, when $\alpha = 1$. This is remarkable, since the high prior trust parameter placed our thumbs firmly on the scale in favor of cooperation. Under our high trust priors, the model predicts the cooperative PE should obtain: benign players should have been willing to cooperate in round 2 even though both types cooperate in round 1, rendering cooperative signals completely uninformative. In short, despite the very best conditions for cooperation – intermediate stakes and high initial trust – tragic conflict among benign actors still occurred about half the time.

Moreover, these results appear to be driven by offensive realist worst-case assumptions among many participants. Under conditions where rational benign types should have cooperated in round 1, defection occurred 43.4% of the time.[25] Proposition 2 of Appendix A demonstrates that, given the actual conditional rates of cooperation among experimental subjects (i.e., accounting for the likelihood of out-of-equilibrium behavior), these players were leaving significant money on the table by not attempting cooperation under high trust.[26]

While risk aversion may have contributed to this behavior, our evidence suggests that many of the less cooperative subjects formed and acted upon irrationally pessimistic beliefs. Examining benign types under conditions where rational, risk-neutral players should have cooperated in the first round, we compared the posterior beliefs of subjects

[24]Qualitative feedback from experimental subjects indicated that they largely understood the game, and mitigates the possibility that this result might be partially due to a lack of experimental realism (see also fn 30, below). Moreover, if subjects either had difficultly comprehending the game's structure or disengaged as the experimental session progressed, we would expect to see changes in the results from early rounds to later rounds within each session. Yet Appendix B shows that subjects' behavior did not meaningfully change across rounds, suggesting that neither learning effects nor engagement fatigue seriously impacted our results.

[25]These conditions are those where either the separating, cooperative pooling, or mixed strategy equilibria were supported *given* the observed rates of cooperation under each combination of parameter values and game history. See Proposition 2 of Appendix A.

[26]For example, under high trust and $\alpha = 0.6$, first-round cooperation yielded an expected payoff of \$0.91, as opposed to \$0.65 for first-round defection.

who cooperated in the first round to those of subjects who defected.[27] Under these conditions, "irrational" defectors formed beliefs in response to cooperative signals that were markedly more pessimistic. On average, first-round defectors' posterior trust only increased 0.189 from their baseline level of prior trust after observing their counterpart cooperate, whereas first-round cooperators' trust increased by 0.254. These defectors' beliefs were also significantly less accurate, as they only guessed their counterpart's type correctly 65% of the time while cooperators guessed correctly 79% of the time. These differences are highly statistically significant. This difference suggests that a substantial portion of our subjects responded to even a relatively small degree of uncertainty as offensive realism implicitly predicts, by forming irrationally pessimistic beliefs and behaving accordingly.[28]

Importantly, not all subjects adopted worst-case assumptions. As Figures 2 and 3 show, many benign players did risk cooperation, and did so more frequently under the conditions the SD game predicts they should have. Thus, our findings do *not* support offensive realism as a rationalist structural theory wherein anarchy dictates uniformly pessimistic beliefs and competitive behavior. Rather, we find limited support for offensive realism as a *behavioral* theory, in which some actors behave irrationally in ways that diminish others' rational incentives for cooperation. As constructivists have long hypothesized (Wendt 1999) and as Kertzer and McGraw (2012) have demonstrated experimentally, worst-case thinking is not a rational calculation, but rather a dispositional trait that characterizes some actors but not others. Therefore, the microfoundations of offensive realist behavior are not found at the systemic level of analysis, but rather at the domestic and individual levels. As such, causal factors like cognitive limitations, psychological biases, and national cultures and institutional structures are most likely to determine whether policymakers will behave more in accordance with offensive realism or rational signaling models.[29]

Indeed, many such factors have long been highlighted in the literature. Alastair Iain Johnston's (1995) classic work on "cultural realism" argues that China's predilection for competitive behavior in the 20th century was deeply rooted in Chinese strategic culture, rather than the rational incentives of the international system. Pioneering applications of

[27]To determine where benign first-round cooperation was rational, we identified conditions where the SE was supported *given* the observed rates of cooperation under each combination of parameter values. These conditions are derived in Proposition 2 of Appendix A. Importantly, differences in guess accuracy are similar under conditions that support the SE in the purely theoretical model.

