You Get What You Give: A Model of Nuclear Reversal

William Spaniel*

September 14, 2015

Abstract

Although research indicates that verification is critical for successful nuclear weapons agreements, some scholars and policymakers are skeptical that transparency can ever be achieved. This paper asks whether states can reach credible agreements without verification. Beyond monitoring institutions, many pacts require destruction of program infrastructure, which adds to the cost of future violations. I show that such cost increases in the form of moderate nuclear reversals incentivize opponents to cut deals that ultimately leave all parties better off. Arms treaties therefore primarily benefit potential proliferators, not their opposition. I apply these findings to help explain features of the recent Joint Comprehensive Plan of Action ("Iran Deal").

*Center for International Security and Cooperation, E212 Encina Hall, Stanford University, Stanford, CA 94305 (williamspaniel@gmail.com, http://williamspaniel.com).

1 Introduction

Scholars of international relations see verification as a major hurdle to reaching arms agreements.[1] Judging by verification's centrality in various treaties, the policymaking community agrees. For example, the recent Joint Comprehensive Plan of Action (JCPOA, or "Iran Deal") between the international community and Iran has many provisions designed to increase the Iranian nuclear program's transparency. This includes adherence to the Nuclear Non-Proliferation Treaty's Additional Protocol—itself a key information-providing treaty—a list of existing nuclear sites, International Atomic Energy Agency inspection of those sites, a three-fold increase in the number of inspectors in the country, and round-the-clock monitoring of specific nuclear materials.

However, questions of compliance remain. Opponents cannot observe everything that occurs inside a target state. Thus, even if they know that potential proliferators are not using declared sites to build a bomb, they must still worry that *unknown* sites exist that are dedicated to that task. These fears prompted the 2003 Iraq War (Debs and Monteiro 2014) and also contributed to a breakdown of the Agreed Framework between North Korea and the United States in the early 2000s (Beal 2005, 121). In the extreme, pessimists might wonder why states bother with weapons inspections at all.

That said, at the very least, weapons inspections act as a nuisance, forcing potential proliferators to develop their weapons less efficiently by having to play cat-and-mouse games with inspectors or by using secondary locations. Such inconvenience provisions are common in weapons agreements. For instance, the JCPOA requires Iran to divest its nuclear infrastructure in part by relinquishing control over many of its centrifuges. Steps like this do not make it impossible for potential proliferators to resurrect their programs. They do, however, increase the cost of reactivation.

It is puzzling that states would willingly reverse course like this. After all, devaluing an outside option in this manner can only hurt a state's outcome should it ultimately proliferate. Meanwhile, rivals would seem to have less incentive to offer generous deals if the potential proliferator seeks to strike a bargain instead.[2] And even if an agreement is possible in theory, rivals appear to face a commitment problem in that they could decrease concessions immediately following the reversal.

[1] See Meyer 1984, Krass 1985, Adelman 1990; Dunn 1990; Gallagher 2003.

[2] To wit, in crisis bargaining with incomplete information, a type with a higher war payoff than a second type must receive at least as great of an overall payoff (Banks 1990; Fey and Ramsay 2011).

This leads to two related questions. First, why do potential proliferators agree to these provisions. And second, if weapons inspections are so ineffective at revealing information, why does the international community continue to press for them? I argue additional burdens to proliferation—whether caused by weapons inspections or divestment—allow the parties to reach agreements that would have been impossible otherwise. This is because opponents, in the absence of a deal and unable to effectively monitor the proliferation process, must waste resources on the "stick" of preventive war. Oftentimes, such wars are mistakes because the potential proliferator opted not to develop the weapons at all. Alternatively, an agreement with the potential proliferator requires giving concessions commensurate with the quality of its nuclear option. Intuitively, rivals pay the "carrots" when these concessions are cheaper than using the stick. One determinant of the optimal choice is the cost to proliferate; the higher it is, the more likely the rival is to offer concessions. The potential proliferator therefore voluntarily undergoes nuclear reversal to induce an agreement.

For clarity, the nuclear reversal in my model—and in most empirical examples—does not terminate the proliferator's ability to build a bomb. Indeed, potential proliferators would never permit such draconian measures, which would lead to bad outcomes for all parties. Instead, the goal is to undergo just enough of a reversal to incentivize negotiations. Although potential proliferators can cheat on these reversals by developing weapons anyway, the negotiated settlements give them incentive not to. Accordingly, as uncomfortable as it may seem, rivals can live with the uncertainty that potential proliferators could secretly nuclearize.

To elucidate the above logic, I develop a model of nuclear reversal, uncertainty over proliferation behaviors, bargaining over weapons, and preventive war. Models have addressed the last three of these components, but never at the same time. However, they have uncovered critical insights that play a role in the mechanism I identify below.[3] In contrast, no model has featured a decision to divest in nuclear infrastructure. Yet I show that all four components are intimately tied together. Thus, the complete interaction is necessary to model to obtain a full understanding of the politics of proliferation and nuclear reversal.

[3]Fearon 1995, Powell 1999, Powell 2006, Chadefaux 2011, and Reed, Wolford, and Arena 2015 contain models of shifting power. Only Debs and Monteiro 2014 has a model of shifting power where opponents cannot directly observe decisions to build weapons. Only Spaniel 2105 has a model of shifting power and bargaining over weapons.

The model reveals four key empirical implications. First, reversals are most likely when the potential proliferator's route to a bomb is cheap and the extent of the possible power shift is greatest—in other words, the conditions seemingly most likely to result in nuclearization. Second, reversals are intentionally limited because further divestment and more onerous weapons inspections decrease the amount of concessions the opponent needs to offer. Thus, limited reversals are *not* proof that the potential proliferator will secretly nuclearize. Third, as the cost of preventive war decreases, the extent of a nuclear reversal increases. This is because the potential proliferator needs to offer more to entice the opponent to propose a deal. Finally, despite having to avoid preventive war and facing potentially substantial costs to nuclearize, the potential proliferator can ultimately capture all of the surplus created by reaching a deal.

This paper contains four additional sections. The next section gives a richer description of non-proliferation agreements, further motivating the four critical components of the model. The third section then describes the model and solves for its equilibria. The fourth section investigates the empirical implications of the model that were outlined in the previous paragraph. A brief conclusion finishes the paper with some key policy implications.

2 Features of Inspection and Reversal Regimes

To date, scholarly research on weapons inspections and arms agreement verification have focused on informational aspects. The earliest treatments build on the iterated prisoner's dilemma mechanism (Axelrod 1984; Downs, Rocke, and Siverson 1986; Ikle 1961, 214-215). Arms races often have a payoff structure like a prisoner's dilemma, where each state individually prefers building but both are worse off with that outcome than had no one increased their arms allotments. Using the shadow of the future, trigger strategies that punish deviations to aggressive actions can facilitate mutual cooperation. However, players must be able to observe their opponents' past actions to effectively implement trigger strategies. This type of information revelation was thus unsurprisingly a key part of arms agreements between the United States and the Soviet Union (Dunn 1990; Gallagher 2003).

The iterated prisoner's dilemma mechanism has two major limitations, though. First, the mechanism only applies to situations where both parties might increase ar-

	Prevent	$\sim$ Prevent
Build	Preventive War; Wasted Costs	Successful Power Shift
$\sim$ Build	Preventive War	Status Quo

Figure 1: A simultaneous move game between a potential proliferator (row player) and its rival (column player).

maments. Yet there are many asymmetric examples where one state tries to convince a second not to build weapons. Second, prisoner's dilemmas ignore richer strategic environments that feature both bargaining and the possibility of preventive war.

Fortunately, Debs and Monteiro (2014) address both these points. Their interaction features a subgame that appears in the model I develop below, so it is worth understanding their underlying mechanism. Suppose (1) a state cannot observe its rival's decision to proliferate, (2) the potential nuclear state finds weapons worth the investment, and (3) the opponent prefers launching preventive war to allowing a successful power shift transpire. Then the states are effectively playing the simultaneous move game substantively described in Figure 1.

Debs and Monteiro's key insight is that optimal strategies allow *all* outcomes to occurr with positive probability. To see why, note that preventive war and building cannot occur with certainty because the potential proliferator would want to deviate to not building to save on the wasted costs. But a mistaken preventive war cannot occur with certainty either because the rival would want to switch to not preventing. Maintaining the status quo with certainty is equally unsustainable because the potential proliferator could sneak in a successful power shift. And a successful power shift cannot be the guaranteed result because the rival would want to switch back to preventing. Thus, each side must mix to stop the other one from exploiting them.

However, the result of this mixing is inefficient. After all, mistaken preventive war occurs with positive probability. Debs and Monteiro show that one solution to this problem is to increase the rival's ability to observe the potential proliferator's move—this deters the opponent from building and minimizes the chances of a preventive war in the absence of an actual weapons investment. Some weapons inspection regimes serve this exact purpose. The Additional Protocol of the Nuclear Non-Proliferation Treaty,

for example, requires signatories to give weapons inspectors broader access to nuclear sites with less forewarning.

