Effectiveness of Counter-proliferation Measures and their Impacts on Security.

Kai Hao Yang*

April 2017

Abstract

This paper investigates the strategic interactions between the counter proliferator and the proliferator in a nuclear proliferation crisis, as well as their impacts on international security and stability. A baseline model of contest with interdependent values is established and its implications are discussed. Furthermore, we characterize the equilibria in a class of models in a "detail-free" fashion and analyze equilibrium outcomes, with particular attentions to international stability and likelihood of a successful development. It thus yields some results and implications that are robust to game forms and model details and provides several generalizations and insights to the effects of various counter-proliferation measures as well as the consequences of nuclear proliferation.

*Department of Economics, University of Chicago; e-mail: khyang@uchicago.edu. I am grateful for the comments and suggestions from Jin Yeub Kim, Roger Myerson, Emerson Niou, as well as the session participants of the 2017 MPSA conference.

1 Introduction

The development of nuclear technology has raised numerous concerns about the issue of nuclear proliferation. This can be reflected by the increasing calls for regulations and establishments of international regimes (Simpson, 1994; Feaver & Niou, 1996; Forland, 2007; Goldblat, 2007). Although it is still debatable whether nuclear proliferation has a negative impact on stability, nuclear proliferation does threaten some certain countries (Feaver & Niou, 1996; Schneider, 1994). As a result, possible countermeasures of the threatened countries (so called the *counter-proliferators*) against the *proliferators* have become a crucial component of their security strategies. During the *nuclear proliferation process*, where the proliferator tries to develop nuclear weapons, the counter-proliferator has several possible countermeasures, both in theory and in practice, including tolerating the proliferator and granting economics incentives, sanctions and *preventive strikes* (Dunn, 1991; Martel, 2001; Feaver & Niou, 1996; Conway, 2003; Benson & Wen, 2011; Baliga & Sjostrom, 2008; Schneider, 2004).[1] Clearly, there are often trade-offs among different strategies. For more "tolerance-orientated" counter-proliferation strategies, it is less likely that conflicts would occur, but more likely that the proliferator may succeed in acquiring nuclear technology. In contrast, for "tougher" counter-proliferation strategies, it is less likely that the proliferator's acquisition would be successful but the risk of unnecessary conflicts might be larger. (Feaver & Niou, 1996; Forland, 2007; Moriarty, 2004; Baliga & Sjostrom, 2008; Benson & Wen, 2011). Thus, the strategies of the counter-proliferator as well as the responses of the proliferator can lead to vastly different outcomes during the nuclear proliferation processes. Moreover, in most of the nuclear proliferation crisises, the counter-proliferator is often uncertain about the proliferator's technology level and hence the time duration for research and development before successfully acquiring nuclear weapons. This further com-

[1] Conventionally, there are some distinctions between *preventive strikes* and *preemptive strikes*. While the former refers to intiating attacks regardless of whether the proliferator has the *intention* to threat, the latter refers to strikes under an *imminent threat*. Under the nuclear proliferation context, although the terminology used in the speech given by president George W. Bush in 2002 was using *preemptive surgical strikes*, the preconditions of strikes did not include salient military preparedness and thus is effectively more similar to *preventive strikes* (Forland, 2007). In general, due to the characteristic of weapons of mass destruction that the time period between preparedness and actual attacks is relatively short, there is no need to distinguish preventive and preemptive strikes (Goldstein, 2006). Therefore, the term "preventive strike" will be used throughout this paper for consistency.

plicates the counter-proliferator's strategical choices, and hence the outcomes of the nuclear proliferation process.

In this paper, we model the nuclear-proliferation process as an allocation problem with interdependent value. We first develop a baseline model that regards the nuclear proliferation process as a brinkmanship, or equivalently, a contest game with interdependent values. This allows us to thoroughly examine the effects of counter-proliferation strategies *during* the proliferation process. Then, we generalize the framework beyond contest games and consider a large class of models for the nuclear proliferation process, including the baseline model as well as the models in seminal papers such as Fearon (1994) and Powell (2003). By using a mechanism design approach, we are able to examine the effectiveness and implications of the counter-proliferation strategies in a "detail-free" fashion and provide further insights to the design of counter-proliferation regimes.