[28]This result cannot be explained by the existence of the all-defection equilibrium noted above, as it describes subjects' beliefs in response to their counterpart's first-round cooperation.

[29]We found no evidence of "learning effects" wherein players' strategies change with experience playing the game. These results are reported in Appendix B.

social psychology to IR have identified numerous mechanisms that promote suboptimally competitive actions, including fundamental attribution error (Stein 2013), loss aversion (McDermott 2001), and out-group bias (Mercer 1996). Given the demonstrated incoherence of offensive realism as a rationalist theory, our results showing beliefs and behaviors consistent with offensive realism are likely driven by variables identified in these non-rationalist theories.

Recent experimental work has further demonstrated how these psychological phenomena cause decisionmakers to systematically depart from rational behavior in international politics. For instance, Gottfried and Trager (2016) show that perceptions of fairness can significantly affect public support for particular negotiated bargains. Renshon, Lee, and Tingley (2017) similarly show that emotions may disrupt individuals' ability to rational respond to commitment problem incentives. Burcu Bayram (2017) demonstrates that "cosmopolitan" social identity can drive behavior that would otherwise, from a purely egoistic standpoint, appear irrational. Rathbun, Kertzer, and Paradis (2017) draw out the microfoundations of such departures from rationality, arguing that psychological factors like epistemic motivation help explain observed variations rational behavior. These studies further corroborate the non-rational underpinnings of offensive realist behavior.

5.1 Generalizability of the Experimental Results

Although our experiment used a sample composed primarily of undergraduate students, there are good reasons to think that our results yield important insights into the behavior of policymakers. First, the direction of bias we would expect from sampling students rather than policymakers actually bolsters our confidence in our core results in support of the model's comparative statics.[30] If the students in our sample possess, on average, lower levels of cognition and experience with security dilemma interactions than foreign policymakers, we would expect our subjects to deviate from the model's predictions more frequently than elites would.[31] Yet the comparative statics results presented above show

[30]Druckman and Kam (2011, 49) note that it is unhelpful to simply ask whether student samples are generalizable. Rather, we need to know "which particular characteristics of student samples might lead us to question whether the causal relationship detected in a student sample experiment would be systematically different from the causal relationship in the general population."

[31]Mintz, Redd and Vedlitz (2006) find that undergraduate students are significantly different from military officers in their choices, degree of information acquisition, decision strategies, and responses to uncertainty. It is worth noting, however, that the applicability of their findings to our study is limited. First, military officers might differ from civilian foreign policymakers in ways that students do not (for example, military officers are more reluctant than students to take no action and they collect less information about the situation at hand). Second, Mintz et al. asked subjects to offer policy responses after reading real-world vignettes, whereas we employed a much more abstract strategic setting that is likely to create more homogeneous

that our student sample acts, in the aggregate, in ways consistent with the rationalist expectations of the SD game. Thus, there is good reason to be confident that our support for the directional effects hypothesized by the SD game generalizes to the presumably more rational population of interest.

Nevertheless, this possible difference between our experimental sample and target population means that our inferences about the model's point predictions must be further qualified. We found that the magnitude of the effects of cooperative signals on the receiver's beliefs and the likelihood of subsequent cooperation were significantly lower than would be expected, even based on a generous interpretation of the theoretical model. Yet if policymakers systematically behave more rationally than our subjects, it is possible that the theoretical predictions would prove to be more accurate than our findings initially suggest.

This possibility, though real, certainly does not vitiate the value of our findings regarding the model's point predictions. First, it is not at all obvious that our subjects differ from policymakers in ways that would significantly attenuate our results. Our sample of majors in a high-ranking economics program may actually enhance the "experimental realism," of our study, i.e., whether participants take the study and treatments seriously and are cognitively capable of understanding the incentives that they face.[32] Thus, although we cannot be certain that our findings regarding the model's point predictions fully generalize to the population of interest, they do increase our confidence that foreign policymakers are less responsive to reassurance signals than the model hypothesizes. In other words, our findings suggest the possibility that offensive realism might retain some descriptive accuracy and explanatory power in the real world, despite its incoherence as a rationalist theory. This shifts the burden of proof toward reassurance optimists to demonstrate that our findings do not generalize to policymakers, and that offensive realism is invalid even as a behavioral theory. Our baseline results now provide crucial motivation for followup studies to assess their robustness with targeted samples that more closely resemble the population of interest.