Unfortunately, though, credible information provision is not always possible. Potential proliferators, it seems, have incentive to continue nuclearizing at undeclared and unknown facilities. Lacking omniscience, rivals simply have to "live with uncertainty" (Dunn 1990). Yet living with uncertainty appears to imply playing Debs and Monteiro's interaction, which still results in mistaken preventive wars some of the time.

If information can never be perfect, is there another way to solve the problem? Indeed, there is. In fact, although verification receives most of the attention, many inspection regimes have secondary clauses geared toward the destruction of existing weapons infrastructure. To wit, consider the following features of the JCPOA:

- At the Arak facility, the reactor under construction will be filled with concrete, and the redesigned reactor will not be suitable for weapons-grade plutonium. Excess heavy water supplies will be shipped out of the country. Existing centrifuges will be removed and stored under round-the-clock IAEA supervision at Natanz.

- The Fordow Fuel Enrichment Plant will be converted to a nuclear, physics, and technology center. Many of its centrifuges will be removed and sent to Natanz under IAEA supervision. Existing cascades will be modified to produce stable isotopes instead of uranium hexafluoride (UF_6). The associated pipework for the enrichment will also be sent Natanz.

- Remaining operational centrifuges will be the IR-1 design, Iran's least efficient model.

- All enriched UF_6 in excess of 300 kilograms will be downblended to 3.67% or sold on the international market.[4]

- All enriched uranium oxide will be fabricated into fuel plates, sold on the international market, or diluted to 3.67%.

These features are not geared toward providing information. Rather, they dismantle Iran's nuclear infrastructure. Such a nuclear reversal does not make acquiring nuclear

[4]Iran has around 12 tons of low-enriched uranium currently.

weapons impossible, just harder. The skill and knowledge to produce reactors, centrifuges, heavy water, and enriched uranium does not go away. But replacing all of those materials—a necessary step to proliferate—would be immensely costly. Further, Iran would suffer a backlash from the international community for having reached an agreement only to backtrack at a later date.

Further, these features are not unique to the JCPOA. Disarmament is a primary pillar of the Nuclear Non-Proliferation Treaty, and manufacturing nuclear weapons materials puts signatories in violation of its terms. Meanwhile, other provisions appear on an *ad-hoc* basis. Libya shipped large portions of its nuclear infrastructure to the United States in the 2003 agreement (ElBaradei 2011, 157; Corera 2006, 221-222). Kazakhstan and Ukraine sold or shipped off their supplies of highly enriched uranium in return for Russian and American aid following the dissolution of the Soviet Union (Cirincione, Wolfsthal, and Rajkumar 2005, 371-375; Jones et. al. 1998, 80; Drezner 1999, 204). START had similar divestment requirements, including the guillotining of hundreds of B-52 Stratofortress bombers. And integration of nuclear scientists into the international community generally creates a brain drain, where technicians immigrate to above-the-board employment opportunities (Hymans 2012).[5]

One might also argue that the inspection provisions are more about costly inconvenience and less about information. Even if IAEA weapons inspectors keep close tabs on known-sites, Debs and Monteiro's work indicates that rivals must be worried about what they *don't* know. While the JCPOA has a provision to address unknown sites, this is irrelevant if the IAEA never discovers them. Meanwhile, if Iran practices its cat-and-mouse games, it could use known facilities as well. In turn, proliferation opponents might not know exactly how Iran will react to the inspection regime. Nevertheless, they can be sure that secretly replacing entire facilities will be a massive economic burden, while cat-and-mouse games will slow progress and be decidedly inefficient for a potential nuclear program.

Further, signing and subsequently breaking a treaty might be costly itself. This could be because treaties increase domestic bargaining power to those they favor (Simmons 2009, 125-147) or due to the coordination power of "naming and shaming" explicit

[5]Similar provisions exist in non-nuclear deals as well: Section II of the Washington Naval Treaty includes a long list ships for the signatories to scrap. Also note that while some of these cases featured joint reductions, not all do. Consequently, a *quid-pro-quo* argument—as seen in the civil war literature (Fortna 2004; Mattes and Burcu 2010)—cannot explain this behavior.

violators of international law (Hafner-Burton 2008; Franklin 2008). And although many of these treaties—including the JCPOA—have sunset clauses (Koremenos 2005), the delay itself is costly.

These points lead to two interrelated questions. First, if proliferation opponents can never truly be certain of a potential nuclear state's actions, will deals ever work? Second, why would a potential nuclear state ever agree to such reversal provisions? Indeed, they only make the proliferation process more expensive. Thus, if the state sought to make a deal, the additional cost would disincentive opponents from offering large concessions; after all, deals reached in coercive bargaining are commensurate with the actors' outside options. Meanwhile, if the state merely wanted to proliferate, it should be making that process as straightforward as possible.

The next section introduces a model that can answer these questions. While logically involved, game theory is the ideal tool for the task. Nuclear reversals involve layers of strategy. A potential proliferator would only want to reverse course if doing so would yield substantial concessions from its rival. A rival might only wish to offer concessions if it expects the proliferator to terminate its program as a result. But the temptation to construct nuclear weapons depends on the amount of additional concessions a potential proliferator would receive if it violated the agreement. Of course, fearing a large shift in power, opponents might terminate the threat with preventive war. And if a potential proliferator is expecting preventive war, it should not waste the costs of a program that has no hope of succeeding. This complex system of interdependent decision making requires some "accounting standards" to ensure that an argument's conclusion follows from its premises (Powell 1999, 32-34). Game theory is uniquely suited to handle this, and so I develop a model with all these features below.

3 The Model

The game consists of two players: state A (the proliferator) and state B (the opponent). As previewed above, there are four phases to the game. The first phase sees A choosing a cost $k \in [\underline{k}, \infty)$. This represents the price A must pay if it attempts to proliferate later in the interaction. Substantively, selecting $\underline{k}$ means A will allow itself the cheapest and easiest path possible to a nuclear weapon. Progressively higher values correspond to the adoption of various barriers to proliferation: policies similar to those in the JCPOA,

dismantling existing nuclear infrastructure (which would need to be rebuilt to produce a bomb), allowing weapons inspectors to effectively shut off civilian infrastructure from contributing to a weapons program, or permitting nuclear scientists to take positions overseas.[6] Put differently, the greater the level of k, the greater the level of nuclear reversal.

In the second phase, B offers a division of the stakes $x \in [0, 1]$ to A. This is a crisis bargaining scenario in which A and B have diametrically opposed preferences on a unit interval. If A rejects, the parties fight a game-ending war. A captures $p_A \in [0, 1)$ portion of the good, B takes the remainder, and the states pay respective costs $c_A, c_B > 0$ in the form of destroyed segments of the bargaining pie. These payoffs persist through the rest of time, with states sharing a common discount factor δ. Thus, standardizing all payoffs by the scalar $1 - \delta$, A's payoff for rejecting is $p_A - c_A$ and B's is $1 - p_A - c_B$.

If A accepts, the game transitions to a third stage that addresses weapons construction and preventive war similar to Debs and Monteiro's model. A first chooses whether to construct a nuclear weapon or not. Without observing A's decision, B must choose whether to fight a preventive war. As Figure 1 illustrated above, four outcomes can result from these moves. If A does not build and B prevents, both receive their war payoffs as though A had rejected the offer x. If A builds and B prevents, the payoffs are identical except A loses an additional k, reflecting that it paid the construction cost. If A does not build and B does not prevent, the states lock in the division B proposed beforehand, giving A the offer x and B the remainder $1 - x$ for the rest of time.

Lastly, if A builds and B does not prevent, A and B enjoy the proposed division (x and $1 - x$) for the period and the game moves to a fourth and final stage. B must once again offer a proposal $y \in [0, 1]$ to A. Accepting locks in the division for the remainder of time; no further weapons construction/preventive war decisions must be made because A already has nuclear weapons, and payoffs accrue for the periods without strategic interaction past that.[7] Again standardizing by $1 - \delta$, this gives the players respective overall payoffs of $(1-\delta)x+\delta y$ and $(1-\delta)(1-x)+\delta(1-y)$. Rejecting

[6]Note that the nuclear restrictions A can impose on itself are completely disconnected from the informational structure. If the inspection regime both increased information and the cost to proliferate, it would be difficult to track which mechanism assists in the creation of a deal. Because A cannot manipulate information revelation here, we can attribute the creation of a deal directly to the increased burdens.