The remaining of this paper is organized as follows: In the next section we discuss several related literature. In the following section, a baseline is established and an equilibrium will be characterized. Then its implications will be discussed. In section 4 we will generalize the theory and discuss both the stability and likelihood of a successful development under a more general framework. Finally, some implications, both theoretical and empirical, will be discussed.

2 Related literature

From the long-run perspective, one of the most salient question about nuclear proliferation is whether it is beneficial for the stability of the international system. There are mainly two distinct perspectives, namely the *optimism* and the *pessimism*. The optimism posits that possessions of nuclear weapons of one country will force other countries to be more cautious, as it was presented in the *strategic stability* relationship between the U.S. and Soviet Union during the Cold War era. Therefore, as more countries posses nuclear weapons, international system would be more stable, because the likelihood of conflicts would be reduced due to cautions of countries (Feaver, 1993; Karl, 1996; Goldstein, 2006:146-163; Waltz, 1990; 2003; Sagan & Waltz, 1995; Bueno de Mesquita & Riker, 1982; Berkowitz, 1985; Measheimer, 1993; Asal & Beardsley, 2007; Gartzke & Jo, 2011). On the other hand, the pessimism posits that since the background and context during the cold war era are significantly different from the post-cold war era and therefore the strategic stability

argument is not applicable nowadays (Dunn, 1982; Kaiser, 1989; Karl, 1996; Goldstein, 2006:15-20; Woods, 2002; Asal & Beardsley, 2007). Furthermore, factors such as miscalculation, irrationality of proliferators, strategic instability, crisis instabliity and the possibility of preventive strikes initiated by the counter-proliferator, might create conflicts after more countries acquire nuclear weapons or during the nuclear proliferation process (Goldstein, 2006: 15-20; Berkowitz, 1985; Feaver, 1993; Feaver & Niou, 1996; Karl, 1996; James, 2000; Sagan, 2003). While the "strategic instability" concern that arises from asymmetric nuclear abilities after proliferation and the "crisis instability" concern that arises from the possession of nuclear weapons are both consequences *after* successful nuclear proliferation (Berkowitz, 1985), this paper focuses on the impacts *during* the nuclear proliferation process.

From the short-run perspective, on the other hand, there are numerous empirical studies about the motivations for the proliferators to acquire nuclear weapons. Singh & Way (2004) indicate that the most important factor the determines whether a country will attempt to develop nuclear weapons is the external threats. Jo & Gartzke (2007) further show that the threat from traditional military forces is more significant than nuclear threats. They also find that one's economic capability is also a crucial determinant. Gartzke & Jo (2011) also point out that it is more likely for a country to capitulate in a crisis when its opponent possesses nuclear weapons.[2] On the other hand, for the theoretical studies, in a seminal paper, Fearon (1994) models international crisises, including a nuclear proliferation crisis, as a two-player contest game and characterizes the equilibria. Although the main result of Fearon (1994) concerns the effects of domestic audience cost on international outcomes, the contest model employed has become one of the workhorse model for international crisis. For instance, Powell (2003) extends this model to study the incentives of small countries to acquire nuclear weapon and derives the implications on likelihood of war. Debs and Monteiro (2013) uses this model for studying the incentives for rising countries in a power-shifting scenario. Among other modeling frameworks, Benson & Wen (2011) studies a finite extensive form game and show that it is the threats together with the overriding interests that motivates one to develop nuclear weapons. Moreover, under certain conditions, the proliferators will resort to *strategic ambiguity*.[3] Baliga & Sjostrom (2008) also study a finite

[2]The coefficient of interest in their probit estimation is -0.602 and is significant with level 1%

[3]Benson & Wen showed that, when the ambiguous strategy is adopted (that is, in the mixed strategy equilibrium), it is more likely that the counter-proliferator will capitulate. Whether this equilibrium exists depends on the interests that proliferators could gain by acquiring nuclear weapons, how insecure the counter-

extensive form game with costly verification and shows that strategic ambiguity can yield a more stable outcome.

