Four Costly Signaling Mechanisms

Kai Quek[*]

Two mechanisms of costly signaling are known in international relations: sinking costs and tying hands. I show that there exist four mechanisms of costly signaling that are equally general as each other. I develop the new mechanisms of installment costs and reducible costs, and contrast them with sunk costs and tied-hands costs. I then conduct experiments to test the four signaling mechanisms. I find that each mechanism can improve credibility when the costs are high, but only reducible costs can improve credibility even when the costs are low.

Version: November 2020

[*] Associate Professor of Politics, Department of Politics and Public Administration, The University of Hong Kong. 940 Jockey Club Tower, Pokfulam Road, Hong Kong. Email: quek@hku.hk.

I am grateful to David Clark, James Fearon and Nicholas Miller for their comments on the paper; Dan Altman, Christina Davis, Brad DeWees, Erik Gartzke, Naima Green, Austin Jordan, Koji Kagotani, Joshua Kertzer, Shuhei Kurizaki, Marika Landau-Wells, Aidan Milliff, Yon Soo Park, Reid Pauly, Alastair Smith, Michael Tomz, Ketian Zhang, and seminar participants at Harvard, MIT and the 2018 Pacific International Politics Conference for their feedback; John Koo and Eddy Yeung for their assistance; and the University of Hong Kong and its OYRA scheme for financial support. I also thank the editors at *APSR* and seven anonymous reviewers for their thoughtful suggestions. The standard disclaimers apply.

Credibility lies at the heart of many strategic questions. When credible signals cannot be sent, bad outcomes, including war (Fearon 1995), can arise. The problem is that a person cannot credibly signal her private information when others know she has an incentive to bluff. The most famous solution to the problem is costly signaling.[1] When a resolved person sends a costly signal that an unresolved person is unlikely to send due to the costs, the resolved person may separate herself from the unresolved.

Two general mechanisms of costly signaling are known in international relations (IR): sinking costs and tying hands (Fearon 1997). In the first, the costs incurred by the signaler are sunk and irrecoverable. In the second, the signal is costly in the future if the signaler reneges, but costless otherwise. Almost all subsequent work on costly signaling in IR have worked off Fearon's (1997) canonical mechanisms. This paper proposes two additional mechanisms that are equally general as the canonical two.

The paper contributes on three fronts. Theoretically, I establish the existence of four mechanisms of costly signaling. Leveraging the dimensions of time horizon and cost contingency that define sunk and tied-hands costs, I derive the two new mechanisms of installment and reducible costs. Substantively, I develop the new mechanisms by analyzing their signaling logic and illustrating their relevance in international politics. Empirically, I provide experimental evidence on whether and how the signaling mechanisms shape the perception of credibility. I find that sunk, tied-hands, and installment costs produce kinked credibility curves, whereby credibility does not improve across costless and low-cost signals, but spikes for a high-cost signal. Reducible costs are the exception, whereby credibility improves significantly across costless,

[1] See, e.g., the arguments in Spence (1973) and Rothschild and Stiglitz (1976) on costly signaling and screening, which laid the foundations for the economic analysis of asymmetric information (Nobel Foundation 2001).

low-cost, and high-cost signals.

I proceed as follows. First, I provide the theoretical framework and develop the substantive implications. Further, I describe a series of controlled experiments that test the credibility effects of the four signaling mechanisms. Finally, I discuss the potential implications and unresolved questions for future research.

Four Mechanisms

Spence (1976, 593) proposed "two qualitatively different types" of signaling mechanisms in economics: "exogenously costly signaling," which is formalized in Spence's (1973) pathbreaking paper, and "contingent contract," which is discussed in Spence (1976). Defined in market terms, the exogenously costly signal is "an activity engaged in by the seller, which has a cost that varies with product quality, independent of the buyer's response to the activity"; and the "contingent contract" is "a menu of options for the seller that are created by virtue of the buyer's subsequent ability to observe the product quality directly, and, to transact with the seller at that point" (Spence 1976, 593). These ideas were clarified and adapted into IR by Fearon (1997) as two general mechanisms of costly signaling: *sinking costs* and *tying hands*.

Fearon (1997, 70) defined sunk-cost signals as "actions that are costly for the state to take in the first place but do not affect the relative value of fighting versus acquiescing in a challenge." Tied-hands signaling involves "an action that increases the costs of backing down if the would-be challenger actually challenges but otherwise entails no costs if no challenge materializes." Almost all subsequent work on costly signaling in IR have worked off these two mechanisms (reviewed in Gartzke et al. 2017).

A large body of research has applied costly signaling across diverse domains. Directly relevant domains in IR include diplomacy, deterrence, and reputation (reviewed in Trager 2016 and Dafoe, Renshon and Huth 2014; see also Sartori 2005). Examples include the recent work on leader-specific reputation (e.g. Lupton 2018; McManus 2018; Wu and Wolford 2018), covert communication (e.g. McManus and Yarhi-Milo 2017; Carson and Yarhi-Milo 2017), alliance commitments (e.g. Fang, Johnson and Leeds 2014), cyberwarfare (e.g. Gartzke and Lindsay 2017), and fait accompli (e.g. Altman 2017; Tarar 2016). Scholars have also applied costly signaling to better understand the conditions that sustain peace after civil wars (e.g. Hartzell and Hoddie 2007; Mattes and Savun 2009; Mattes and Vonnahme 2010; see also Reiter 2009), the implications of trade interdependence (e.g. Gartzke and Westerwinter 2016), the dynamics of trust and reassurance (e.g. Acharya and Ramsay 2013; Blankenship 2020; Chan 2012; Kydd 2005; Kydd and McManus 2017; Haynes and Yoder 2020), and the signaling functions of international institutions such as the International Monetary Fund and United Nations (e.g. Dai, Snidal and Sampson 2017; Fang 2008; Simmons 2000; Tago and Ikeda 2015; Voeten 2005). Scholars have also applied costly signaling across the social sciences, from economics and sociology to anthropology and archaeology (reviewed in Riley 2001, Gambetta 2009, Bliege Bird and Smith 2005, and Quinn 2019).

Previous research has sharpened our understanding of costly signaling in important ways, but several puzzling aspects have also emerged. Conceptually, if sunk costs are sunk and have no time horizon by definition, how do we interpret situations where the signaler is committed to fixed costs over a future time horizon? Practically, we rarely see policymakers citing the logic of sinking costs when they send costly signals. Are there other logics of signaling that are more salient to policymakers? Empirically, recent research provides mixed evidence on the effectiveness of costly signaling (e.g.

Fuhrmann and Sechser 2014; Quek 2016; Snyder and Borghard 2011; Trachtenberg 2012; Yarhi-Milo, Renshon and Kertzer 2018). Yet, costly signaling is ubiquitous and governments often send costly signals to one another in the real world. Are there important theoretical aspects of costly signaling that have yet to be uncovered?

These puzzles raise the question of whether our existing theory of costly signaling is complete. Specifically, are there only two general mechanisms of costly signaling?

I check the logical completeness of the system in three steps. The first step returns to first principles: the nature of a costly signaling mechanism depends on the nature of the *costs* of the signal.

Next, I identify and generalize the nature of the costs that define each signaling mechanism. Fearon's (1997) sunk-cost mechanism is defined by costs with two properties: the costs are ex-ante because they are already incurred in the present, and non-contingent because they are incurred regardless of whether the signaler fulfills the threat. The tied-hands mechanism is also defined by two properties: the costs are ex-post because they are incurred in the future, and contingent because they are only incurred if the signaler does not fulfill the threat.[2] The dual properties of time horizon and cost contingency are the theoretical pivots in Fearon's (1997) mechanisms which lend each mechanism its particular character. These properties not only differentiate one mechanism from the other, but also provide the common denominators that unify both into the same framework.

Finally, I check if this framework is complete with only two signaling mechanisms. Because there are two properties, and because each property is binary, there are four

[2] A tied-hands signal is costly "if the would-be challenger actually challenges but otherwise entails no costs if no challenge materializes" (Fearon 1997, 70).

equally general mechanisms in the family. A two-by-two yields four possibilities, but only sunk and tied-hands costs are discussed in the literature (Table 1).

Table 1: Four Costly Signaling Mechanisms

	Non-Contingent	**Contingent**
Ex Ante	Sunk Costs	*Reducible Costs*
Ex Post	*Installment Costs*	Tied-Hands Costs

The theoretical universe for costly signaling is therefore larger than what the literature has assumed. Beside sunk costs and tied-hands costs, there are also installment costs and reducible costs. I define the new mechanisms and explain their signaling logic below.