[7]Later, I show the main results extend to a fully repeated game.

leads to a game-ending war. This is identical to before, except A now takes $p'_A \in (p_A, 1]$ of the good, leaving the remainder for B. The value $p'_A > p_A$ reflects A's enhanced bargaining position with nuclear weapons.[8] War remains costly, giving the players overall respective payoffs of $(1-\delta)x + \delta(p'_A - c_A)$ and $(1-\delta)(1-x) + \delta(1-p'_A - c_B)$ after the appropriate standardization.[9]

To recap, the timing is:

1. A selects $k \in [\underline{k}, \infty)$, its cost to proliferate
2. B makes an offer $x \in [0, 1]$; A accepts or rejects
3. If A accepts, A and B respectively and simultaneously choose whether to build or not and launch preventive war or not
4. If a power shift occurs, B makes a second offer $y \in [0, 1]$; A accepts or rejects

Note that B is strategically vulnerable in this setup: A can take B's concessions for the period and proliferate anyway. This addresses an important concern policymakers have about the JCPOA and similar agreements. Iran, for example, could enjoy sanctions relief and an improved stature in the international community for the moment, develop nuclear weapons, and gain the security benefits later. Proliferation could be a self-fulfilling prophecy as a result—potential proliferators nuclearize because they are not receiving concessions, and rivals do not offer concessions because the opponent is nuclearizing. If deals are possible despite A's temptation, it would assuage this policy concern.

[8]Note that the model only assumes that war outcomes with nuclear weapons are better for A and worst for B. Although a large literature debates whether nuclear weapons permit credible compellent threats (Schelling 1960, 193-199; Schelling 1966, 69-72 Betts 1987; Trachtenberg 1985; Pape 1996, 35-38; Gaddis 1987, 108-110; Beardsley and Asal 2009; Kroenig 2013), even skeptics recognize that opponents must limit their war aims when facing a nuclear opponent (Sechser and Fuhrmann 2013, 177-178). Shifting the war outcome in the proliferator's favor—just as the model does—reflects this power of deterrence.

[9]One might imagine that the costs of war will change due to the nuclear device. For simplicity, I keep c_A and c_B the same regardless of when the war occurs. Analogous theoretical results would still hold if they changed.

3.1 The Preventive War Calculus

Since this is an extensive form game with complete information, subgame perfect equilibrium (SPE) is the appropriate solution concept. SPE ensures that all equilibrium promises and threats are credible. I therefore begin at the end of the interaction, with post-nuclear bargaining. The solution here is simple:

Lemma 1. *Suppose A has successfully nuclearized. In every SPE, B offers* $y = p'_A - c_A$ *and A accepts.*

See the appendix for all omitted proofs. Lemma 1 is a straightforward application of Fearon's seminal bargaining game. After A has acquired a nuclear weapon, no further power shifts can occur. War remains costly, though, so B offers just enough to induce A to accept. Note that B must concede $p'_A - c_A$ to reflect A's newfound power, whereas A would only receive $p_A - c_A$ if it fought a war before the shift. This incentivizes A to build the weapon, which in turn motivates B to take measures to stop it.

Knowing the solution to the final step, we can solve for each state's preventive war/build decision. This is strategically compelling because B can manipulate both sides' incentives in this third step by changing its offer x in the second stage of the game. Thus, solving the preventive war/build problem for all different values of x allows B to find the x that optimizes its earlier offer decision.

Going forward, I focus on the parameter space where $\underline{k} > \delta(p'_A - c_A)$ and $p'_A > \frac{p_A + c_B}{\delta} + c_A$.[10] This is purely for analytical convenience; similar theoretical results appear in the other parameter spaces as well, but the proofs operate differently. As a reminder, for notational simplicity, I also standardize all payoffs by the scalar $1 - \delta$ throughout.

Lemma 2. *Suppose* $x \in \left[p'_A - c_A - \frac{k(1-\delta)}{\delta}, p_A + c_B\right]$. *Then A does not build and B does not prevent.*

The proof is straightforward and builds on the intuition of the model, so I describe it here. Figure 2 depicts the strategic form game being played in the third stage, using

[10] Substantively, the first condition guarantees that B prefers making a deal to taking everything up front and allowing a power shift to occur if A chooses no reversal. Here, A would want to undergo a reversal to induce B to make an agreement because the alternative is needlessly inefficient. The second condition ensures that B would not permit a power shift to occur if it knew A was building even if B took the entire good in the first stage. Again, because the alternative would result in needless inefficiency, the states would avoid this outcome in equilibrium.

	Prevent	$\sim$ Prevent
Build	$p_A - c_A - (1-\delta)k, 1 - p_A - c_B$	$(1-\delta)x + \delta(p'_A - c_A) - (1-\delta)k,$ $(1-\delta)(1-x) + \delta(1 - p_A + c_A)$
$\sim$ Build	$p_A - c_A, 1 - p_A - c_B$	$x, 1 - x$

Figure 2: The simultaneous move subgame for a generic offer x. A is the row player; B is the column player. Post-shift payoffs filled in using strategies from Lemma 1.

the payoffs from Lemma 1. Note that if $x \geq p'_A - c_A - \frac{k(1-\delta)}{\delta}$, A's build strategy is dominated.[11] This is because A strictly prefers to not build if B prevents—preventive war negates the possibility of acquiring the bomb, so A is better off not wasting the costs of proliferation. Meanwhile, A only wants to build if B does not prevent. That is, B can make A's stake in the status quo (x) so good that a successful power shift becomes unprofitable.[12] This is because B can manipulate A's opportunity cost by immediately making most of the concessions it would make if A proliferated, at which point paying the cost to build weapons is unnecessary. Given that B does not prevent, comparing A's payoff for building versus not building generates the x value that triggers the dominated strategy:

$$x \geq (1-\delta)x + \delta(p'_A - c_A) - (1-\delta)k$$
$$x \geq p'_A - c_A - \frac{(1-\delta)k}{\delta}$$

Thus, even though B cannot observe A's actions in this model, it can infer A's exact plan of action. Moreover—and this will be important later—B can always convince A not to build. All it must do is offer $x \geq p'_A - c_A - \frac{(1-\delta)k}{\delta}$. Whether B is willing to make such an offer remains to be seen.

From here, B can use a sequence of iterated elimination of dominated strategies to pick its best response. Because it can infer A will not build, B's choice is to prevent and earn $1 - p_A - c_B$ or not prevent and earn $1 - x$. Consequently, it does not prevent

[11]For brevity, I assume that an actor always chooses the efficient action when indifferent in this proof. In fact, with substantially more effort, it is possible to show that overall equilibrium to the game is unique. The logic is similar to why the ultimatum game has a unique SPE.

[12]As I explain below, this would also be true if the game were infinite horizon.

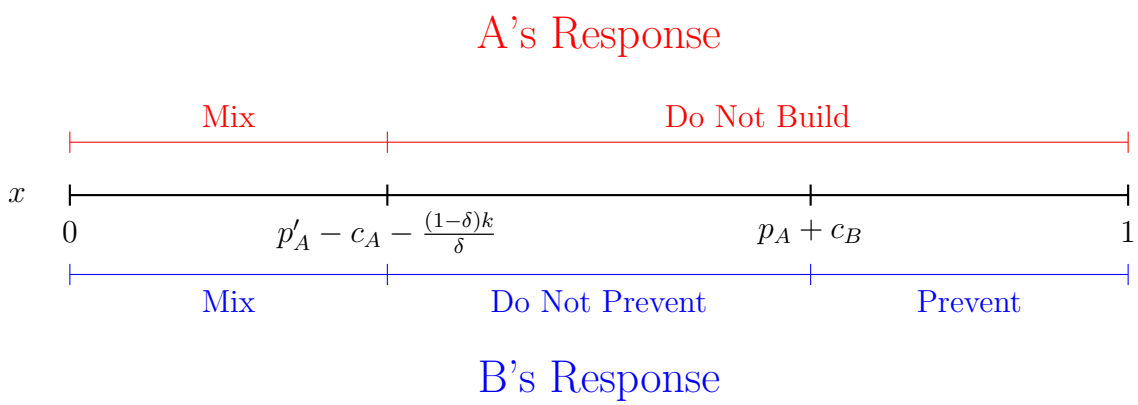

Figure 3: Each state's optimal response to different offer values x. Note that if $p_A+c_B < p'_A - c_A - \frac{(1-\delta)k}{\delta}$, the middle range ceases to exist, and the states adopt the strategies from the rightmost range.

if $x \leq p_A + c_B$. This holds for the parameters of Lemma 2.

Lemma 3. *Suppose $x > p_A + c_B$. Then B prevents and A does not build.*

This is also the result of iterated elimination of dominated strategies. Here, the amount B left for itself is so low that it prefers fighting a war regardless of A's build decision. Leaving such a small amount for itself in this manner is an obviously poor strategic decision. Accordingly, B will never put itself in this situation in equilibrium.