As for the effectiveness and consequences of counter-proliferation strategies, Mazarr (1995) highlights the futility of sanctions and the importance of package deals by investigating the case of North Korean nuclear crisis. Montgomery (2005) also emphasizes that providing economic aids to the proliferator in exchange for their capitulation is more effective than preventive strikes. It is also noteworthy that both of them argued that the deals offered to the proliferators must be equivalent to the gains of possessing nuclear weapons in terms of strategic interests. In addition, Goldstein (2006) argues that the asymmetry in nuclear power between proliferator and counter-proliferator will increase the likelihood of preventive strike. Feaver & Niou (1996) also show that whether a proliferator would attempt to acquire nuclear weapons and whether the counter-proliferator would assist their development or initiating a preventive strike depend on the prior belief of both on each other's type, their diplomatic relationships, the discrepancy of their military power and the degree to which counter-proliferators have already developed.

It is noteworthy that among the papers above above, Fearon (1994) and Powell (2003) are the closest to this paper in terms of modeling choices. Specifically, the models of both Fearon (1994) and Powell (2003) can be regarded as a brinkmanship model. Equivalently, they use a contest game with private values to model nuclear proliferation crisis. In contrast, the baseline model of this paper is a contest game with interdependent values. In particular, the proliferator's payoff also depend on the counter-proliferator's technology, which is private information to the proliferator. In terms of motivations, Benson & Wen (2011) is the closest to this paper, as they also consider the trade-off between tolerance and preventive strikes. In contrast to Benson & Wen (2011), this paper focuses on the effectiveness of each policy during the nuclear proliferation process and their impacts on security as well as proliferation outcomes, as opposed to how these different strategies affect the proliferator's incentive of acquiring nuclear weapons.

proliferators perceive, and the prior belief of the proliferator being "zealous".

3 The Baseline Model

3.1 Set up

In this section, we consider a specific model that describes the nuclear proliferation process. We analyze an incomplete information game with two players, each of them can choose a certain value of time. For the counter-proliferator, the time he chooses (t_1) represents the time he would tolerate whereas for the proliferator, the time being chosen (t_2) represents the time she would keep developing nuclear weapons. If the time that the proliferator is willing to tolerate is less than the time counter-proliferator would keep developing, then a preventive strike occurs. In contrast, if the time of tolerance is larger than the time of developing, then the counter-proliferator would capitulate, provided that she had not yet acquired the nuclear weapons. Finally, if the time of tolerance is large enough so that the proliferator could acquire nuclear weapons if she chooses to, then the development would be successful.

Formally, the model is set up as follows: The set of players is $\{1, 2\}$, where player 1 represents counter-proliferator and player 2 represents proliferator. The action spaces of both players are nonnegative real numbers $\mathbb{R}_+$, which stands for the time they choose. At the beginning, nature draws two values, c and τ according to CDF F_1 and F_2 respectively and reveals c to player 1 and τ to player 2. The value of c stands for the private information of player 1 about the cost of initiating a preventive strike,[4] while the value of τ stands for the threshold of time needed for a successful development, that is, player 2 will acquire nuclear weapons successfully if the time devoted into development if larger than τ, provided that a preventive strike had not yet occur until τ. After Nature's move, both players simultaneously choose $t_1 \geq 0$ and $t_2 \geq 0$. The outcomes and payoffs are then determined by the realizations of c and τ, as well as the chosen t_1 and t_2.