Installment Costs and Reducible Costs

Installment Costs

Installment costs (IC) are fixed costs that will be incurred in the future. They can be incurred in one installment in the short or long-term future, or in a stream of installments over a definite or open-ended time horizon. These costs are ex-post because they are paid in the future (which differentiates them from sunk costs), and non-contingent because they are incurred even if the receiver does not respond or challenge (which differentiates them from tied-hands costs).[3]

[3] For a tying-hands signal, the signal is costly "if the would-be challenger actually challenges but otherwise entails no costs if no challenge materializes" (Fearon 1997, 70).

Let's deploy an example to make things concrete. Suppose the United States builds a nuclear base in a foreign country to signal its resolve to defend that country. The ex-ante costs of building the base are sunk costs; the ex-post costs of maintaining the nuclear base are installment costs. *At the point of signaling*, the credibility perception of the receiver would be based on two different types of costs incurred by the signaler: sunk costs and installment costs.[4] Because the signal involves costs of two kinds, the overall credibility effect can be broken up into two parts: a credibility effect that comes from the ex-ante costs, which we know as sunk costs, and a credibility effect from the ex-post costs, in the form of installment costs. As we will see later, the two have different theoretical properties and should not be conflated as one.

In reality, the U.S. has been maintaining nuclear weapons in several foreign countries since the Cold War. These weapons can help to signal U.S. commitment to defend those countries, even though they are militarily superfluous in that they are unlikely to change the outcome of a nuclear war, given long-distance delivery capabilities and mutually assured destruction (see O'Neill 1990; Slantchev 2011, 36; Fuhrmann and Sechser 2014, 924). To maintain these weapons, costs are incurred in installments on a recurrent basis over time. The installment costs include material costs – for example, what the U.S. must pay to protect the nuclear base, ensure the operational effectiveness of the weapons, and remunerate the specialized personnel. These costs are not trivial. In the case of U.S. bases in Western Europe, the ex-post installment

[4] In fact, there will always be some positive amount of ex-post installment costs involved regardless of what the U.S. does. Even if the U.S. decides to leave, dismantling the nuclear base will incur costs. Withdrawal of weapons and personnel involve substantial costs and risks, and redeploying nuclear weapons back to the U.S. may not be much cheaper than deploying them to Europe in the first place. Such costs constitute installment costs that are irrevocable. But there is also a second kind of installment costs that involve some degree of a commitment problem, which are not completely non-contingent insofar as the expected amount of costs varies according to the receiver's probabilistic beliefs. We will discuss the two kinds of installment costs at the end of this section.

costs should be at least comparable to the ex-ante sunk costs.[5] In addition, installment costs can also include non-material costs – such as the risks of nuclear accident, abuse, misuse, and the loss of control.[6] These may be more salient to the receiver (see, e.g., Schelling 1966) than the simple monetary costs of creating a nuclear base.

This example is a mixed case: the same signal can contain different cost components with each invoking a different mechanism. Here, sunk costs and installment costs co-occur in the same case, with each contributing a separate effect that combine into the overall credibility of the signal. That a signal has a sunk-cost component does not preclude it from having an audience-cost or installment-cost component. In fact, real-world cases of sinking costs in IR are usually mixed cases. Troop mobilization is the commonly cited example of sinking costs, but it is clearly a mixed case (Fearon 1997; Slantchev 2005). Fearon (1997, 70) argued there are "few examples of the pure case." Slantchev (2005) developed the argument further, showing mobilization not only sinks costs, but also ties hands by increasing the probability of victory. And unless mobilization is conducted in perfect secrecy, it will also involve international and domestic audience costs.[7]

[5] According to the U.S. Department of Defense, building the specialised infrastructure to hold nuclear weapons within existing local airbases costs US$384 million in total. In comparison, the U.S. nuclear arsenal costs US$33.4 billion in 2019 and the estimate for the next 10 years is US$494 billion. Assuming the costs are spread out proportionally over the number of operational nuclear warheads (3,800 as of 2019) and using a conservative estimate of 20 warheads in one NATO nuclear base, the cost of maintaining a NATO nuclear base would be about US$260 million per year. See Congressional Budget Office, "Projected Costs of U.S. Nuclear Forces, 2019 to 2028," January 2019; Department of Defense, "Military Construction Program FY 2020 Budget: Justification Data Submitted to Congress," March 2019.

[6] For example, a Blue Ribbon Review investigation, focusing on nuclear bases in Europe, was triggered by a "notorious incident in August 2007 when the U.S. Air Force lost track of six nuclear warheads for 36 hours as they were flow[n] across the United States without the knowledge of the military personnel in charge ...." Hans M. Kristensen, "USAF Report: Most Nuclear Weapon Sites In Europe Do Not Meet US Security Requirements," Federation of American Scientists Blog, 19 June 2008.

[7] Fearon (1997) argued that President Clinton had signaled his resolve to intervene in Haiti with "a massive military mobilization that created very significant audience costs" (71).

Installment costs have not been systematically studied in the IR signaling literature. Unlike sunk costs, which are by definition ex-ante (sunk), installment costs are by definition ex-post. It is important to analyze them separately because their underlying logics differ. As Fearon (1997, 70) argued on differentiating sinking costs from tying hands – despite the two are often coupled together in reality – "[i]t is important to see ... that two distinct mechanisms are at work, and we need to analyze them separately as ideal types to understand the strategic logic of mixed cases."[8]

It is important to see that installment costs and sunk costs differ, so we can analyze their effects separately to understand the true character of mixed cases. A convenient way to distinguish installment costs from traditional sunk costs is to formulate the former as a form of "expected future sunk costs" at the point of signaling. But this formulation creates a logical contradiction. Sunk costs, by definition, are sunk. Future sunk costs are therefore future past costs, which is a contradiction: the same costs cannot be already sunk (ex ante) and yet paid in the future (ex post). One might ignore the contradiction and apply an expected utility framework nonetheless. Notice, however, the expected utility framework falls into the economics and psychology of time horizon that define installment costs but not sunk costs (since there is no future time horizon in sunk costs).

There are several remarkable points to note about installment costs. First, although the costs would be incurred in the future, they impact the receiver's credibility calculation in the present. The subtlety is that whereas sunk costs are calculated as given (because they are ex-ante), installment costs are calculated on *beliefs* about the future

[8] Troop mobilization is commonly cited as an example of a sunk-cost signal, and audience costs as an example of a tied-hands signal. Mobilization is not a pure case of a sunk-cost signal (Fearon 1997; Slantchev 2005). Recent research suggests audience costs are not a pure case of a tied-hands signal either (Kertzer and Brutger 2016).

(because they are ex-post). As a consequence, the credibility effect of installment costs is qualitatively different from that of sunk costs. The IC signal opens a time horizon in the receiver's credibility assessment. In fact, all costly signals of a continuous nature will have some time horizon – only a pure sunk-cost signal such as "burning money" is of a one-shot discrete nature. But because the existing literature has worked from the idealized model of a pure sunk-cost signal (which is discrete), many costly signals of a continuous nature in the real world have been taken as discrete, and their time-horizon effects on credibility omitted from view.

Second, IC signaling changes the structure of interaction as we move from the one-shot nature of standard sunk-cost signaling to a time-horizon structure in IC signaling. This allows us to examine intertemporal dynamics whereby beliefs and time discounts evolve with time. We will return to this later; the key point is that thinking dynamically about signaling costs is useful in many realistic settings where the costs are not incurred at the point of signaling, but committed over a future time horizon. In many cases of state-to-state signaling, there is a time gap between announcement and realization, signaling and implementation.[9]

Third, the time horizon is a central element in installment costs, and it opens a gradation of possibilities that are not open to sunk costs. An ex-ante cost has no future time horizon. However, the ex-post nature of installment costs can have varying degrees of gradation, based on the varying shape and duration of the time horizon. Thus, two actions can fall into the same signaling ideal-type, but their effects on credibility can differ in degree. The ideal types are distinct, but real-world cases can have gra-

[9] For example, when the U.S. announces that it would build a nuclear base, the signal does not involve a sunk cost incurred immediately at the point of signaling, but an installment cost committed for the future. It is the construction of the nuclear base at the implementation stage that is a sunk-cost signal.

dations in degree within. Just as sunk costs and tied-hands costs are “a distinction between two ‘ideal types”’ (Fearon 1997, 69), so are the distinctions between the four mechanisms.