Lemma 4. *Suppose $x < p'_A - c_A - \frac{k(1-\delta)}{\delta} < p_A + c_B$. Then both players mix in equilibrium. Specifically, B prevents with probability $\frac{p'_A - c_A - \frac{(1-\delta)k}{\delta} - x}{p'_A - c_A - x}$ and A builds with probability $\frac{x - p_A - c_B}{\delta(x - p'_A + c_A)}$.*

From a strategic standpoint, Lemma 4's outcome is the most interesting. It corresponds to the case that Figure 1 described earlier. Because $x < p'_A - c_A - \frac{k(1-\delta)}{\delta}$, A prefers building if B does not prevent; because $x < p_A+c_B$, B prefers not preventing if A does not build. Therefore, no pure strategy Nash equilibrium exists. Each player must mix to stop the other side from exploiting it. The appendix uses the mixed strategy algorithm to derive the mixed strategies described in Lemma 4.

In review, how A and B act in this simultaneous move subgame depends on the offer x B issued beforehand. Figure 3 provides a convenient reference with each player's response. Sufficiently offers (i.e., $x \geq p'_A - c_A - \frac{(1-\delta)k}{\delta}$) convince A not to build; A mixes

otherwise. B, meanwhile, fights a war to gain back the losses it foolishly handed to A if the offer is too high (i.e., $x > p_A + c_B$). More moderate offers between $p'_A - c_A - \frac{(1-\delta)k}{\delta}$ and $p_A + c_B$ hit the sweet spot where B is happy to let the deal pass unmolested. But too small offers (i.e., $x < p'_A - c_A - \frac{(1-\delta)k}{\delta}$) causes B to distrust A and results in a mixing behavior.

3.2 Bargaining to Avoid War

The previous subsection showed the prospects for peace, proliferation, and preventive war depend on the proposal x. Now I address how B best manipulates that stage of the game through its offer in the bargaining stage. These cases are important because they explain how the world would operate without nuclear regimes like the IAEA that to manipulate proliferation incentives. Further, A's reversal decision will depend on the outcomes associated with particular cost levels.

The first case is straightforward:

Proposition 1. *Suppose the cost of proliferation is great relative to the extent of the power shift (i.e., $k \geq \frac{\delta(p'_A - p_A)}{1-\delta}$). In equilibrium, B offers $x = p_A - c_A$ and A accepts. The states then play the strategies in Lemma 2.*

Intuitively, A views proliferation in this model as an investment in the future—it pays a cost today to enjoy enhanced bargaining power in the future. At some point, though, the cost outweighs whatever potential gains may come. In turn, A prefers rejecting low offers to proliferation. B therefore offers the minimum amount necessary to buy A's peaceful compliance. A does not build and B does not prevent afterward.

Proposition 1's outcome is analogous to the standard bargaining model of war. When the cost to build weapons is too great, A cannot credibly threaten to do so. As such, B can treat the scenario as a static bargaining game. Potential proliferation does not factor in at all.

The remaining cases feature richer strategic environments. As it turns out, the game hinges on the value $x = p'_A - c_A - \frac{(1-\delta)k}{\delta}$. Recall that this is the minimum amount B must offer to make building a dominated strategy for A. Consequently, if B wishes to guarantee that A will not build, it ought to offer that amount; anything more is a needless concession. In turn, B can guarantee itself a payoff of $1 - p'_A + c_A + \frac{(1-\delta)k}{\delta}$ if it so wishes.

What is the alternative? This has a surprisingly simple answer: *any* smaller offer yields a payoff equal to its war value of $1 - p_A - c_B$. To understand why, note that two cases are possible. If B makes an offer according to Lemma 3, it always rejects during the build/prevent subgame. Thus, it earns its war payoff.[13] If B makes an offer according to Lemma 4, both states mix. It may then seem that B has a complicated payoff. However, recall that a player must be indifferent between his strategies to mix. Further, note from Figure 2 that B's payoff equals $1-p_A-c_B$ regardless of A's strategy if it prevents. Therefore, either way, B earns $1 - p_A - c_B$.

Consequently, despite the seemingly complicated web of possibilities, B's choice comes down to whether it would prefer earning $1 - p'_A + c_A + \frac{(1-\delta)k}{\delta}$ or $1 - p_A - c_B$. Note that a specific value of k determines which outcome B prefers:

$$1 - p'_A + c_A + \frac{(1-\delta)k}{\delta} = 1 - p_A - c_B$$

$$k^* = \frac{\delta(p'_A - p_A - c_A - c_B)}{1-\delta}$$

How the game ends for the remainder of the parameter space depends on whether k falls above or below that threshold:

Proposition 2. *Suppose the cost of proliferation falls in a medium range (i.e., $k \in \left[k^*, \frac{\delta(p'_A - p_A)}{1-\delta}\right]$ holds). In equilibrium, B offers $x = p'_A - c_A - \frac{(1-\delta)k}{\delta}$ and A accepts. The states then play the strategies in Lemma 2.*

In this case, the payoff for the remainder of the smallest acceptable bribe is greater than B's payoff for war. As such, B ensures that the deal succeeds. Note that the result here is efficient. A never builds, and the parties never go to war. This is despite the fact that B would go to war if A built. However, the sizable concessions assuage B's concern that A might take that route. Again, B is not offering such a large amount to be friendly—rather, offering a lower amount destroys this trust, leads to B having to fight a war, and ultimately gives B a lower payoff.

Nevertheless, B may prefer that route:

Proposition 3. *Suppose the cost of proliferation is low (i.e., $k < k^*$). The subgame*

[13] Because A earns its war payoff here as well, it is indifferent between advancing to the subgame and rejecting outright. However, rejection still gives B its war payoff.

has multiple SPE. B receives its war payoff in every SPE. Preventive war, mistaken preventive war, and successful proliferation are supported in SPE.

In effect, B can handle the proliferation problem using carrots or sticks. When the cost of proliferation is low, B only needs to offer moderate inducements (carrots) to obtain A's compliance. It thus chooses that route. But progressively decreasing the cost of proliferation requires giving A more and more carrots to keep it satisfied. When that cost eventually falls below k^*, B reaches the parameters of Proposition 3, and it becomes cheaper to fight a war instead. Thus, B opts for that strategy.

Note that, unlike the outcomes for Propositions 1 and 2, inefficiency results in this last case. This is because the states fight a war with positive probability and B may pay the costs of proliferation as well. The deadweight loss here leads to the surprising result that A sometimes prefers increasing its cost to proliferate.

3.3 Endogenous Nuclear Reversal

Given the many different paths the game can take, A has the opportunity to influence the final outcome by shifting its value of k at the beginning of the game. As discussed before, it may seem that A would always want to maintain the status quo cost to proliferate $\underline{k}$, as any increase would only harm its outside option. Yet the final proposition says otherwise:

Proposition 4. *In every SPE, if the smallest possible cost to build is sufficiently low (i.e., $\underline{k} < k^*$), A artificially inflates its cost to $k = k^*$.*

In words, if A's natural cost of proliferation is low enough that preventive war may occur, it ratchets up its cost to the minimum level necessary to induce B to offer a deal.

This result is counterintuitive and worth further explaining. Recall that the cost of proliferation determines whether B is willing to cut a deal with A. If the cost of proliferation is extremely large, B only needs to offer the potential proliferator a minimal amount and would thus prefer taking the remainder to fighting a war. But if the cost of proliferation is extremely cheap—which would be the case if A refused to divest—B prefers its war payoff and is unwilling to make the bribe.

In between these extremes, a particular cost of proliferation makes B indifferent between bribing A and fighting. Thus, at that cost, B is willing to take the efficient

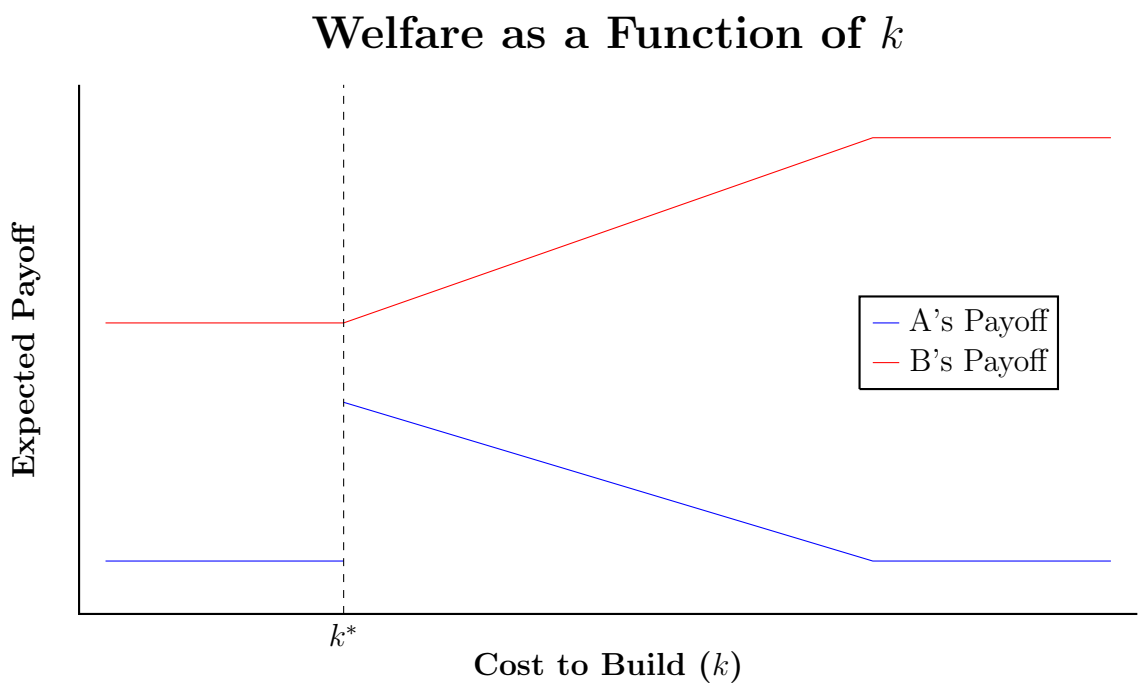

Figure 4: Each player's payoff as a function of the cost to build k. Note that the values past k^* Pareto dominate values below. Of these, A maximizes its payoff at k^*, which is why it divests to that value in equilibrium.