To model the costs and gains of tolerance, let $\pi_H(\cdot)$ and $\pi_L(\cdot)$ be two functions of time. For any $t \geq 0$, $\pi_H(t)$ represents the amount of costs (gains if negative) to the counter-proliferator, as well as the amount of gains (costs if negative) for the proliferator if the total time of tolerance is t and if there is no preventive strike at the end. On the other hand, $\pi_L(t)$ represents the costs/gains if the total time of tolerance is t and a preventive

[4] As in the international crisis literature, this can also be interpreted as the resolution level of the counter-proliferator.

strike occurs at the end. Furthermore, assume that the sum of gains/costs to the counter-proliferator and the proliferator is zero so that the values of $\pi_H(t), \pi_L(t)$ can be interpreted as the amount of transfers during the proliferation process. Notice that the possibility of different value of costs/gains in different outcomes means that the costs/gains when there is a preventive strike can be different from those when there is no preventive strike at the end. Such flexibility captures the feature that some of the costs/gains (e.g. economic aids or sanctions) can be retrieved when the counter-proliferator decides to initiate a preventive strike. Finally, let $w > 0$ be the cost to counter proliferator when the proliferator acquires nuclear weapons successfully, $s > 0$ denote the cost for proliferator of being attacked by the preventive strike and $\nu > 0$ denote the gain for the proliferator if she possesses nuclear weapons. Furthermore, we make the following assumptions:

Assumption 1. c and τ are independent and $c \sim F_1$, $\tau \sim F_2$. Furthermore, for each $i \in \{1, 2\}$, F_i has the following properties:

1. F_i has full support on $\mathbb{R}_+$.
2. F_i admits density $f_i > 0$.
3. F_i induces a random variable with finite second moment.

Assumption 2. The functions π_H and π_L are continuously differentiable.

To formalize the payoffs, let u_1 and u_2 be the (ex-post) payoffs of player 1 and 2 respectively. According to the scenarios described above, u_1 and u_2 are specified as follows:

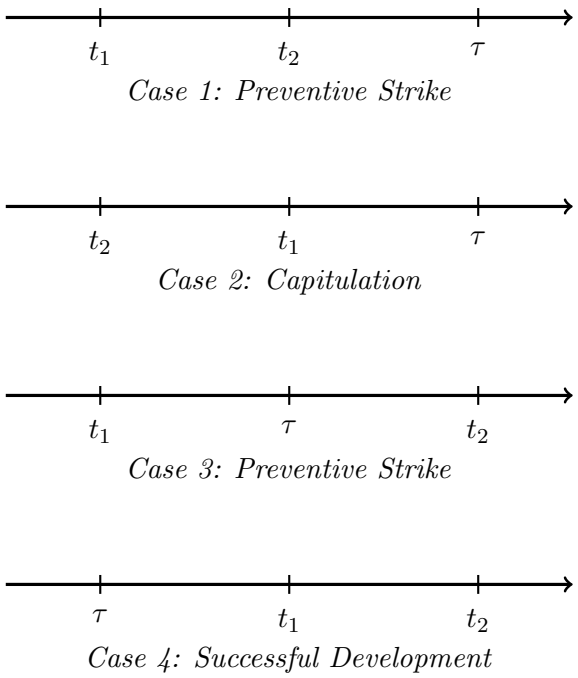
$$u_1(t_1, t_2, c, \tau) := \begin{cases} -\pi_H(t_2) & \text{if} \quad t_2 \leq \tau \quad \text{and} \quad t_1 \geq t_2 \\ -c - \pi_L(t_1) & \text{if} \quad t_2 \leq \tau \quad \text{and} \quad t_1 < t_2 \\ -\pi_H(\tau) - w & \text{if} \quad t_2 > \tau \quad \text{and} \quad t_1 \geq \tau \\ -c - \pi_L(t_1) & \text{if} \quad t_2 > \tau \quad \text{and} \quad t_1 < \tau \end{cases}. \tag{1}$$

$$u_2(t_1, t_2, \tau) := \begin{cases} \pi_H(t_2) & \text{if} \quad t_2 \leq \tau \quad \text{and} \quad t_1 \geq t_2 \\ \pi_L(t_1) - s & \text{if} \quad t_2 \leq \tau \quad \text{and} \quad t_1 > t_2 \\ \pi_H(\tau) + \nu & \text{if} \quad t_2 > \tau \quad \text{and} \quad t_1 \geq \tau \\ \pi_L(t_1) - s & \text{if} \quad t_2 > \tau \quad \text{and} \quad t_1 < \tau \end{cases}. \tag{2}$$