Finally, there are two types of installment costs. One is a pure type that does not involve a commitment problem because the sender is irrevocably committed to those costs. For example, there is a physical object that simply cannot be dismantled in the short term. Or dismantling it is more expensive than just maintaining it in the foreseeable future, so there is no incentive to renege. For this kind of installment costs, there are time-horizon effects on the receiver’s credibility calculation, but no time inconsistency. The second type of installment costs involves time inconsistency – there is a positive probability that the installments may not be fully paid in the future. Unlike the first type of installment costs, the second type is not completely non-contingent if the exact amount of costs incurred varies according to the receiver’s probabilistic beliefs.[10] IC signals of this kind open up new possibilities in the intertemporal interaction. One implication is that the sender may need some form of tying-hands mechanism to support IC signaling of this type. So, very interestingly, this form of IC signaling is distinguished from sunk-cost signaling by the need for an exogenous tying-hands mechanism.[11] It is remarkable that while tying hands solves commitment problems, installment costs of this kind create commitment problems which in turn need tying hands to solve.

To summarize, an IC signal differs from a sunk-cost signal in at least three ways:

[10] This type of IC signal may or may not be able to achieve credibility on its own, depending on the magnitude of the commitment problem. If the magnitude is small, this kind of IC may stand on its own. But if the commitment problem is severe, it will need to be supported by another signaling mechanism (e.g. tying hands) to achieve credible commitment.

[11] I thank a reviewer for this observation.

1. Installment costs have a time horizon but sunk costs do not. Because the economics and psychology of time horizons apply to installment costs but not to sunk costs, the two mechanisms should have different substantive effects on credibility.

2. Time inconsistency is possible for installment costs but impossible for sunk costs. It is not possible for sunk costs to have a commitment problem because they are already sunk.

3. Installment costs and sunk costs are distinct by definition. A logical contradiction arises if we take them as equivalent: it means that the same cost is both ex-ante and ex-post, which is contradictory.

Reducible Costs

Reducible costs (RC) are costs that have been paid but which can be offset in the future contingent on the signaler's action.[12] These costs are ex-ante because they are paid at the point of signaling, but contingent because they can be offset if the signaler fulfills her threat or promise. Because the signaler can offset the costs incurred in the past if the threat is fulfilled in the future, the threat becomes more credible.[13]

Many cases of real-world signaling involve some degree of reducible costs, which we will see later when we discuss the specific pathways of RC signaling. Because RC signaling is quite ubiquitious in practice, some of its features have been discussed by

[12] The original term I used was "recoverable costs", but it seemed less intuitive.

[13] The RC signal is a credible threat because the signal makes it less costly to fulfill the threat (there are three ways whereby RC signaling achieves this, as discussed later). An interesting possibility is an inverse form of reducible costs, which make it more costly to fulfill the threat and thereby reduce credibility. We will not pursue this possibility in this paper, as our focus is on the credible communication of private information.

scholars and applied by policymakers; the literature did not actually ignore it. Rather, because only two signaling mechanisms were known, the literature has conflated it with sunk-cost and tied-hands signaling. It is useful to disaggregate these different classes of signal and explore the distinct implications of each.

While both involve contingent costs, an RC signal is an ex-ante cost whereas a tied-hands signal is an ex-post cost.[14] While both are ex-ante costs, an RC signal involves contingent costs and benefits, but a sunk-cost signal does not.[15] Reducible costs are costs that have been paid in the past but which can be offset in the future. The signaling costs are reducible in that they are offsettable. What RC signaling does is to intentionally change the cost calculation for fulfilling the threat or promise.

RC signaling only requires changing the ex-post cost calculation for the signaler; the mechanism does *not* require changing the balance of power or probability of winning. In crisis diplomacy, RC signaling works by changing the signaler's ex-post cost calculation by making it cheaper to fight, but it does not need to change the probability of victory (which will in turn change the receiver's war payoff). RC signaling differs from existing ways of conceptualizing military signals such as mobilization, which involve changing the probability of winning.[16]

[14] A tied-hands signal is defined as "an action that increases the costs of backing down if the would-be challenger actually challenges but otherwise entails no costs if no challenge materializes" (Fearon 1997, 70).

[15] Here we may think in terms of the distinction between preference and behavior. The benefits of the sunk-cost signal to the sender – such that honest types will send it – depend on the sender's underlying *preferences*. The benefits of the RC signal, however, will also depend on the sender's *behavior*. The honest type also benefits from RC signaling when it is challenged and follows through. I thank a reviewer for suggesting this distinction. By implication, RC signaling also differs in strategic terms from sunk-cost signaling, which we will discuss later.

[16] In Fearon (1997) and Slantchev (2005), a military signal (such as mobilization) is one that involves a change in the probability of winning. In Fearon (1997), "the probability of winning a conflict [with a militarily relevant signal] should increase with the size of the signal" (82). In Slantchev (2005), "mobilization simultaneously sinks costs, because it must be paid for regardless of the outcome, and

In the context of crisis diplomacy, the ex-ante costs of the RC signal can be offset ex-post through three pathways:

1. **The costs are reduced computationally.** The signaler prepays a part of the total cost of fighting, making it cheaper to fulfill its threat to fight in the future.[17] Because one part of the total cost is prepaid in advance, the total cost is reduced *computationally* – by simple calculation in a purely accounting sense. Prepaying some of the costs of fighting allows a state to signal its willingness to fight, and a state which prepays a fraction of the cost of fighting (e.g. by buying weapons) is more credible than one which doesn't. If a war occurs, weapons would be needed anyway – the costs that would have been paid later are prepaid in advance with RC signaling. An important difference is that prepaying the costs can transmit informational value toward credible deterrence, but "postpaying" the same costs in a war has no such informational value. Indeed, a good way to show that one is willing to pay the costs of fighting is to have already paid some of the costs upfront – just as a good way to show that one is serious about doing a PhD is to have already taken graduate courses as an undergraduate; and a good way to threaten to break up and move out after a quarrel is to have already packed the bags. The same signaling logic is used by some leaders in real-world crises. For example, in the 1962 Sino-Indian border crisis, Premier

ties hands, because it increases the probability of winning should war occur" (533). Specifically, the military signal has to change "the probability of prevailing in an armed conflict ... it is not enough that the action affects one's own expected value of war, it must also affect one's opponent's value of war" (Slantchev 2011, 66-67). RC signaling may be expanded to include cases where both the signaler's and receiver's cost calculations change with the signal – but for the signaling mechanism to work, the former is sufficient.

[17] I thank Dan Altman for a discussion on this point, and a reviewer for nudging me to clarify if reducible costs are "recouped costs" and/or "offsetting benefits" (the answer is both). Here, credibility and the cost of implementing the threat are related in that "[i]f part of the implementation cost is already paid in sending the costly signal ... then the signaled threat (e.g. the use of force) becomes cheaper to implement, and thus more credible" (Quek 2013, 31).

Zhou Enlai argued to his colleagues that "to fight a bit ... would cause some people to understand things more clearly" (Garver 2006, 115). Here, the cost of fighting is reduced in a computational or accounting sense: fighting becomes cheaper simply because some of the costs were already paid in advance. Note that here fighting becomes cheaper – and the informational value of the signal holds – even if the probability of victory and balance of power are unchanged.

2. The costs are reduced substantively. The signaler pays for particular preparations or investments in the present that would substantively (rather than computationally) reduce the cost of fighting in the future – even though these expenditures may not change the probability of victory. This can apply in several ways. Firstly, there are types of investments that decrease the cost of war but do not affect success in the battlefield. Examples include launching a civil defense program, building bomb shelters for civilians, or evacuating people living near the border during a territorial crisis. Civil defense programs may contribute little to military success in the battlefield, but they have an important signaling function – "one little recognized and much underrated, that could prove enormously important in a crisis ... if one wants demonstrations" (Schelling 1966, 239). Likewise, as U.S. Air Force Chief of Staff Carl Spaatz argued, "whether the Soviet Union is building civilian shelters for its own people [can] be one of the most significant indicators of its intentions" (United States Congress House Committee on Armed Services 1963, 3044). The underlying mechanism here is RC signaling, whereby the costs of fulfilling the threat of war are reduced through these moves. Secondly, there could be cases where military investments – whether early or late – lead to a similar fighting capacity, but early investment can substantively reduce the total costs incurred. For instance, it could be that mobilizing troops today or tomorrow would lead to a roughly similar tactical situation or fighting capacity (and thus a comparable probability of victory), but an early mobilization is less costly than a late mobilization. Finally, there are situations where a change to the balance of power by one side triggers a strategic response by the other side that neutralizes the change. If one's action puts the opponent at a disadvantage, the opponent may respond with a counteraction to eliminate the disadvantage and return the probability of victory back to the previous status-quo. In short, even if the probability of winning is held constant, RC signaling can make the threat to fight more credible because it makes it cheaper for the signaler to fulfill her threat. And because it does not shift the posterior balance of power and the probability of winning, RC signaling may communicate resolve without increasing the risk of war as much as in existing models of military signaling based on a tied-hands logic (Fearon 1997; Slantchev 2005).[18]

3. The costs are reduced indirectly. The net costs of fighting can also be reduced in an indirect way by increasing the benefits of fighting at the domestic level, even when these benefits do not change the probability of winning a war. For example, a government that increases taxes for military defense spending may suffer domestic political costs at the first instance, but these domestic costs may be offset in the future if indeed a war occurs (affirming the foresight of the government doing the right thing). Like before, because the net costs of exercising the option to fight are reduced, the threat to fight becomes more credible.