route. Here, B still receives its war payoff. Meanwhile, all of the costs that would have otherwise been wasted on war and proliferation can efficiently go to A. This is more than it would receive if k were lower and the parties mixed instead. Thus, *the potential proliferator gains from increasing its cost to build a nuclear weapon.*

Figure 4 illustrates both states' payoffs as k increases, holding the other parameters at $p_A = .2$, $p'_A = .7$, $c_A = .1$, $c_B = .25$, and $\delta = .9$. When costs are high, B treats the bargaining problem as though power were static and extracts the entire surplus. When costs fall in the middle range, B offers concessions, inducing A not to build despite the monitoring problem. When costs are low, massive inefficiencies from war and proliferation result. The deadweight loss ensures that moving to the middle range will increase both sides' payoffs.

4 Empirical Implications: Understanding Weapons Inspections and Nuclear Reversals

How do nuclear reversals, restrictions, and weapons inspections alter the bargaining environment? For weapons inspections in particular, the straightforward interpretation is that they provide information to rivals. Suppose a potential proliferator cannot effectively hide their programs from the prying eyes of inspectors. If rivals would launch preventive war short of a signal not to, a potential proliferator might wish to invite weapons inspectors into its country and allay its opponents' fears. Since refusal to admit inspectors inherently signals violations, the rivals could efficiently sort out the proliferators from the non-proliferators.

However, weapons inspectors are an imperfect solution. Proliferators have a home field advantage—violations could be anywhere in the country, leaving weapons inspectors with a lot of ground to cover without sufficient manpower. Strong intelligence alleviates some of the problem by giving inspectors likely locations of infractions, but that can be lacking as well since the absence of a smoking gun is not direct proof of compliance. In the extreme, cat-and-mouse games with inspectors might prevent inspections from providing any relevant informational content whatsoever.

Fortunately, even such ineffective weapons inspections have a hidden secondary effect that partially solves the proliferation problem. Although intelligence can be imperfect, the most efficient locations to construct weapons are often evident. Weapons inspections effectively shut down these avenues to proliferation since inspectors can investigate such known sites and report violations. In turn, proliferating states must seek alternative means to develop their weaponry. But since the weapons inspectors close off the most efficient means, the alternative methods are inherently costlier. Combined with divestment of technology, equipment, and materials, potential proliferators can use these international regimes to transform the bargaining environment to the Pareto dominant range.

Consequently, imperfect monitoring—that is, the inability to know with certainty whether a state has invested in a nuclear proliferation—is perfectly acceptable. Weapons inspectors do not act isolation; they are part of a greater negotiation strategy. While the cost of proliferation might still prove worth the finished product, the additional inefficiency incentives the rival to offer a deal, thus removing the incentive to prolifer-

ate. In turn, potential nuclear states accept inspectors and divest their infrastructure to make their non-proliferation commitments credible.

For practical purposes, it is important to note that weapons inspections might appear weak or ineffectual under these circumstances. For example, the JCPOA grants Iran some advanced notice before inspectors can visit certain facilities. While that time cannot mask radioactive signatures, policymakers worry that Iran could instead develop infrastructure that could disappear (Schumer 2015). This makes sense in light of Proposition 4. *Some* additional barriers to proliferation increase the potential builder's welfare, but too many is detrimental. In turn, it is unsurprising that potential proliferators do not agree to full reversals of their programs.[14] But absence of full reversal is not evidence of intention to proliferate.

Along those lines, it is important to note that the model does *not* predict that weapons inspections will be welcome under all circumstances. Indeed, as Figure 5 illustrates, nuclear restrictions are only mutually agreeable when proliferation and war would occur in equilibrium; reversal leads to Pareto improvement in that case. However, once the parameters reach Proposition 2's region, Figure 4 shows that the potential proliferator's payoff decreases in k. Consequently, the potential proliferator would be unwilling to go that far.

One can see how this logic played out in negotiations between the United States and some of its Cold War allies. Australia, Japan, and South Korea each at times considered developing their own weapon. Meanwhile, the United States sought to curb the spread of nuclear bombs. However, given the relatively warm relations between the United States and each of these countries, preventive war was not credible threat. Such a constraint pushes the allies into Proposition 2's parameters. Correspondingly, the United States offered economic and security concessions (Drezner 1999, 255; Campbell and Sunohara 2004, 222-224; Reynolds 2000, 195-196). Critically, though, the allies did *not* adopt reversal measures, just as the model would predict; divestment would only harm their bargaining position.

In the remaining case, reversal is irrelevant. If the potential proliferator's minimum proliferation cost $\underline{k}$ forces it to choose values exclusively in the range from Proposition 1, the exact choice does not matter—it has no desire to build weapons regardless of the

[14]By the same token, it should not be surprising that opponents push for full nuclear reversals—their payoffs weakly increasing in the cost to build.

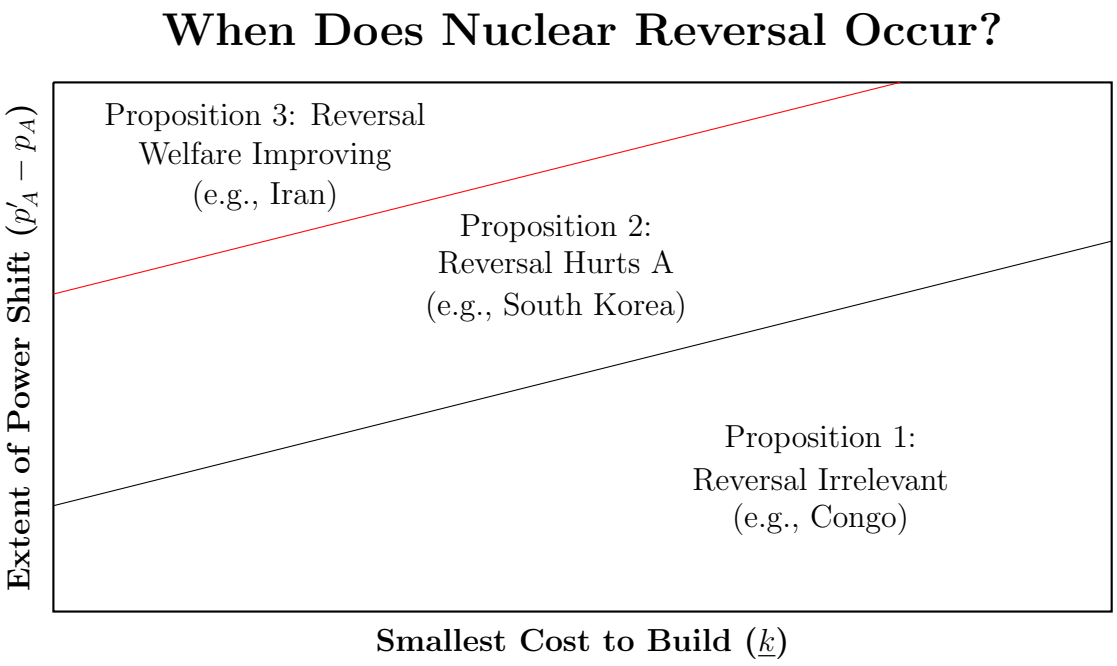

Figure 5: The utility of nuclear reversal as a function of the extent of the power shift and the smallest possible proliferation cost A can pay. The red line denotes the value k^*. If A starts with proliferation cost $\underline{k}$, it would endogenously reverse course to the red line.

specific circumstance. These parameters cover states that would not reasonably build nuclear weapons under any conditions, either because they lack major security threats (e.g., Iceland) or they do not have the economic means (e.g., the Democratic Republic of the Congo).