In sum, the possibility that policymakers would act more rationally and hew more

incentive structures across diverse populations. Our higher degree of internal experimental control should therefore mitigate the external validity concerns that Mintz et al. highlight.

[32]Generally speaking, these subjects are more likely than the general population to think strategically and understand the incentives of the model. As Druckman and Kam (2011, 46) point out, this experimental realism is far more important for establishing external validity than whether the subjects match the population to which the results are intended to generalize ("mundane realism"). On overemphasis of mundane realism relative to experimental realism in experimental social science, see also McDermott (2002), Morton and Williams (2008, 345).

closely to the model's predictions bolsters our main findings in support of Kydd's comparative statics predictions, while reducing our confidence that the findings regarding the model's point predictions generalize to the population of interest. Still, the latter findings suggest that policymakers might be less responsive to cooperative signals than the reassurance literature would predict, and that offensive realism might retain viability as a behavioral theory.

6 Conclusion

This paper has presented a first-cut experimental test of Andrew Kydd's canonical model of reassurance under the security dilemma. The model predicts that benign states should be able to confidently identify each other and build trust by first cooperating on issues where the stakes are low enough that benign states will risk cooperation, but high enough that hostile types will not. Moreover, the model shows that this range of stakes exists no matter how distrustful the actors initially are, such that the security dilemma can always be overcome. This theoretical finding is devastating to the logic of offensive realism, which holds that states' intentions under anarchy are inherently unknowable and, consequently, that competition and conflict are unavoidable between rational benign states. Yet to date there have been no systematic attempts to evaluate these propositions empirically.

Our experimental results support Kydd's hypothesized directional effects. Actors with benign intentions were best able to reassure each other and achieve second-round cooperation under intermediate first-round stakes. However, the credibility subjects assigned to cooperative signals and the rates of second-round cooperation were markedly lower than the model predicts. Even under ideal conditions of high initial trust and intermediate stakes, successful reassurance only occurred around half the time, and a significant proportion of subjects who played suboptimally revealed overly pessimistic beliefs in keeping with offensive realism's claim that actors adopt "worst-case" assumptions. Furthermore, contrary to Kydd and consistent with offensive realist critiques, our results show that benign players were significantly less likely to achieve cooperation when initial trust was low. Thus, while our results imply that Kydd's reassurance mechanism does indeed mitigate the security dilemma, offensive realism might remain a viable theory of human behavior under anarchy.

One need not look far to find cases of offensive realist thinking in the contemporary world. US foreign policy under Donald Trump has exhibited a nearly unambiguous offensive realist worldview concerning both allies and potential rivals. For example, despite

widespread uncertainty and disagreement among experts about China's intentions (Yoder 2020), the Trump administration has expressed high confidence that China is hostile to the US (Trump 2017). Correspondingly, the administration has adopted policies of economic containment and military competition toward China. Our experimental findings suggest that while these pessimistic, worst-case attitudes are not a rational response to international anarchy, they are also not necessarily atypical, and may explain a great deal of competitive behavior in international politics.

References

Acharya, Avidit and Kristopher Ramsay. 2013. "The Calculus of the Security Dilemma." *Quarterly Journal of Political Science* 8(2):183–203.

Bayram, A. Burcu. 2017. "Due Deference: Cosmopolitan Social Identity and the Psychology of Legal Obligation in International Politics." *International Organization* 71(S1):S137–S163.

Brooks, Stephen. 1997. "Dueling Realisms." *International Organization* 51(3):445–477.

Copeland, Dale. 2000. "The Constructivist Challenge to Structural Realism: A Review Essay." *International Security* 25(2):187–212.