A recurrent puzzle in signaling research is that states sometimes signal resolve with actions that do not seem too costly or likely to improve military outcomes. Jervis (1970) highlighted the puzzle, observing that the actions chosen by President Kennedy

[18] See footnote 16.

during the Berlin crisis in 1961 “did not put the US in an appreciably better military position and they were cheap” (23); and his naval blockade in the Cuban Missile Crisis in 1962 was a “signal which could have been a bluff and did not involve any strong proof” (22). Thus, “receivers can be expected to at least partially discount them and one might therefore expect signals to be relatively rare or unimportant. But in fact a great deal of modern international relations consists of signals” (Jervis 1970, 23).

IC and RC signaling may shed light on this puzzle. First of all, many signaling actions are mixed cases. They involve more than one signaling mechanism. For example, Kennedy’s announcement of the blockade in the Cuban Missile Crisis did not only involve sunk costs and audience costs. It also involved IC signaling in the form of the costs and risks of maintaining the naval blockade in the foreseeable future. The ExComm expected the crisis to continue over time (Naftali and Zelikow 2001). Robert Kennedy believed the blockade could last for months (Fursenko and Naftali 1997, 231); President Kennedy’s announcement of the blockade warned of “many months of sacrifice ... months in which both our patience and our will will be tested” (Kennedy 1962).

The blockade was not the only signal sent by President Kennedy – nor did it seem to be sufficient on its own.[19] Kennedy announced three military moves on October 22: the first was the blockade and the second was the surveillance of Cuba by the USAF. The third reinforced the U.S. base at Guantanamo and “evacuated today the dependents of our personnel there” (Kennedy 1962) – an instance of RC signaling that would substantively reduce the ex-post costs of war. And the GRU (Soviet Main Intelligence Directorate) report of an order issued to mobilize U.S. hospitals to prepare

[19] Soviet sources suggest that it did not immediately convince Khrushchev to back down. See Fursenko and Naftali (1997, 247-8).

for taking casualties – another instance of an RC signal – was one of the key pieces of intelligence Khrushchev received right before he decided to write his October 26 letter proposing a peaceful settlement,[20] as he found "today's batch of intelligence ... too unequivocal to ignore" (Fursenko and Naftali 1997, 262).

In a similar vein, President Kennedy in the 1961 Berlin crisis "reaffirmed the Western resolve to remain in Berlin [when he] asked Congress to expand American military forces, inaugurated a civil defense program, and made plans for fall-out shelters" (Palmer, Colton and Kramer 2002, 956).[21] The shelters, the civil defense program, and the announcement to increase military defense spending were unlikely to improve battlefield outcomes *in* the current crisis itself – they are, as Jervis (1970, 22) put it, "gestures which would have little impact on the outcome of hostilities should they occur." But these gestures also involve reducible costs, which can make the threat to fight more credible despite the limited sunk costs and tied-hands costs.

In short, many signaling actions may involve more than one signaling mechanism. Sunk costs or tied-hands costs alone may not suffice to separate a signal of resolve from a bluff. However, a signal may also include installment costs or reducible costs that augment its credibility. IC and RC signaling may help explain why states can sometimes achieve credibility with actions that do not seem too costly (in terms of sunk costs or audience costs),[22] or with gestures that seem to have little direct im-

[20] The other two pieces of intelligence were Pentagon's putting U.S. forces on DEFCON 2 and an American journalist's leak to Soviet agents about the U.S. government on the verge of military intervention in Cuba (Fursenko and Naftali 1997, 262).

[21] These signals challenged Khrushchev's priors about U.S. resolve. As Gaddis (1997, 146) observed, the signals were "rather more than Khrushchev had expected. "Only a mad man can declare war today," he told John McCloy the next day."

[22] For instance, Trachtenberg's (2012) study of a dozen great power crises found "little evidence that the audience costs mechanism played a 'crucial' role in any of them. Indeed, it is hard to identify any case in which that mechanism played much of a role at all" (32).

plication on military success. Understanding that four costly signaling mechanisms exist will allow us to decompose a mixed case more precisely, and to uncover the additional sources of credibility latent in the case.

Four Experiments

The previous section explained the mechanics of how installment costs and reducible costs work. In this section, I turn to the empirical tests.[23] I will test the effects of installment costs and reducible costs on credibility, since credibility is the central concern in signaling.[24] For completeness, I will also test the credibility effects of sunk costs and tied-hands costs (Quek 2016; Yarhi-Milo, Renshon and Kertzer 2018; Kertzer, Renshon and Yarhi-Milo 2019).

I use experiments to test the credibility effects. A randomized experiment can isolate causal effects cleanly (Holland 1986; Morton and Williams 2010; Mutz 2011). Experimentation is especially useful for studying signaling, because real-world information environments are often noisy and saturated with a large number of different information variables, many of which are unobservable or unmeasurable.

I fielded four experiments with 1,707 American adults on 5-6 December 2017 on Amazon Mechanical Turk (AMT). Past studies have used AMT to test theories of signaling

[23] The IR literature on costly signaling has largely focused on tying hands and audience costs. See, for example, the special issue in *Security Studies*, Vol. 21, Issue 3. Recent research suggest that audience costs are relevant not only to democracies, but also in authoritarian contexts (Weeks 2008; Weiss 2013; Kurizaki and Whang 2015; Quek and Johnston 2018; Weiss and Dafoe 2019; Chen and Li 2020).

[24] Experiments on audience costs focus on two major areas: the measurement of audience costs in public opinion (e.g. Tomz 2007; Trager and Vavreck 2011; Levendusky and Horowitz 2012; Davies and Johns 2013; Levy et al. 2015; Kertzer and Brutger 2016; Quek 2017) and the effect of audience costs on credibility (Yarhi-Milo, Renshon and Kertzer 2018; Kertzer, Renshon and Yarhi-Milo 2019).

and reputation.[25] Similar to past experiments, a national sample is used. How humans make sense of costly signaling in general is an important question in itself. Answering this question is also a necessary first step to establish a baseline, so that we can subsequently compare how different sub-populations respond similarly or differently to costly signaling. A recent study found that a super-elite sample responded similarly to costly signaling as a national sample (Yarhi-Milo, Renshon and Kertzer 2018), which suggests that elites may process costly signals not too differently from normal human beings.

Costly signaling experiments in IR include Quek (2016), Yarhi-Milo, Kertzer and Renshon (2018), Kertzer, Renshon and Yarhi-Milo (2019), and Kertzer, Rathbun and Rathbun (2020). Three of these studied the credibility of a costly threat. Quek (2016) used game-theoretic experiments to show that signalers are more likely to sink costs when they are randomly assigned with high resolve, but receivers are equally likely to acquiesce with or without the sunk-cost signal. Yarhi-Milo, Kertzer and Renshon (2018) and Kertzer, Renshon and Yarhi-Milo (2019) used survey experiments with vignettes to show that sinking costs and tying hands can improve credibility.

The experiments here differ from previous work in investigating new signaling mechanisms. They also differ with a design that can identify how credibility changes when the costs of costly signals change. Past experiments on costly signaling used binary comparisons of a costly signal with a costless signal or no signal (Quek 2016; Yarhi-Milo, Kertzer and Renshon 2018; Kertzer, Renshon and Yarhi-Milo 2019), which pre-

[25] See, e.g., Brutger and Kertzer (2018); Kertzer, Rathbun and Rathbun (2020); Quek (2016); Renshon, Dafoe and Huth (2018). Mattes and Weeks's (2019) recent work on hawks and doves also used AMT. Coppock (2019) conducted 15 replication experiments on AMT and found similar results as the original experiments, corroborating earlier studies that tested the validity of AMT. For recent investigations of external validity, see Lupton (2019) and Tomz, Weeks and Yarhi-Milo (2020).

vent identification of how credibility perception evolves when the signal cost changes across levels of costliness.

Design

I designed four experiments to investigate the four signaling mechanisms. Each participant was randomly assigned to only one of the four, under a between-subjects design. All experiments began with the same scenario: "Two countries – Country X and Country Z – have a dispute over a piece of territory. Country X sends a Signal to Country Z threatening to go to war if Z does not withdraw from the territory."