To sum up, given a value of τ, the decisions of two players constitute four cases. In the first case, $t_2 \leq \tau$ and $t_1 \geq t_2$, which means the proliferator chooses not to aim for successful

development and the counter-proliferator tolerates until the proliferator capitulates. In this case, the proliferator will not acquire nuclear weapons nor will there be any preventive strikes. In the second case, $t_2 \leq \tau$ and $t_1 < t_2$, which means that the proliferator does aim for successful development but the counter-proliferator initiates a preventive strike before the porliferator capitulates. In the third case, $t_2 > \tau$ and $t_1 \geq \tau$, which means that the porliferator aims for developing nuclear weapons and the counter-porliferator tolerates until the development is successful. In forth case, $t_2 > \tau$ and $t_1 < \tau$, which means that the proliferator aims for a successful development and the counter-proliferator initiates a preventive strike before the proliferator succeeds. These four cases are illustrated by the figure below:

Figure 1.



It is noteworthy that the model described above can be regarded as a contest game with interdependent values. More specifically, the time chosen by the players t_1 and t_2 can be interpreted as "bids", and the outcome is determined by which player's bid is higher. Unlike the standard brinkmanship model, the contest model above is of interdependent values since player 1's payoff also depends on player 2's private information τ. The (private) threshold τ captures the nature of the nuclear proliferation process: From the counter-proliferator's

perspective, it is uncertain how long it takes the proliferator for research and development before successfully acquiring nuclear weapons.

3.2 Equilibrium

In what follows, we first show that a Bayes Nash equilibrium equilibrium in which the proliferator uses a cutoff strategy always exists. Then we impose further conditions on π_H and π_L, as well as F_1, F_2 and derive more implications from this class of equilibria.

Proposition 1. *Under Assumption 1 and Assumption 2, there exists functions $\underline{\gamma} < \bar{\gamma}$, real numbers $0 \leq \underline{c} \leq \bar{c}$, and $0 \leq t^* \leq \hat{t}$, with $\bar{\gamma}(\bar{c}) = \hat{t}$ such that (t_1^*, t_2^*) is a Bayes Nash equilibrium, where*

$$t_1^*(c) = \begin{cases} \underline{\gamma}(c) & \text{if} \quad c \leq \underline{c} \\ \bar{\gamma}(c) & \text{if} \quad \underline{c} < c \leq \bar{c} \\ \hat{t} & \text{if} \quad c > \bar{c} \end{cases}, \tag{3}$$

$$t_2^*(\tau) = \begin{cases} \tau & \text{if} \quad \tau \leq \hat{t} \\ t^* & \text{if} \quad \tau > \hat{t} \end{cases}. \tag{4}$$

Without further specifications of π_H, π_L and F_1, F_2, we can already see some properties of the equilibrium strategies and their implications. The first implication is that for both players, the chosen time for tolerance and development are always bounded by $\hat{t}$, ruling out the possibility of the counter-proliferator's *appeasement* (i.e. always tolerate). For any realized cost of attack c, time of tolerance chosen by the counter-proliferator is always below $\hat{t}$, which is also the critical value for the proliferator between choosing to aim for a successful development ($t_2(\tau) = \tau$) or not ($t_2(\tau) = t^*$). Indeed, in equilibrium, with $\hat{t}$ being a common knowledge for both players, it is suboptimal for the counter-proliferator to tolerate more than $\hat{t}$ since shifting from tolerating $t' > \hat{t}$ to $\hat{t}$ makes no (ex-ante) difference for the counter-proliferator, regardless of the cost c. Second, it is noteworthy that the proliferator is essentially adopting a cutoff strategy. That is, when the realized value of technology threshold τ is less than the critical value $\hat{t}$, she will always aim for a successful development (i.e. $t_2(\tau) = \tau$), whereas when the threshold is greater than $\hat{t}$, she will not aim for a successful development and always choose some optimal development time t^* that is less than the threshold τ and is independent of τ. This suggests that whenever $t^* > 0$,[5] there