Figure 5 also reveals a counterintuitive empirical implication. From the outset, one might think that reversal is most likely to occur when weapons are most costly and have the least impact on the balance of power—in other words, when nuclear weapons are the bad investments. In fact, reversal is irrelevant under those circumstances because the potential proliferator would never want to build. Instead, reversal is most likely to occur under the exact opposite conditions: when weapons are cheap and the extent of the power shift is the greatest. Under such circumstances, it would seem that potential proliferators would be hard-pressed to reverse their programs because nuclear weapons are great investments here. In turn, one might expect these parameters to be the most likely to lead to bargaining breakdown and the corresponding inefficient outcomes. But precisely due to those concerns, the potential proliferator opts for reversal.

Before concluding, a couple of other important empirical implications follow from Proposition 4. To begin, not all nuclear reversals are equal:

Corollary 1. *When nuclear reversals are welfare increasing (i.e., $\underline{k} < k^*$), the extent of the reversal is decreasing in B's cost of preventive war.*

The proof is as simple as noting that if the potential proliferator wants to reverse course, the optimal divestment sets k equal to $k^* = \frac{\delta(p'_A - p_A - c_A - c_B)}{1-\delta}$. Thus, increasing c_B decreases that optimal value, meaning A divests less. Intuitively, decreasing the cost of preventive war makes B less willing to cut a deal. Because the consequences of no agreement are disastrous, A must further divest to incentivize a bargain.

Corollary 1's prediction matches the United States' experiences in the two most recent nuclear deals. In December 2003, Libya announced it would terminate its program. At the time, the United States was at the height of its post-September 11 geopolitical strength; Washington had quickly dispatched unfriendly regimes in Afghanistan and Iraq but was not yet experiencing the full-force of the insurgencies that would follow. Libya's program was relatively small and therefore an easy target. Combined with the fact that the Bush administration generally had hawkish tendencies, the United States' effective cost of war at the time was low. Correspondingly, Libya dismantled virtually

all of its nuclear infrastructure (ElBaradei 2011, 157; Corera 2006, 221-222; Bowen 2006, 72-77).

War against Iran—while perhaps plausible (Kroenig 2012; Moneteiro and Debs 2014, 50-51)—would not be so simple. While tensions exist between Iranian leadership and its populace, the country is comparatively more united than Libya was prior to the 2011 civil war. Moreover, Iran has invested in fortifying its nuclear sites; a death knell to the program would need significantly greater military effort, likely requiring a ground assault of some kind. Meanwhile, the United States has endured more than a decade of insurgencies and risks further polarizing the Muslim world with another conflict in the region. As the model would predict, the JCPOA only calls for Iran to moderately reverse its program.

Regardless, the winner of divestment may come as a surprise:

Corollary 2. *If $\underline{k} < k^*$, A captures the entire surplus in equilibrium.*

From a game-theoretical perspective, Corollary 2's is exceptionally strange. Normally, in bargaining games in which only one individual has proposal power (e.g., ultimatum games), the proposer captures the entire surplus. Here, the proposer is B. Yet, in equilibrium, A's endogenous nuclear reversal leads to a cost that makes B indifferent between making a deal and going to war. Thus, B's payoff equals $1 - p_A - c_B$; because the result is efficient, this implies that A takes the remainder, or $p_A + c_B$. The surplus goes to the receiver.[15]

From an empirical perspective, from the outset, it would appear that A is the more vulnerable state in this type of interaction. After all, it has not acquired a nuclear weapon yet, to do so requires paying a cost, and the opponent can stop that shift with preventive war. This intuition is wrong. A can manipulate B's incentives so that A receives all of the benefits from reaching a deal. Arms treaties therefore primarily benefit potential proliferators, not their opposition.[16]

Corollary 2 also demonstrates why the model is robust to a number of substantive concerns that might otherwise seem to derail reversal efforts. I detail a few of these

[15]These results do not extend to the other parameter spaces. For the parameter space of Proposition 2, the states split the surplus; for Proposition 1, B captures the entire surplus.

[16]Although I assume that the cost of proliferation is strictly positive, Corollary 2 shows that Proposition 4's result would hold even if the cost for proliferation was negative, perhaps due to "prestige" arguments or domestic politics. The only requirement is that conflict (in the form of positive probability of building and war) be be inefficient—i.e., $c_A + c_B + k > 0$.

concerns below.

Inspections as a Power Shift. Some potential proliferators might be reluctant to allow reversal regimes into their countries, as information leaks could give rivals tactical advantages. Other states worry that weapons inspectors—in the process of shutting down the most efficient paths to proliferation—will steal trade secrets (Schiff 1983, 94).[17] This is, in part, why rivals delegate weapons inspections to international organization and not bilateral task forces (Schiff 1983, 95). But international organizations are not perfect. To wit, the United Nations Special Commission on Iraq, the regime tasked with monitoring Iraqi compliance following the Persian Gulf War, ceased operations in 1999 amid allegations that Western intelligence agencies had infiltrated it (Blix 2004, 36-37; ElBaradei 2011, 32-33).

Fortunately, espionage does not inherently doom inspection regimes. At its core, Proposition 4's result stems from the fact that inflating k past k^* yields Pareto improvement. As Corollary 2 explained and Figure 4 illustrated, these additional benefits flow to the potential proliferator. In fact, the power shift caused by inspector espionage would have to exceed the sum costs of war to make a deal impossible. This seems highly implausible. Consequently, those states would still happily reverse course—the substantial gains more than offset whatever tactical problems come with them.

Costly Weapons Inspections. Reversal regimes and their associated weapons inspectors are not free—someone must pay for their employment, administrative support, and logistics. As such, reversal regimes can only lead to welfare improvement if their cost is less than the inefficiency from warfare and weapons construction.[18]

Given how economically disruptive and destructive war is, reversal is easily the cheaper option. The JCPOA, for example, calls for a maximum of 150 weapons inspectors. In Iraq, UNSC Resolution 986 funded weapons inspections with just 0.8% of the revenue brought in from the food-for-oil program. Meanwhile, the total yearly budget for IAEA inspections is about $120 million (ElBaradei 2011, 80). Many states share the cost of this burden, as is standard with information providing institutions (Keohane 1984). Even in marginal cases, third party states have incentive to contribute to avoid

[17]Accordingly, the JCPOA (via the Additional Protocol) allows Iranian intelligence to deny visas to weapons inspectors of its choosing.

[18]This is essentially a costly peace argument (Powell 2006; Coe 2012).

war's negative externalities. Thus, although these costs add to the difficulty of reaching an agreement, they certainly do not render reversals impossible.

Possible Detection. As stated from the outset, the model looked at the worse-case scenario where B has no capability of observing A building. As it turns out, relaxing this assumption merely shifts around surplus. A noisy signal revealing that A has built effectively makes nuclear weapons more expensive—some portion of the time, B will see it and declare war, thereby forcing A to pay the cost of proliferation without any benefits. This gives rise to regions analogous to Proposition 2 and 3, as seen in Figure 5. Thus, if the proliferation cost remains high enough, B is willing to buy off A (at a lower price), and states avoid the inefficient mixing outcome. But when the cost is sufficiently low, the parties still end up mixing.

Once more, this logic helps explain why potential proliferators demand limits on information revelation to their opponents. With a perfectly informative signal, the credible threat of preventive war eliminates nuclearization as an option entirely, causing the potential proliferator to lose out on all the surplus.[19] But even smaller measures hurt the value of its outside option. In turn, *some* information revelation proves useful if it convinces the rival to cut a deal and preserve the efficient outcome. However, any further information provision can only hurt the potential proliferator. Once again, as a consequence, rivals cannot use refusal to provide information as proof positive that the potential proliferator will nuclearize.

Infinite Horizon. For simplicity, the model analyzed a two-period interaction. However, it is easy to verify that the same welfare improvement mechanism works if the states bargained repeatedly. The key to seeing this is recognizing that an efficient agreement—without war or proliferation—requires A receive at least $p'_A - c_A - \frac{(1-\delta)k}{\delta}$ and B receive at least $1 - p_A - c_B$ by accepting a settlement in each period. However, choosing a k value less than k^* ensures that the sum of these minimal payoffs exceed 1, making it impossible to reach such a settlement. Inefficient behavior must result.

Further, for the same reason as in the two-period model, B must expect to receive $1 - p_A - c_B$ in the absence of a deal. Due to the inefficiency, A must receive strictly less than $p_A + c_B$.

[19] See Spaniel 2015 for a model with perfect information.

However, raising k to at least k^* permits B to make acceptable offers. In turn, A can take the surplus. This is preferable to keeping k below k^* and receiving strictly less than the remainder of B's reservation value. Hence reversal would still occur in an infinite horizon setup.

5 Conclusion

This paper explored the role of divestment strategies in combatting nuclear arms programs. Whereas research normally focuses on the monitoring and verification clauses of arms agreements, I showed that reversing course can entice opposing states to offer generous terms to terminate nuclear programs. Consequently, potential proliferators are the primary beneficiaries of such agreements despite their apparently precarious bargaining position.