Because our aim is to identify the effect of *each distinct mechanism* of signaling, the experiments will be confounded if we use real-world signals that conflate the different mechanisms (see the previous section). To cut through the identification problem, the experiments here are designed to approximate "pure signaling": the information environment is sterilized by design to eliminate noise, and a third-person perspective with generic countries is used to remove the effects of national identities, prior beliefs and reputation. While highly stylized, the pure-signaling baseline is the priority for a first investigation. Upon identifying how signal cost shapes credibility, we can test how credibility changes when the signal cost co-occurs with specific identities, relative capabilities and prior beliefs, where these effects can be identified separately from the effect of the costly signal.

As a formal definition will not be comprehensible to most people, each experiment distills a signaling mechanism into its most basic form, using simple language:

Experiment 1 (sunk-cost signal): In sending the signal, "X has paid a cost of

> [0/2/10]. The cost cannot be recovered regardless of whether X fulfills its threat."[26]
>
> **Experiment 2 (tied-hands signal):** In sending the signal, "X will pay a cost of 0 if X fulfills its threat, and a cost of [0/2/10] if X does NOT fulfill its threat."
>
> **Experiment 3 (IC signal):** In sending the signal, "X will pay a total cost of [0/2/10] in installments over time. Once the cost is paid it cannot be recovered regardless of whether X fulfills its threat."[27]
>
> **Experiment 4 (RC signal):** In sending the signal, "X has paid a cost of [0/2/10]. X gets back a value of [0/2/10] if X fulfills its threat, and 0 if X does NOT fulfill its threat."[28]

Respondents were randomly divided into three conditions under a between-subjects design: *costless* (cost = 0), *low-cost* (cost = 2), and *high-cost* (cost = 10).[29] Respondents interpreted these conditions on a scale from "costless" (0) to "extremely costly" (10). Thus, respondents responded to the same standardized scale, and were made aware of the spectrum of costly signals that the signaler could send, instead of evaluating a costly signal in isolation without a relative comparison of costliness. To hold constant

[26] Sunk-cost signaling involves an ex-ante fixed cost ("sunk") that carries a signaling effect independent of the substantive effect of the action. A pure sunk-cost signal does not in itself make it easier or harder for the signaler to fulfill the substance of the signal (threat or promise). Thus, Spence (1973, 364) assumed that "education [the sunk-cost signal] does not contribute to productivity", and Fearon (1997, 70) emphasized that sunk-cost signals "do not affect the relative value of fighting versus acquiescing in a challenge".

[27] Here the signal is open-ended as many IC signals have open-ended time horizons in the real world. Future work can randomize different time horizons and time inconsistencies to examine the effects of these variations.

[28] As this is a first test of the mechanism, I apply a sharp test with a scenario where ex-ante costs can be fully recovered (or lost) if the signaler follows through (or reneges). Future work can randomize different amounts of reducible costs to track how credibility changes across the different quantities.

[29] The chosen values allow for comparability with previous experiments, as we will see later. In particular, existing sunk-cost signaling experiments in IR that specified the cost of the signal used cost = 0 for the costless signal and cost = 2 for the costly signal, on a 0 to 10 scale (Quek 2016).

the signal type (threat), the experimental comparison is between costly and costless threats, rather than a costly threat and no threat.

After reading the scenario, respondents were asked whether they thought X was likely or unlikely to fulfill the threat if Z did not withdraw, yielding a credibility score on a seven-point scale from 0 (“very unlikely”) to 6 (“very likely”). Appendix 1 shows the wording of the experimental instrument. Appendix 2 shows the experimental groups are identical in their demographic characteristics upon randomization.

This experimental design enables us to answer three interrelated questions on signal effectiveness:

1. Do installment and reducible costs change the perception of credibility?
2. Are the credibility effects significant only when the costs are high?
3. Do people respond to installment and reducible costs in the same way?

Results

Because the treatments are randomized, confounding by omitted variables is ruled out by design. To identify treatment effects, we can simply look at the difference in means between the experimental groups. We will consider each experiment and signaling mechanism separately, before discussing them in conjunction.

Sunk-cost signal (Experiment 1). For respondents randomly assigned to the high-cost condition, the average credibility score is 4.27 on a seven-point scale (0 to 6) compared to 3.42 in the costless condition ($p < 0.001$, $n = 288$).[30] The credibility score

[30] All t-tests are two-tailed.

in the low-cost condition is 2.99 compared to 4.27 in the high-cost condition ($p < 0.001$, $n = 282$). These results show that the credibility of the signal sharpens when it carries high sunk costs.

Experiment 1 also found that a costless threat can still be slightly credible in itself. This result is consistent with what previous experiments found on cheap talk (Tingley and Walter 2011) and sinking costs (Quek 2016). In the existing literature, the only other research that explicitly specified a threat of cost = 0 are the sunk-cost signaling experiments in Quek (2016). Using incentivized signaling games over the Internet and in the laboratory, Quek (2016) found that among those who received a costless threat, 51% (in the Internet experiment with an AMT sample) and 56% (in the laboratory experiment with an MIT sample) were successfully deterred by the threat. Despite a different instrument and setting, the outcome from Experiment 1 is remarkably consistent: 54% of respondents in the costless condition found the threat credible.

Tied-hands signal (Experiment 2). The average credibility score in the high-cost condition is 4.91 compared to 4.45 in the costless condition ($p = 0.035$, $n = 275$). The credibility score in the low-cost condition (4.18) is also significantly lower than in the high-cost condition (4.91) ($p = 0.001$, $n = 298$). These results show that tying hands can improve credibility.

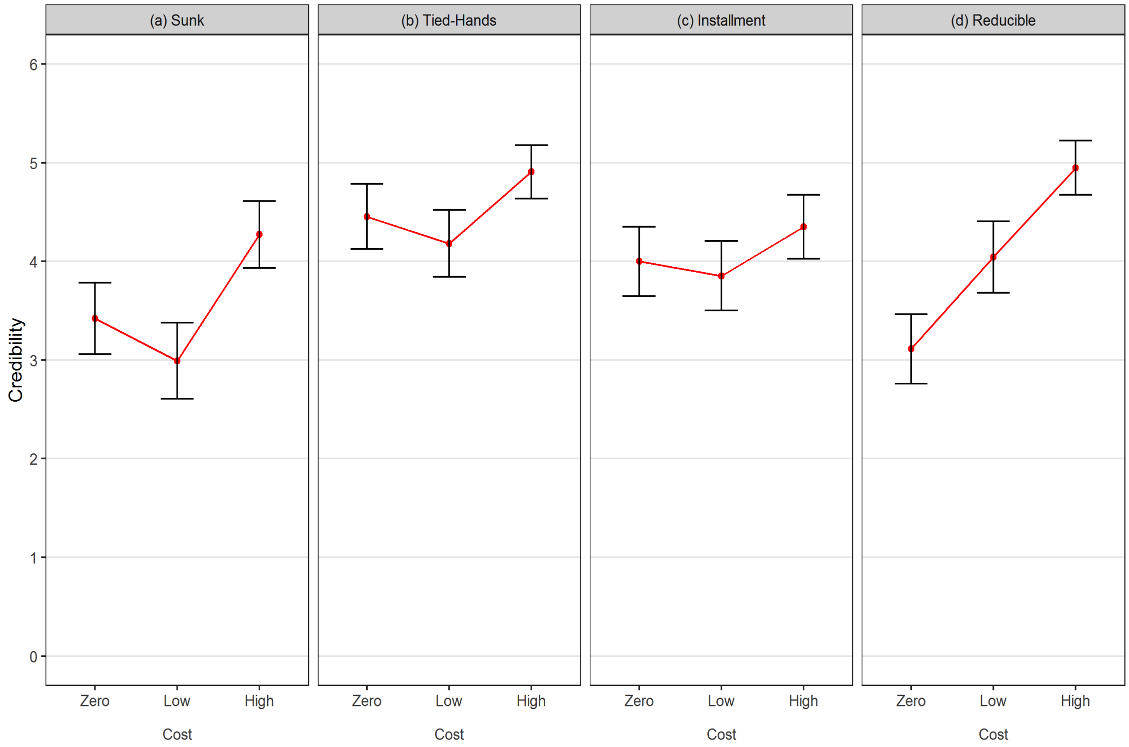
The results also suggest that the credibility effect is non-monotonic, whereby there is no significant difference in credibility between costless (4.45) and low-cost signals (4.18) ($p = 0.257$, $n = 287$). In fact, although their effects on credibility differ, tying hands and sinking costs generate credibility curves with a similar kinked pattern, whereby credibility does not increase across costless and low-cost levels, but spikes at the high-cost level. Similar to the case for tying hands, there is no significant difference in credibility between the costless (3.42) and low-cost (2.99) conditions for sinking costs (p = 0.111, n = 284).[31] The finding is consistent with the results from two earlier sunk-cost signaling experiments (Quek 2016), which compared a costless signal (cost = 0) with a (low) sunk-cost signal (cost = 2) that generated a unique separating equilibrium.