Davis, Douglas D. and Charles A. Holt. 1993. *Experimental Economics*. Princeton, NJ: Princeton University Press.

Dickson, Eric. 2009. "Do Participants and Observers Assess Intentions Differently During Bargaining and Conflict?" *American Journal of Political Science* 53(4):910–930.

Druckman, James and Cindy Kam. 2011. Students as Experimental Participants: A Defense of the 'Narrow Data Base'. In *Cambridge Handbook of Experimental Political Science*, ed. James Druckman, Donald Green, James Kuklinski and Arthur Lupia. New York: Cambridge University Press chapter 4, pp. 41–57.

Edelstein, David. 2002. "Managing Uncertainty: Beliefs about Intentions and the Rise of Great Powers." *Security Studies* 12(1):1–40.

Fearon, James. 1997. "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." *The Journal of Conflict Resolution* 41(1):68–90.

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

Glaser, Charles. 1997. "The Security Dilemma Revisited." *World Politics* 50(1):171–201.

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

Gottfried, Matthew S and Robert F Trager. 2016. "A Preference for War: How Fairness and Rhetoric Influence Leadership Incentives in Crises." *International Studies Quarterly* 60(2):243–257.

Hafner-Burton, Emilie, Stephen Haggard, David Lake and David Victor. 2017. "The Behavioral Revolution and International Relations." *International Organization* 71(Supplement):S1–S31.

Haynes, Kyle. 2019. "A Question of Costliness: Time Horizons and Interstate Signaling." *Journal of Conflict Resolution* 63(8):1939–1964.

Haynes, Kyle and Brandon K. Yoder. 2020. "Offsetting Uncertainty: Reassurance with Two-Sided Incomplete Information." *American Journal of Political Science* 64(1):38–51.

Hyde, Susan. 2015. "Experiments in international relations: Lab, survey, and field." *Annual Review of Political Science* 18:403–424.

Ikenberry, G. John. 2001. *After Victory: Institutions, Strategic Restraint, and the Rebuilding of Order After Major Wars*. Princeton, NJ: Princeton University Press.

Jervis, Robert. 1978. "Cooperation Under the Security Dilemma." *World Politics* 30(2):167–214.

Johnston, Alastair Iain. 1995. *Cultural Realism: Strategic Culture and Grand Strategy in Chinese History*. Princeton, NJ: Princeton University Press.

Johnston, Alastair Iain. 2003. "Is China a Status Quo Power?" *International Security* 27(4):5–56.

Kertzer, Joshua D. 2017. "Resolve, Time, and Risk." *International Organization* 71(S17):S109–S136.

Kertzer, Joshua D., Brian Rathbun and Nina Srinivasan Rathbun. Forthcoming. "The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations." *International Organization* .

Kertzer, Joshua D. and Kathleen M. McGraw. 2012. "Folk Realism: Testing the Microfoundations of Realism in Ordinary Citizens." *International Studies Quarterly* 56(2):245–258.

Kydd, Andrew. 1997. "Sheep in Sheep's Clothing: Why Security Seekers Do Not Fight Each Other." *Security Studies* 7(1):114–155.

Kydd, Andrew. 2000. "Trust, Reassurance, and Cooperation." *International Organization* 54(2):325–357.

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

Layne, Christopher. 1993. "The Unipolar Illusion: Why New Great Powers Will Rise." *International Security* 17(4):5–51.

Ledyard, John O. 1995. Public Goods: A Survey of Experimental Research. In *Handbook of Experimental Economics*, ed. John Kagel and Alvin Roth. Princeton, NJ: Princeton University Press pp. 111–194.

Martin, Lisa L. 2017. "International Institutions: Weak Commitments and Costly Signals." *International Theory* 9(3):353–380.

McDermott, Rose. 2001. *Risk-taking in international politics: Prospect theory in American foreign policy*. Ann Arbor, MI: University of Michigan Press.

McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis* 10(4):325–342.

McDermott, Rose, Jonathan Cowden and Cheryl Koopman. 2002. "Framing, Uncertainty, and Hostile Communications in a Crisis Experiment." *Political Psychology* 23(1):133–149.