I conclude with four policy implications. First, rivals of states with nuclear programs should not treat proposed reversals as obvious traps. Although such agreements might seem too good to be true, potential proliferators have good reason to divest their infrastructure and avoid facing preventive war. Moreover, North Korea nonwithstanding, such reversal deals have a good historical track record. This does not mean that the JCPOA or other future deals are guaranteed to work—other mechanisms may lead to bargaining failure—just that policymakers should not immediately dismiss them as inherently incredible.

However, it is important to note that reversals are mere skipping stones to an agreement. Potential proliferators do not divest out of generosity—they do because the expect to receive concessions in return. Despite apparent commitment problems, these concessions are credible because potential proliferators maintain some ability to restart their programs. Thus, rivals should not see reversal as a victory in itself—concessions must follow, or the nuclear outcome will result anyway.

Third, these agreements cannot be fleeting. Again, this is a matter of credibility. Potential proliferators do not choose to only partially divest because they want to backtrack later—they maintain some capabilities because the *threat* to backtrack later keeps the concessions flowing. Demanding complete divestment can therefore cause the house of cards to crumble. But long-lasting concessions throughout time incentivize the potential proliferator to remain non-nuclear, allowing everyone to win.

Finally, the model gives reason to be cautiously optimistic about the future of nuclear non-proliferation. A consistent policy concern is that the nuclear club will continue to expand as the cost of proliferation declines (e.g., Reiss 2004, 4). While technology may ultimately increase everyone's latent nuclear capacity, the model indicates that states still have incentive to erect barriers on their own. Indeed, potential proliferators optimally adopt ever *more* hurdles as the ease of nuclearizing increases. Thus, in some cases, the apparent proliferation problem may solve itself.

6 Appendix

This appendix covers all proofs that were missing from the main text.

6.1 Proof of Lemma 1

The proof is a trivial application of the bargaining model of war. All payoffs from the first period are sunk. Ignoring those and working backward, A receives δy if it accepts and $\delta(p'_A - c_A)$ if it rejects. Therefore, A is willing to accept if $y \geq p'_A - c_A$. B receives $\delta(1 - p'_A - c_B)$ if it induces rejection and $\delta(1 - y)$ if it induces acceptance. Note that the optimal acceptable offer yields $1 - p'_A + c_A$, which is greater than inducing rejection. Therefore, in all SPE, B offers $y = p_A - c_A$ and A accepts $y \geq p_A - c_A$. □

6.2 Proof of Lemma 4

If $x < p_A + c_B$, it is trivial to show that no pure strategy Nash equilibria exist and that the only way one player can be willing to mix is if the other also mixes. Thus, consider mixed strategy Nash equilibria by deriving the indifference conditions.

To begin, let σ_p be the probability B prevents. Then A's expected utility for building equals:

$$\sigma_p[p_A - c_A - k(1-\delta)] + (1-\sigma_p)[(1-\delta)x + \delta(p'_A - c_A) - (1-\delta)k]$$

$$\sigma_p[p_A - c_A - (1-\delta)x - \delta(p'_A - c_A)] + (1-\delta)x + \delta(p'_A - c_A) - (1-\delta)k$$

And A's expected utility for not building equals:

$$\sigma_p(p_A - c_A) + (1 - \sigma_p)x$$

$$\sigma_p(p_A - c_A - x) + x$$

Setting these two equations equal to each other and solving for σ_p yields:

$$\sigma_p = \frac{p'_A - c_A - \frac{(1-\delta)k}{\delta} - x}{p'_A - c_A - x}$$

Now consider B's indifference condition. B's expected utility for preventing is $1 - p_A - c_B$ regardless of R's strategy. Meanwhile, letting σ_b be the probability A builds, B's expected utility for not preventing equals:

$$\sigma_b[(1 - \delta)(1 - x) + \delta(1 - p'_A + c_A)] + (1 - \sigma_b)(1 - x)$$

$$\sigma_b[\delta(x - p'_A + c_A)] + 1 - x$$

Setting these two values equal to each other and solving for σ_b yields:

$$\sigma_b = \frac{p_A + c_B - x}{\delta(p'_A - c_A - x)}$$

Note that if B is mixing, it is indifferent between preventing and not preventing. Since preventing yields a flat $1 - p_A - c_B$, B receives its war payoff in this case. □

6.3 An Additional Case

In addition to the parameters of the lemmata, there is still one unlisted case: when $x = p_A + c_B < p_A - c_A - \frac{(1-\delta)k}{\delta}$. Here, A does not build and B prevents with probability $\sigma_p \in [\frac{\delta(p_A+c_B-p'_A+c_A)+(1-\delta)k}{\delta(p_A+c_B-p'_A+c_A)}, 1]$.

To see this, begin by noting that preventing now weakly dominates not preventing. Not building is the best response to preventing, so preventing and not building is an equilibrium.

No other pure strategy Nash equilibria exist. Not preventing and not building is not an equilibrium; A needs at least $p'_A - c_A - \frac{(1-\delta)k}{\delta}$ to not want to deviate to building, but $p_A + c_B$ is less than that. Building and not preventing is not an equilibrium since B prefers to prevent. Lastly, preventing and building is not an equilibrium because A

could deviate to not building and save the investment cost.

Now consider mixed strategy Nash equilibria. Since preventing weakly dominates not preventing, if A mixes, B must prevent as a pure strategy. But since A strictly prefers not building in this case, A cannot optimally mix.

Thus, the remaining cases to consider require B to mix and A to select a pure strategy. B is indifferent between preventing and not preventing only if A does not build. For A to be willing to not build, Lemma 4 shows that $\sigma_p \geq \frac{\delta(p_A+c_B-p'_A+c_A)+(1-\delta)k}{\delta(p_A+c_B-p'_A+c_A)}$.

Thus, in equilibrium, A does not build and B prevents with probability at least $\frac{\delta(p_A+c_B-p'_A+c_A)+(1-\delta)k}{\delta(p_A+c_B-p'_A+c_B)}$. Since B is indifferent between preventing and not preventing in all of these cases, it earns its war payoff.

6.4 Proof of Proposition 1

Because $p_A - c_A > p'_A - c_A - \frac{(1-\delta)k}{\delta}$ in this parameter space, by Lemma 2, the outcome of the simultaneous move subgame if B offers $x \geq p_A - c_A$ is not build/not prevent. Note that B's optimal offer in this range is $p_A - c_A$, leaving $1 - p_A + c_A$ for B.

If $x < p_A - c_A$, every outcome in the simultaneous move subgame generates a payoff less than $p_A - c_A$ except prevent/build. But this outcome cannot occur with certainty because B would deviate to not preventing. Therefore, A earns strictly less than $p_A - c_A$ in the simultaneous move subgame. But A can earn $p_A - c_A$ by rejecting x instead. Thus, it rejects $x < p_A - c_A$.

In turn, B earns $1 - p_A - c_B$ if it offers $x < p_A - c_A$. This is less than the $1 - p_A + c_A$ it earns by offering $x = p_A - c_A$. Consequently, B offers that amount, and the states play the simultaneous move game according to Lemma 2. □

6.5 Proof of Proposition 2

Here, A can respond in one of four ways to B's offer. If $x < p'_A - c_A - \frac{(1-\delta)k}{\delta}$, A finds itself mixing in the simultaneous move subgame according to Lemma 4. In either case, B obtains its war payoff. If $x \in \left[p'_A - c_A - \frac{(1-\delta)k}{\delta}, p_A + c_B\right]$, A accepts and the parties reach an agreement according to Lemma 2. B takes the remainder, or $1 - x$. Finally, if $x > p_A + c_B$, A's exact choice becomes irrelevant per Lemma 3. B receives its war payoff.

Consequently, B maximizes its payoff subject to these constraints. In the first and last case, it receives its war payoff. In the second case, B's optimal offer equals $p'_A - c_A - \frac{(1-\delta)k}{\delta}$, earning it $1 - p'_A + c_A + \frac{(1-\delta)k}{\delta}$, which is greater than $1 - p_A - c_B$ in this parameter space. Since B's payoff is decreasing in x, its optimal offer equals $p'_A - c_A - \frac{(1-\delta)k}{\delta}$. A accepts, and the parties play the simultaneous move game according to Lemma 2. □

6.6 Proof of Proposition 3

If $k < k^*$, the range of offers from B can end the game in the mixing parameters of Lemma 4 or in the guaranteed war parameters of Lemma 3. Recall that Lemma 4 generates an expected payoff for B equal to its war value. This, of course, is identical to the payoff for always fighting according to Lemma 3. As such, B is indifferent between the two. It can therefore choose either as a pure strategy or mix freely in equilibrium. The parties then play according to the respective lemma, which allows for preventive war and proliferation to occur on the path. □

6.7 Proof of Proposition 4

Proving Proposition 4 is a matter of calculating A's payoff for each value it can choose for k. The parameter space from the proposition means there are three options: pick a value such that the states play the strategies from Proposition 1 afterward, a value such that they play the strategies from Proposition 2 afterward, or a value such that they play the strategies from Proposition 3. The first case generates a flat war payoff of $p_A - c_A$ regardless of the specific value. The second case generates a payoff equal to the value B offers, or $p'_A - c_A - \frac{(1-\delta)k}{\delta}$. This value is decreasing in k, so the optimal choice in this range is $k = k^*$, which yields:

$$p'_A - c_A - \frac{(1-\delta)\frac{\delta(p'_A - p_A - c_A - c_B)}{1-\delta}}{\delta} = p_A + c_B$$

This is better than selecting the first case.