IC signal (Experiment 3). The average credibility score is 4.35 in the high-cost condition compared to 4.00 in the costless condition, with no statistically significant difference (p = 0.147, n = 286). The difference between the credibility scores in the low-cost condition (3.85) and the high-cost condition (4.35) is marginally significant (p = 0.040, n = 279). The lack of sharpness in the statistical evidence suggests the credibility of installment costs is discounted to some extent.

Departing from previous signaling experiments that compared one costly signal with either a costless signal or no signal, these experiments measured credibility effects at three different cost levels. Figure 1 shows how the credibility of each type of costly signal changes across different levels of costliness.

[31] See footnote 32.

Figure 1: Perceived Credibility in the (a) Sunk-Cost, (b) Tied-Hands, (c) Installment-Cost and (d) Reducible-Cost Signaling Experiments



Across the three mechanisms (Figures 1a–1c), one interesting commonality stands out: costless signals are as credible as low-cost signals.[32] If the receiver knows the sender is deliberately choosing to incur less costs than more, the costly signal would not improve perceived credibility, and low-cost signals are no better than costless signals. This result supports Fearon's (1997) conjecture that sending a smaller, half-hearted signal may demonstrate that one's resolve is insufficient to support the larger signal, and thus reveals a *lack* of resolve.[33] Some leaders in the past seem to share a

[32] IC signal: 4.00 (costless) versus 3.85 (low-cost) (p = 0.564, n = 281). Tied-hands signal: 4.45 (costless) versus 4.18 (low-cost) (p = 0.257, n = 287). Sunk-cost signal: 3.42 (costless) and 2.99 (low-cost) (p = 0.111, n = 284). There is an interesting pattern whereby credibility is lower in absolute (but not statistically significant) terms in the low-cost condition compared to the costless condition. Future work may investigate if this is a coincidence or a systematic phenomenon. See also footnote 34.

[33] I thank a reviewer for highlighting this.

similar conjecture, though the hypothesis has not been systematically tested in previous work to my knowledge.[34]

Credibility of RC signals (Experiment 4). In contrast, Figure 1(d) shows that the credibility curve for the RC signal has a positive slope. First, the average credibility score is 3.11 in the costless condition compared to 4.04 in the low-cost condition ($p < 0.001$, $n = 286$). Further, there is also a sharp difference between the high-cost (4.95) and costless (3.11) conditions ($p < 0.001$, $n = 291$). Finally, the RC signal is also more credible in the high-cost condition than in the low-cost condition, with a credibility score of 4.95 in the former compared to 4.04 in the latter ($p < 0.001$, $n = 277$).

Low-cost signals can significantly improve credibility under RC signaling, unlike in the other mechanisms. Future research may test the different reasons for why the mechanism is effective for both low and high-cost signals. One conjecture is the prospect of the sender recovering the costs (whether low or high) from fulfilling the threat makes it reasonable to believe the threat. The costs are already paid, but can be recovered if the sender fulfills the threat, regardless of whether ex-ante costs were low or high. This cost-recovery paradigm might be particularly salient to receivers.[35] Future work may test this conjecture against alternative explanations.

To conclude, we turn back to the big picture. I designed the experiments to answer three questions. We close by summarizing the answers:

1. Reducible costs have a strong and significant effect on credibility. However, the credibility effect of installment costs is more subdued, which implies a time

[34] Alfred Vagts, for example, "warns, cogently citing Disraeli and Churchill on his side, against the demonstration that falls short of the mark and signals the opposite of stern intent" (Schelling 1966, 239).

[35] See the discussion in the next section.

discount in credibility.

2. RC signals are credible at both low and high-cost levels. For IC signals, there is no change in credibility between costless and low or high-cost signals, despite a marginally significant difference between low and high-cost signals.

3. The evidence suggests that people do not respond to installment and reducible costs in the same way. The credibility curve for the RC signal has a positive slope, whereas the credibility curve for the IC signal is kinked.

This is a first investigation into whether the new signaling mechanisms are effective in shaping assessments of credibility. While some tentative interpretations are offered, *why* the patterns of credibility differ the way they do is a deeper question that would require more research.

Discussion

This paper proposes the existence of four mechanisms of costly signaling, develops the new mechanisms of installment costs and reducible costs, and tests their effectiveness in shaping credibility. Existing research assumes that only two costly signaling mechanisms exist. I show there are four mechanisms logically distinct and equally general as each other, with a unified framework that brings together what would otherwise be unrelated mechanisms. Existing research has not examined the implications of installment and reducible costs on signaling and credibility. I develop these mechanisms and show how they differ from sunk and tied-hands costs. Existing research has not identified empirically how credibility effects change when the costs of costly signals change across levels of costliness. I identify the credibility effects that emerge from the four signaling mechanisms, both new and old.

For scholars studying signaling, it is useful to recognize that a signaling action may invoke more than one signaling mechanism, and then decompose the different sources of credibility in the signal. Does the signal have different cost components? For each component, in which of the four cells of the 2x2 matrix (Table 1) does it fall? If IC signaling is involved, what are the time-horizon effects on credibility? If RC signaling is involved, are the signaling costs reducible computationally, substantively, or indirectly if the signaler fulfills the threat? The four signaling mechanisms provide ideal-type differentiations that can help us organize and extend our understanding of empirical cases. In cases where the sunk and tied-hands costs appear limited, scholars can check whether installment and reducible costs are present. By breaking down the signal into different cost components and analyzing the mechanism in each component, the different sources of credibility latent in the signal can be identified precisely.

Many unresolved questions remain. Because the mechanisms are identified here for the first time, there is still much about them that we do not know. Signaling theory is complex; its implications sometimes subtle. The purpose of this paper is not to establish final conclusions, but to present possibilities and implications for future research.

The new mechanisms have many potential implications. It is useful to note that while costly signaling was first formalized in economics, the idea has been adopted across the social sciences and applied in different ways. Some researchers apply costly signaling in the rationalist tradition, while others explore instead its behavioral and psychological aspects (e.g. Jervis, Lebow and Stein 1985; Herrmann and Fischerkeller 1995; O'Neill 1999; Larson 2000; Hall and Yarhi-Milo 2012; Quek 2016; Yarhi-Milo, Kertzer and Renshon 2018; Acharya and Grillo 2019; Kertzer, Rathbun and Rathbun

2020). The new mechanisms are relevant not only in the rationalist tradition, but also for researchers in the psychological and behavioral tradition. We shall consider both the *rationalist* and the *behavioral* significance of the new signaling mechanisms.

Reducing costs. These signals double-dip into credibility through the effect of the initial costs incurred and the effect of the reduction of these costs by fulfilling the threat. Their impact on credibility should be stronger than the impact from a single determinant of credibility. From a rationalist perspective, credibility can be enhanced, but depending on the game variant, there may or may not be a tradeoff similar to that found in IR models of tied-hands signaling, where the risk of war also increases (Fearon 1997; Slantchev 2005).[36] Future research may explore how the formal implications of this mechanism change across different models. As an example of one way by which the strategic setting might shift, Appendix 3 illustrates how the game and payoffs in Fearon (1997) change when we move from sunk-cost or tied-hands signaling to RC signaling.[37]

From a behavioral perspective, prospect theory and the endowment effect would predict that signalers place greater weight on recovering losses and take greater risks to do so (Kahneman and Tversky 1979; Levy 1992; McDermott 2004; Butler 2007). Even if the absolute value of the costs and gains are the same, recovering costs has a different psychological salience compared to making gains. The cost-recovery paradigm implies signalers should be more likely to fulfill their threats in order to offset the

[36] Fearon (1997) suggests tying hands involves a greater risk of war than sinking costs. Slantchev's (2005) model suggests war may become preferable for both sides under uncertainty.

[37] For example: (1) Unlike military threats (Slantchev 2005), which affect the war payoffs of both Defender (D) and Challenger (C), reducible costs directly affect only D's war payoff but not C's. (2) Unlike tying hands (Fearon 1997), the cost of fighting can be shaped by signal cost. Thus, the risk of war and equilibrium level of signal cost can change as a consequence. (3) In the full recovery condition, signaling is costless if C challenges and D fights, but costly if C does not challenge. In tying hands, signaling is costless if C does not challenge, but costly if C challenges and D does not fight.

costs incurred. RC signaling may thus not only rationally reveal the signaler's ex-ante resolve, but also make her psychologically more resolved ex-post. This behavioral effect can reinforce the rationalist effect on the likelihood of the signaler fulfilling the threat, amplifying further the salience of the RC signal.