[5] In particular, when $\pi_H, \pi_L \geq 0$ and $\pi_H(0) = \pi_L(0)$.

are some cases when the proliferator is *not* aiming for successful development, yet due to the nature that this intention is observationally unidentifiable for the counter-proliferator—in the sense that he cannot observe the proliferator's choice $t_2 = t^* \geq 0$ and her private information τ so that he cannot distinguish between $t_2(\tau) = \tau$ and $t_2(\tau) = t^*$—it is possible that the preventive strike would be "unnecessary" in the sense that the proliferator has no intention to complete the development but the counter-proliferator still embarks a preventive strike.

In addition, if we further specify the properties of π_H, π_L and F_1, F_2, we would be able to derive more relevant implications from the equilibrium strategies. To see this, consider the following assumptions:

Assumption 3.

1. π_H, π_L are increasing and convex
2. For any $T > 0$, $i \in \{1, 2\}$, the function $\frac{F_i(T) - F_i(t)}{f_i(t)}$ is strictly decreasing on $[0, T]$. [6]

With Assumption 3, first notice that the first order conditions for the players' equilibrium strategies give:

$$\begin{aligned}
&\left(\underline{\gamma}(c) - t^*\right)\left(\pi_L'(\underline{\gamma}(c))\left(\frac{1 - F_2(\underline{\gamma}(c)}{f_2(\underline{\gamma}(c)}\right) + w - c\right) = 0,\\
&0 < \underline{\gamma}(c) \leq t^*,\\
&\left(\pi_L'(\underline{\gamma}(c))\left(\frac{1 - F_2(\underline{\gamma}(c)}{f_2(\underline{\gamma}(c)}\right) + w - c\right) \geq 0,
\end{aligned}$$

and

$$\begin{aligned}
&\left(\bar{\gamma}(c) - \hat{t}\right)\left(\pi_L'(\bar{\gamma}(c))\left(\frac{F_2(\hat{t}) - F_2(\bar{\gamma}(c)}{f_2(\bar{\gamma}(c)}\right) + w - c\right) = 0,\\
&t^* < \bar{\gamma}(c) \leq \hat{t},\\
&\left(\pi_L'(\bar{\gamma}(c))\left(\frac{F_2(\hat{t}) - F_2(\bar{\gamma}(c)}{f_2(\bar{\gamma}(c)}\right) + w - c\right) \geq 0.
\end{aligned}$$

With Assumption 3, it then follows that $\underline{\gamma}(c), \bar{\gamma}(c)$ are increasing in c and that $\bar{\gamma}(\bar{c}) = \hat{t}$. Therefore, the equilibrium strategy $t_1(c)$ is increasing and $t_1(c) = \hat{t}$ for any $c \geq \bar{c}$. Figure 2 plots the equilibrium strategies under Assumption 3.

[6] The exponential distribution, for example, satisfies this property.