Mearsheimer, John. 2001. *The Tragedy of Great Power Politics*. New York: W.W. Norton and Co.

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

Mintz, Alex, Steven B. Redd and Arnold Vedlitz. 2006. "Can We Generalize from Student Experiments to the Real World in Political Science, Military Affairs, and International Relations?" *Journal of Conflict Resolution* 50(5):757–776.

Mitzen, Jennifer. 2006. "Ontological Security in World Politics: State Identity and the Security Dilemma." *European Journal of International Relations* 12(3):341–370.

Montgomery, Evan Braden. 2006. "Breaking Out of the Security Dilemma: Realism, Reassurance, and the Problem of Uncertainty." *International Security* 31(2):151–185.

Morton, Rebecca B. 1999. *Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science*. New York: Cambridge University Press.

Morton, Rebecca B. and Kenneth Williams. 2008. Experimentation in Political Science. In *The Oxford Handbook of Political Methodology*, ed. Janet Box-Steffensmeier, Henry Brady and David Collier. Oxford, UK: Oxford University Press chapter 14, pp. 339–356.

Quek, Kai. 2016. "Are Costly Signals More Credible? Evidence of Sender Receiver Gaps." *Journal of Politics* 78(3):925–940.

Quek, Kai. 2017*a*. "Rationalist Experiments on War." *Political Science Research and Methods* 5(1):123–142.

Quek, Kai. 2017*b*. "Type II Audience Costs." *Journal of Politics* 79(4):1438–1443.

Rathbun, Brian C., Joshua D. Kertzer and Mark Paradis. 2017. "Homo Diplomaticus: Mixed-Method Evidence of Variation in Strategic Rationality." *International Organization* 71(S1):S33–S60.

Renshon, Jonathan, Julia J Lee and Dustin Tingley. 2017. "Emotions and the Micro-Foundations of Commitment Problems." *International Organization* 71(S1):189–218.

Rosato, Sebastian. 2015. "The Inscrutable Intentions of Great Powers." *International Security* 39(3):48–88.

Schweller, Randall. 1996. "Neorealism's Status-Quo Bias: What Security Dilemma?" *Security Studies* 5(3):90–121.

Stein, Janet Gross. 2013. Psychological Explanations of International Decision Making and Collective Behavior. In *Handbook of International Relations*, ed. Walter Carlnaes, Thomas Risse and Beth Simmons. London: Sage pp. 195–219.

Tang, Shiping. 2009. "The Security Dilemma: A Conceptual Analysis." *Security Studies* 18(3):587–623.

Tingley, Dustin. 2011. "The Dark Side of the Future: An Experimental Test of Commitment Problems in Bargaining." *International Studies Quarterly* 55:521–544.

Tingley, Dustin and Barbara Walter. 2011. "The Effect of Repeated Play on Reputation Building: An Experimental Approach." *International Organization* 65(2):343–365.

Trump, Donald J. 2017. "National Security Strategy of the United States of America." *Office of the President of the United States* .

Weiss, Jessica Chen. 2013. "Authoritarian Signaling, Mass Audiences, and Nationalist Protest in China." *International Organization* 67(1):1–35.

Wendt, Alexander. 1999. *Social Theory of International Politics*. Cambridge, UK: Cambridge University Press.

Yarhi-Milo, Keren. 2014. *Knowing the Adversary: Leaders, Intelligence Organizations, and Assessments of Intentions in International Relations*. Princeton, NJ: Princeton University Press.

Yoder, Brandon K. 2019*a*. "Hedging for Better Bets: Power Shifts, Credible Signals, and Preventive Conflict." *Journal of Conflict Resolution* 63(4):923–949.

Yoder, Brandon K. 2019*b*. "Retrenchment as a Screening Mechanism: Power Shifts, Strategic Withdrawal, and Credible Signals." *American Journal of Political Science* 63(1):130–145.

Yoder, Brandon K. 2020. "How Informative are China's Foreign Policy Signals? IR Theory and the Debate about China's Intentions." *Chinese Journal of International Politics* Forthcoming.