Installing costs. These signals differ from sunk-cost signals because they have a time horizon. The time horizon is a vibrant subject of research in psychology and economics, generating many implications for IC signaling. The classic rationalist implication, following the discounted utility paradigm (Samuelson 1937), is that an IC signal (with costs incurred over a future time horizon) suffers a *time discount*, whereas a sunk-cost signal (with costs incurred in the present) does not. As cost magnitude affects signal credibility, the IC signal will be discounted in credibility depending on the degree of the time discount. Another rationalist implication is that the IC signal can involve a *commitment discount* to credibility, if there is a probability that the signaler may not pay the installment costs in the future. Obviously, because the costs are already sunk for sunk-cost signaling, time and commitment discounts cannot apply.

From a rationalist perspective, it is therefore apparent that installing costs differs substantively from sinking costs. However, the specific differences can take many forms. Scholars have modeled the economics of time horizons in many variants,[38] and installment costs with and without commitment problems can be formulated in different ways. The combination of the economics of time horizons and time inconsistencies will generate further possibilities. This is a complex but very interesting area for

[38] Early rationalist treatments of the time horizon in IR include Axelrod (1984) and Oye (1985). Toft (2006) highlighted that differences in time horizon between states can constitute a rationalist explanation for war. Kertzer (2017) argued that individual dispositional differences in time preferences can explain differences in resolve. Haynes (2019) developed a model in which the receiver's uncertainty about the sender's time horizon influences reassurance credibility and cooperation possibilities.

research. Appendix 3 offers some suggestions on how to think about installing costs in its basic form. The dynamic nature of installing costs may be especially relevant for general deterrence.

From a behavioral perspective, the story becomes more nuanced. Hyperbolic discounting theory in behavioral economics suggests that signalers and receivers would not apply a linear time discount (Frederick, Loewenstein and O'Donoghue 2002). Instead, future cost and value fluctuate over the time horizon, falling quickly for earlier periods and slowly for later periods. Since the credibility of costly signaling depends not on real costs but *perceived* costs, the credibility of an IC signal should also fluctuate depending on the shape and duration of the time horizon. Psychological research on temporal construals offers a different behavioral interpretation (Trope and Liberman 2003; Streich and Levy 2007; Krebs and Rapport 2012). This research suggests that people tend to think about near-term phenomena in concrete terms and long-term phenomena in abstract terms. Abstract thinking produces more optimistic perceptions because it can transcend uncomfortable details. One implication is that optimism – at *both* the sender and receiver ends – is more likely when installment costs are spread over a longer time horizon. Senders may be more optimistic that their signals are credible, whereas receivers may be more optimistic that their priors (whether true or false) are correct. The impact on credibility will thus be influenced by the initial priors held by the receivers.

Our discussion suggests there are many open questions that require investigation. Future work may find it useful to think about both rationalist and behavioral implications, rather than one in isolation of the other. This can help formulate more robust expectations based on whether the implications move in similar directions. It also opens up new hypotheses for each signaling mechanism. Which implications are right under what circumstances – how they combine with one another and to what degree – are empirical questions that need more research. Although addressing them involves hard work, it will clarify our understanding of whether and when costly signaling matters, and why.

References

Acharya, Avidit, and Edoardo Grillo. 2019. "A Behavioral Foundation for Audience Costs." *Quarterly Journal of Political Science* 14(2): 159–190.

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

Altman, Dan. 2017. "By Fait Accompli, Not Coercion: How States Wrest Territory from their Adversaries." *International Studies Quarterly* 61(4): 881–891.

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

Blankenship, Brian. 2020. "Promises under Pressure: Statements of Reassurance in US Alliances." *International Studies Quarterly.* Forthcoming.

Bliege Bird, Rebecca, and Eric Smith. 2005. "Signaling Theory, Strategic Interaction, and Symbolic Capital." *Current Anthropology* 46(2): 221–248.

Brutger, Ryan, and Joshua Kertzer. 2018. "A Dispositional Theory of Reputation Costs." *International Organization* 72(3): 693–724.

Butler, Christopher. 2007. "Prospect Theory and Coercive Bargaining." *Journal of Conflict Resolution* 51(2): 227–250.

Carson, Austin, and Keren Yarhi-Milo. 2017. "Covert Communication: The Intelligibility and Credibility of Signaling in Secret." *Security Studies* 26(1): 124–156.

Chan, Steve. 2012. "Money Talks: International Credit/Debt as Credible Commitment." *Journal of East Asian Affairs* 26(1): 77–103.

Coppock, Alexander. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7(3): 613–628.

Dafoe, Allan, Jonathan Renshon, and Paul Huth. 2014. "Reputation and Status as Motives for War." *Annual Review of Political Science* 17: 371–393.

Dai, Xinyuan, Duncan Snidal, and Michael Sampson. 2017. "International Cooperation Theory and International Institutions." *Oxford Research Encyclopedia: International Studies*. https://doi.org/10.1093/acrefore/9780190846626.013.93.

Davies, Graeme, and Robert Johns. 2013. "Audience Costs among the British Public: The Impact of Escalation, Crisis Type, and Prime Ministerial Rhetoric." *International Studies Quarterly* 57(4): 725–37.

Fang, Songying. 2008. "The Informational Role of International Institutions and Domestic Politics." *American Journal of Political Science* 52(2): 304–321.

Fang, Songying, Jesse Johnson, and Brett Ashley Leeds. 2014. "To Concede or to Resist? The Restraining Effect of Military Alliances." *International Organization* 68(4): 775–809.

Fearon, James. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88(3): 577–592.

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

———. 1997. "Signaling Foreign Policy Interests: Tying Hands Versus Sinking Costs." *Journal of Conflict Resolution* 41(1): 68–90.

Frederick, Shane, George Loewenstein and Ted O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review." *Journal of Economic Literature* 40(2): 351–401.

Fuhrmann, Matthew, and Todd Sechser. 2014. "Signaling Alliance Commitments: Hand-Tying and Sunk Costs in Extended Nuclear Deterrence." *American Journal of Political Science* 58(4): 919–935.

Fursenko, Aleksandr, and Timothy Naftali. 1997. *One Hell of a Gamble: Khrushchev, Castro, and Kennedy, 1958-1964.* New York: W. W. Norton.

Gaddis, John Lewis. 1997. *We Now Know: Rethinking Cold War History*. New York: Oxford University Press.

Gambetta, Diego. 2009. "Signaling." In *Oxford Handbook of Analytical Sociology*, eds. Peter Hedstrom and Peter Bearman. New York: Oxford University Press, 168–194.

Gartzke, Erik, and Oliver Westerwinter. 2016. "The Complex Structure of Commercial Peace Contrasting Trade Interdependence, Asymmetry, and Multipolarity." *Journal of Peace Research* 53(3): 325–343.

Gartzke, Erik, Shannon Carcelli, J. Andres Gannon, and Jiakun Jack Zhang. 2017. "Signaling in Foreign Policy." *Oxford Research Encyclopedia of Politics*. https://doi.org/10.1093/acrefore/9780190228637.013.481.

Gartzke, Erik, and Jon Lindsay. 2017. "Thermonuclear Cyberwar." *Journal of Cybersecurity* 3(1): 37–48.

Garver, John. 2006. "China's Decision for War with India in 1962." In *New Directions in the Study of China's Foreign Policy*, eds. Alastair Iain Johnston and Robert Ross. Stanford: Stanford University Press, 86–130.

Hall, Todd, and Keren Yarhi-Milo. 2012. "The Personal Touch: Leaders' Impressions, Costly Signaling, and Assessments of Sincerity in International Affairs." *International Studies Quarterly* 56(3): 560–573.

Hartzell, Caroline, and Matthew Hoddie. 2007. *Crafting Peace: Power-sharing Institutions and the Negotiated Settlement of Civil Wars.* University Park: Penn State University Press.

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

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

Herrmann, Richard, and Michael Fischerkeller. 1995. "Beyond the Enemy Image and Spiral Model: Cognitive-Strategic Research after the Cold War." *International Organization* 49(3): 415–450.

Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945–960.

Jervis, Robert. 1970. *The Logic of Images in International Relations*. Princeton: Princeton University Press.

Jervis, Robert, Richard Ned Lebow, and Jane G. Stein, eds. 1985. *Psychology and Deterrence.* Baltimore: Johns Hopkins University Press.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 42(2): 263–92.

Kennedy, John F. 1962. Radio and Television Address to the American People on the Soviet Arms Build-Up in Cuba, 22 October 1962. John F. Kennedy Presidential Library. https://www.jfklibrary.org/asset-viewer/archives/JFKWHA/1962/JFKWHA-142-001/JFKWHA-142-001.

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

Kertzer, Joshua, and Ryan Brutger. 2016. "Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory." *American Journal of Political Science* 60(1): 234–49.