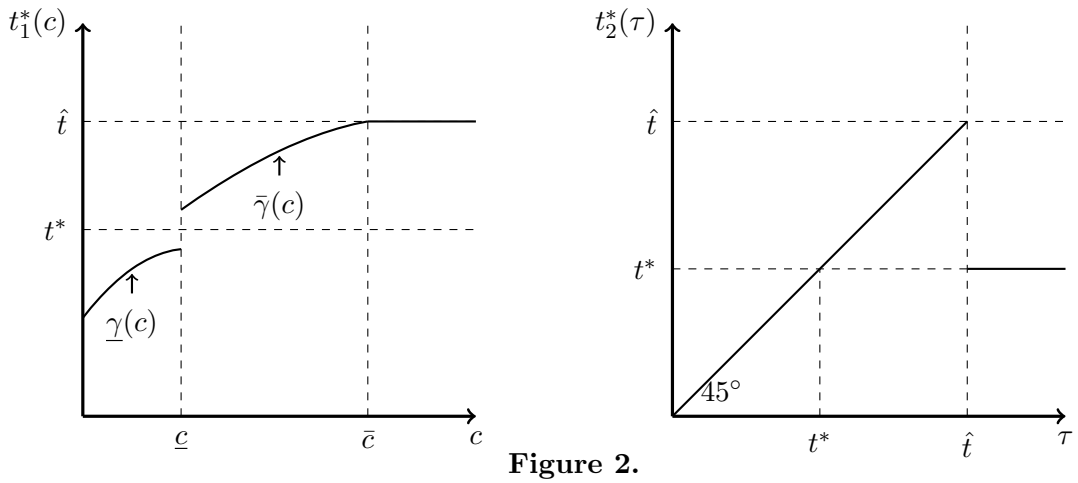


Figure 2.

Furthermore, combing the setups and the equilibrium strategies under assumptions 1,2 and 3, the type spaces for counter-porilferator and proliferator can be partitioned into segments as depicted in Figure 3. The first region indicates the combination of (c, τ) where the proliferator successfully obtains nuclear weapons. Qualitatively, this is when the cost for counter-proliferator is relatively high or the technology threshold of the proliferator is relatively low. The second region is where a preventive strike occurs in equilibrium. This occurs when the cost is relatively low or the threshold is relatively high. It is also noteworthy that when the cost is low enough and the threshold is high enough (specifically, when $c \leq \underline{c}$ and $\tau > \hat{t}$), such a preventive strike is unnecessary. The third region stands for the case where the proliferator capitulates, which occurs when both the cost and threshold are high.

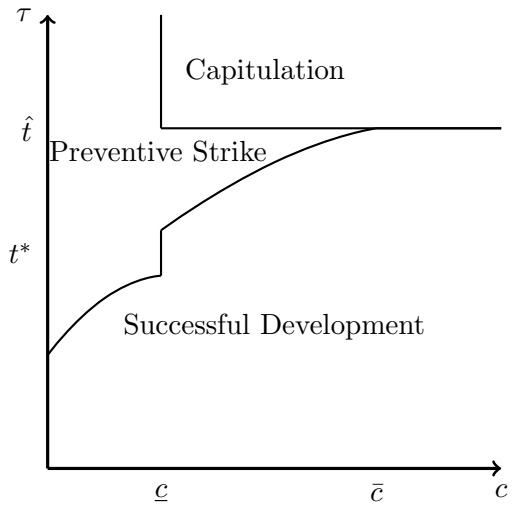


Figure 3.

Several implications can be derived from these equilibrium strategies. First, the toler-

ance time adopted by the counter-proliferator is increasing in the cost of conflicts. This property corresponds to the claim in the literature that when the relative military power between the two is more asymmetric (reflected by lower c), it is more likely for the counter-proliferator to initiate a preventive strike. Indeed, not only the tolerance time is lower when c is lower, the (interim) probability of preventive strike is also higher when c is lower, as seen in Figure 3. However, we should be cautious in generalizing this result since monotonicity of the counter-proliferator's strategies hinges on Assumption 3. That is, when the cost (gain) of tolerating is increasing (decreasing) in time for the counter-proliferator and the distribution of types have some certain properties in their hazard rates, counter-proliferator's tolerance time will be increasing in cost in equilibrium, but it is still unclear if the similar property holds in more general environments without Assumption 3.[7] Second, in the region where $c \geq \bar{c}$, the game can never end with a preventive strike. When $c \geq \bar{c}$, the equilibrium outcome will be either that the proliferator acquires nuclear weapons successfully or the proliferator capitulates. In these cases, the process of nuclear proliferation is "stable" in the sense that it will never lead to conflicts. Third, the unnecessary preventive strike can occur only when the cost is sufficiently small and threshold is sufficiently large ($c \leq \underline{c}$, $\tau > \hat{t}$).