The remaining case either leads to guaranteed war or the mixing strategies from Lemma 4. In the first case, A once again receives $p_A - c_A$, so that cannot be optimal. In the second case, recall that the indifference conditions ensure that B receives $1 - p_A - c_B$.

Preventive war and building occur in equilibrium. Both of these are deadweight loss. The exact amount of deadweight loss built into the system could depend on the mixing probabilities, but $L > 0$ can represent this amount regardless. Since the total value of the good equals 1, A's payoff equals:

$$1 - (1 - p_A - c_B) - L$$

$$p_A + c_B - L$$

Because $L > 0$, this is worse than selecting k^*. Therefore, A artificially inflates its cost to k^*. □

Works Cited

Adelman, Kenneth L. 1990. "Why Verification Is More Difficult (and Less Important)." *International Security* 14 (4): 141-146.

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

Banks, Jeffrey S. 1990. "Equilibrium Behavior in Crisis Bargaining Games." *American Journal of Political Science* 34 (3): 599-614.

Beal, Tim. 2005. *North Korea: The Struggle against American Power.* Ann Arbor: Pluto Press.

Beardsley, Kyle and Victor Asal. 2009. "Winning with the Bomb." *Journal of Conflict Resolution* 53 (2): 235-55.

Betts, Richard K. 1987. *Nuclear Blackmail and Nuclear Balance.* Washington, DC: Brookings Institute.

Blix, Hans. 2004. *Disarming Iraq.* New York: Random House.

Bowen, Wyn Q. 2006. *Libya and Nuclear Proliferation: Stepping Back from the Brink.* New York: Routledge.

Campbell, Kurt M. and Tsuyoshi Sunohara. 2004. "Japan: Thinking the Unthinkable." In *The Nuclear Tipping Point: Why States Reconsider Their Nuclear Choices* Eds. Kurt M. Campbell, Robert J. Einhorn, and Mitchell B. Reiss. Washington D.C.: Brookings Institution Press.

Chadefaux, Thomas. 2011. "Bargaining over Power: When Do Shifts in Power Lead to War?" *International Theory* 3 (2): 228-253.

Cirincione, Joseph, Jon B. Wolfsthal, and Miriam Rajkumar. 2005. *Deadly Arsenals: Nuclear, Biological, and Chemical Threats.* Washington DC: Carnegie Endowment for International Peace.

Coe, Andrew J. 2012. "Costly Peace and War." Manuscript, Harvard University.

Corera, Gordon. 2006. *Shopping for Bombs: Nuclear Proliferation, Global Insecurity, and the Rise and Fall of the A.Q. Khan Network.* Oxford: Oxford University Press.

Downs, George W., David M. Rocke, and Randolph M. Siverson. 1986. "Arms Races and Cooperation." In *Cooperation under Anarchy* Ed. Kenneth A. Oye. Princeton: Princeton University Press.

Drezner, Daniel W. 1999. *The Sanctions Paradox: Economic Statecraft and International Relations.* Cambridge: Cambridge University Press.

Dunn, Lewis A. 1990. "Arms Control Verification: Living with Uncertainty." *International Security* 14 (4): 165-175.

ElBaradei, Mohamed. 2011. *The Age of Deception: Nuclear Diplomacy in Treacherous Times.* New York: Metropolitan Books.

Fearon, James D. 1995. "Rationalist Explanations for War." *International Organization* 49 (3): 379-414.

Fey, Mark and Kristopher W. Ramsay. 2011. "Uncertainty and Incentives in Crisis Bargaining: Game-Free Analysis of International Conflict." *American Journal of Political Science* 55 (1): 149-169.

Fortna, Virginia Page. 2004. *Peace Time: Cease-Fire Agreements and the Durability of Peace.* Princeton: Princeton University Press.

Franklin, James C. 2008. "Shame on You: The Impact of Human Rights Criticism on Political Repression in Latin America." *International Studies Quarterly* 52 (1): 187-211.

Gaddis, John Lewis. 1987. *The Long Peace: Inquiries into the History of the Cold War.* Oxford: Oxford University Press.

Gallagher, Nancy W. 2003. *The Politics of Verification.* Baltimore: Johns Hopkins University Press.

Hafner-Burton, Emilie M. 2008. "Sticks and Stones: Naming and Shaming the Human Rights Enforcement Problem." *International Organization* 62 (4): 689-716.

Hymans, Jacques E.C. 2012. *Achieving Nuclear Ambitions: Scientists, Politicians, and Proliferation.* Cambridge: Cambridge University Press.

Ikle, Fred Charles. 1961. "After Detection—What?" *Foreign Affairs* 39 (1): 208-220.

Jones, Rodney W., Mark G. McDonough, Toby F. Dalton, and Gregory D. Koblentz. 1998. *Tracking Nuclear Proliferation: A Guide in Maps and Charts, 1998.* Washington DC: Carnegie Endowment for International Peace.

Keohane, Robert O. 1984. *After Hegemony: Cooperation and Discord in the World Political Economy.* Princeton: Princeton University Press.

Koremenos, Barbara. 2005. "Contracting around International Uncertainty." *American Political Science Review* 99 (4): 549-565.

Krass, Allan S. 1985. *Verification: How Much Is Enough?* Lexington: Lexington Books.

Kroenig, Matthew. 2012. "Time to Attack Iran." *Foreign Affairs* 91 (1): 76-86.

Kroenig, Matthew. 2013. "Nuclear Superiority and the Balance of Resolve: Explaining Nuclear Crisis Outcomes." *International Organization* 67 (1): 141-171.

Mattes, Michaela and Burcu Savun. 2010. "Information, Agreement Design, and the Durability of Civil War Settlements." *American Journal of Political Science* 54 (2): 511-524.

Meyer, Stephen M. 1984. "Verification and Risk in Arms Control." *International Security* 8 (4): 111-126.

Monteiro, Nuno P. and Alexandre Debs. 2014. "The Strategic Logic of Nuclear Proliferation." *International Security* 39 (2): 7-51.

Pape, Robert A. 1996. *Bombing to Win: Air Power and Coercion in War.* Ithaca: Cornell University Press.

Powell, Robert. 1999. *In the Shadow of Power: States and Strategies in International Politics.* Princeton: Princeton University Press.

Powell, Robert. 2006. "War as a Commitment Problem." *International Organization* 60 (1): 169-203.

Reed, William, Scott Wolford, and Philip Arena. 2015. "Gambling on Diplomacy: Bargaining in the Shadow of Uncertain Shifting Power." Manuscript, University of Maryland.

Reiss, Mitchell B. 2004. "The Nuclear Tipping Point: Prospects for a World of Many Nuclear Weapons States." In *The Nuclear Tipping Point: Why States Reconsider Their Nuclear Choices* Eds. Kurt M. Campbell, Robert J. Einhorn, and Mitchell B. Reiss. Washington D.C.: Brookings Institution Press.

Reynolds, Wayne. 2000. *Australia's Bid for the Atomic Bomb.* Melbourne: Melbourne University Press.

Schelling, Thomas C. 1960. *The Strategy of Conflict.* Cambridge: Harvard University Press.

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

Schiff, Benjamin N. 1983. *International Nuclear Technology Transfer: Dilemmas of Dissemination and Control.* Totowa: Rowman & Allanheld.

Schumer, Chuck. 2015. "My Position on the Iran Deal." `https://medium.com/@SenSchumer/my-position-on-the-iran-deal-e976b2f13478`. Accessed 8/16/2015.

Sechser, Todd S. and Matthew Fuhrmann. 2013. "Crisis Bargaining and Nuclear Blackmail." *International Organization* 67 (1): 173-195.

Simmons, Beth A. 2009. *Mobilizing for Human Rights: International Law in Domestic Politics.* Cambridge: Cambridge University Press.

Spaniel, William. 2015. "Arms Negotiations, War Exhaustion, and the Credibility of Preventive War." Forthcoming, *International Interactions.*

Trachtenberg, Marc. 1985. "The Influence of Nuclear Weapons in the Cuban Missile Crisis." *International Security* 10 (1): 137-163.