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

Kertzer, Joshua, Jonathan Renshon, and Keren Yarhi-Milo. 2019. "How Do Observers Assess Resolve?" *British Journal of Political Science*. Forthcoming.

Krebs, Ronald, and Aaron Rapport. 2012. "International Relations and the Psychology of Time Horizons." *International Studies Quarterly* 56(3): 530–543.

Kurizaki, Shuhei, and Taehee Whang. 2015. "Detecting Audience Costs in International Disputes." *International Organization* 69(4): 949–980.

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

Kydd, Andrew, and Roseanne McManus. 2017. "Threats and Assurances in Crisis Bargaining." *Journal of Conflict Resolution* 61(2): 325–348.

Larson, Deborah Welch. 2000. *Anatomy of Mistrust: US-Soviet Relations During the Cold War*. Ithaca: Cornell University Press.

Levendusky, Matthew, and Michael Horowitz. 2012. "When Backing Down Is the Right Decision: Partisanship, New Information, and Audience Costs." *Journal of Politics* 74(2): 323–338.

Levy, Jack. 1992. "An Introduction to Prospect Theory." *Political Psychology* 13(2): 171–186.

Levy, Jack, Michael McKoy, Paul Poast, and Geoffrey Wallace. 2015. "Backing Out or Backing In? Commitment and Consistency in Audience Costs Theory." *American Journal of Political Science* 59(4): 988–1001.

Li, Xiaojun, and Dingding Chen. 2020. "Public Opinion, International Reputation, and Audience Costs in an Authoritarian Regime." *Conflict Management and Peace Science.* Forthcoming.

Lupton, Danielle. 2018. "Signaling Resolve: Leaders, Reputations, and the Importance of Early Interactions." *International Interactions* 44(1): 59–87.

——— . 2019. "The External Validity of College Student Subject Pools in Experimental Research: A Cross-Sample Comparison of Treatment Effect Heterogeneity." *Political Analysis* 27(1): 90–97.

McDermott, Rose. 2004. "Prospect Theory in Political Science: Gains and Losses from the First Decade." *Political Psychology* 25(2): 289–312.

McManus, Roseanne. 2017. "The Impact of Context on the Ability of Leaders to Signal Resolve." *International Interactions* 43(3): 453–479.

——— . 2018. "Making it Personal: The Role of Leader-Specific Signals in Extended Deterrence." *Journal of Politics* 80(3): 982–995.

McManus, Roseanne, and Keren Yarhi-Milo. 2017. "The Logic of 'Offstage' Signaling: Domestic Politics, Regime Type, and Major Power-Protege Relations." *International Organization* 71(4): 701–733.

Mattes, Michaela, and Burcu Savun. 2009. "Fostering Peace after Civil War: Commitment Problems and Agreement Design." *International Studies Quarterly* 53(3): 737–759.

Mattes, Michaela, and Greg Vonnahme. 2010. "Contracting for Peace: Do Nonaggression Pacts Reduce Conflict?" *Journal of Politics* 72(4): 925–938.

Mattes, Michaela, and Jessica Weeks. 2019. "Hawks, Doves, and Peace: An Experimental Approach." *American Journal of Political Science* 63(1): 53–66.

Morton, Rebecca, and Kenneth Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab.* New York: Cambridge University Press.

Mutz, Diana. 2011. *Population-Based Survey Experiments.* Princeton: Princeton University Press.

Naftali, Timothy, and Philip Zelikow. 2001. *The Presidential Recordings: John F. Kennedy, Volume 2*. New York: W. W. Norton.

Nobel Foundation. 2001. The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2001: Information for the Public. https://www.nobelprize.org/prizes/economic-sciences/2001/popular-information/.

Oye, Kenneth. 1985. "Explaining Cooperation under Anarchy: Hypotheses and Strategies" *World Politics* 38(1): 1–24.

O'Neill, Barry. 1990. "The Intermediate Nuclear Force Missiles: An Analysis of Coupling and Reassurance." *International Interactions* 15: 345–363.

———. 1999. *Symbols, Honor, and War*. Ann Arbor: University of Michigan Press.

Palmer, R. R., Joel Colton and Lloyd Kramer. 2002. *A History of the Modern World*. New York: A. A. Knopf.

Quek, Kai. 2013. "Rationalist Causes of War: Mechanisms, Experiments, and East Asian Wars." PhD Dissertation, Massachusetts Institute of Technology.

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

———. 2017. "Type II Audience Costs." *Journal of Politics* 79(4): 1438–1443.

Quek, Kai, and Alastair Iain Johnston. 2018. "Can China Back Down? Crisis De-escalation in the Shadow of Popular Opposition." *International Security* 42(3): 7–36.

Quinn, Colin. 2019. "Costly Signaling Theory in Archaeology." In *Handbook of Evolutionary Research in Archaeology,* ed. Anna Marie Prentiss. Cham: Springer, 275–294.

Reiter, Dan. 2009. *How Wars End.* Princeton: Princeton University Press.

Renshon, Jonathan, Allan Dafoe, and Paul Huth. 2018. "Leader Influence and Reputation Formation in World Politics." *American Journal of Political Science* 62(2): 325–339.

Riley, John. 2001. "Silver Signals: Twenty-Five Years of Screening and Signaling." *Journal of Economic Literature* 39: 432–478.

Rothschild, Michael, and Joseph Stiglitz. 1976. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Quarterly Journal of Economics* 90(4): 629–649.

Sartori, Anne. 2005. *Deterrence by Diplomacy.* Princeton: Princeton University Press.

Schelling, Thomas. 1966. *Arms and Influence*. New Haven: Yale University Press.

Slantchev, Branislav. 2005. "Military Coercion in Interstate Crises." *American Political Science Review* 99(4): 533–547.

———. 2011. *Military Threats: The Costs of Coercion and the Price of Peace*. New York: Cambridge University Press.

Simmons, Beth. 2000. "International Law and State Behavior: Commitment and Compliance in International Monetary Affairs." *American Political Science Review* 94(4): 819–835.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics* 87(3): 355–374.

———. 1976. "Informational Aspects of Market Structure: An Introduction." *Quarterly Journal of Economics* 90(4): 591–597.

Streich, Philip, and Jack Levy. 2007. "Time Horizons, Discounting, and Intertemporal Choice." *Journal of Conflict Resolution* 51(2): 199–226.

Tago, Atsushi, and Maki Ikeda. 2015. "An 'A' for Effort: Experimental Evidence on UN Security Council Engagement and Support for US Military Action in Japan." *British Journal of Political Science* 45(2): 391–410.

Tarar, Ahmer. 2016. "A Strategic Logic of the Military Fait Accompli." *International Studies Quarterly* 60(4): 742–752.

Tingley, Dustin, and Barbara Walter. 2011. "Can Cheap Talk Deter? An Experimental Analysis." *Journal of Conflict Resolution* 55(6): 996–1020.

Toft, Monica Duffy. 2006. "Issue Indivisibility and Time Horizons as Rationalist Explanations for War." *Security Studies* 15(1): 34–69.

Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61(4): 821–840.

Tomz, Michael, Jessica Weeks, and Keren Yarhi-Milo. 2020. "Public Opinion and Decisions about Military Force in Democracies." *International Organization* 74(1): 119–143.

Trachtenberg, Marc. 2012. "Audience Costs: An Historical Analysis," *Security Studies* 21(1): 3–42.

Trager, Robert. 2016. "The Diplomacy of War and Peace." *Annual Review of Political Science* 19: 205–228.

Trager, Robert, and Lynn Vavreck. 2011. "The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party." *American Journal of Political Science* 55(3): 526–545.

Trope, Yaacov, and Nira Liberman. 2003. "Temporal Construal." *Psychological Review* 110(3): 403–421.

United States Congress House Committee on Armed Services. 1963. *Hearings 1963-64, No. 10-20, Pt. 2, Vol. 1.* U.S. Government Printing Office.

Voeten, Eric. 2005. "The Political Origins of the UN Security Council's Ability to Legitimize the Use of Force." *International Organization* 59(3): 527–557.

Weeks, Jessica. 2008. "Autocratic Audience Costs: Regime Type and Signaling Resolve." *International Organization* 62(1): 35–64.

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

Weiss, Jessica Chen, and Allan Dafoe. 2019. "Authoritarian Audiences, Rhetoric, and Propaganda in International Crises: Evidence from China." *International Studies Quarterly* 63(4): 963–973.

Wu, Cathy Xuanxuan, and Scott Wolford. 2018. "Leaders, States, and Reputations." *Journal of Conflict Resolution* 62(10): 2087–2117.

Yarhi-Milo, Keren, Joshua Kertzer, and Jonathan Renshon. 2018. "Tying Hands, Sinking Costs, and Leader Attributes." *Journal of Conflict Resolution* 62(10): 2150–2179.