Although the analyses above provide us some relevant implications, there is still a crucial limitation, as commonly seen in most of the game-theoretic models in the study of international conflicts. That is, the implications above hinges on certain specification of the game form and certain assumptions, and yet there might be many other specifications that are also suitable for modeling a nuclear proliferation process. For instance, one might argue that even if the counter-proliferator's tolerance time is less than the proiferator's development time, preventive strikes might still not occur for sure as the two parties should be able to negotiate or embark certain bargaining process at this moment. Alternatively, one might argue that the cost (gain) of tolerance can also be determined endogenously, rather than being given exogenously as functions π_H and π_L. After all, the level of economic aids and/or the severity of sanctions are mostly results of bargaining and negotiations in reality. To address this problem, we will consider a more general setup in the next section and will derive some results that are robust to the particular specifications of game forms and

[7] In fact, in more general settings, the interim probability of preventive strike is indeed nonincreasing in c, which will be shown in the next section.

detailed assumptions.

4 A General Model for Nuclear Proliferation

4.1 Mechanisms and the Revelation Principle

In this section, we consider a large class of possible game forms for nuclear proliferation and derive implications that are robust to the details of the models. As commonly applied the literature of mechanism design, we use the revelation principle to reduce the complexity inherited in detailed specifications of the game forms.[8] As a brief introduction to this methodology, we consider a class of games that capture the nature of scenario of nuclear proliferation. Specifically, we consider a class of games characterized by the rules that determines outcomes—the cost/gain of tolerance, the probability of a successful developments, and the probability of preventive strike. However, unlike the previous section, we do not specify how these outcomes are determined, nor the actions available to the players. Instead, we study the results that holds for *all* games with this structure. As noted above, this seemingly complex problem can be simplified by applying the revelation principle, which enables us to focus only on a (much smaller) class of incentive compatible direct mechanisms.

Formally, for $i \in \{1, 2\}$, let A_i denote the action space available for player i, T_i denote the type space for player i and let $A := A_1 \times A_2$, $T := T_1 \times T_2$. Also, let $\mathcal{O}$ denote the set of all possible *outcomes*, $\omega : A \times T \to \mathcal{O}$ is a map called *outcome function*, which determines the outcome given the player's type and their actions. Finally, player i's payoff is given by a von Neumann-Morgenstein utility function $u_i : T \times \mathcal{O} \to \mathbb{R}$. Together, a *mechanism* is defined by a tuple $\mathcal{M} := (A_i, T_i, u_i, \mathcal{O}, \omega)_{i=1}^2$. With the notations above, a *direct mechanism* is a mechanism with action spaces being identical to type spaces, denoted as $\mathcal{D} = (T_i, u_i, \mathcal{O}, \omega)_{i=1}^2$. A direct mechanism $\mathcal{D}$ is called *incentive compatible* if truth-telling is a Bayes Nash equilibrium. The revelation principle is then stated as follows:

Theorem 1 (Revelation Principle)**.** *For any mechanism $\mathcal{M} := (A_i, T_i, u_i, \mathcal{O}, \omega^{\mathcal{M}})_{i=1}^2$ and a Bayes Nash equilibrium $\sigma^* = (\sigma_1^*, \sigma_2^*)$ in $\mathcal{M}$, there exists an incentive compatible direct mechanism $\mathcal{D} = (T_i, u_i, \mathcal{O}, \omega)_{i=1}^2$ such that for any $t = (t_1, t_2) \in T$, the equilibrium outcome in $\mathcal{D}$ is the same as in $\mathcal{M}$, that is: $\omega(t, t) = \omega^{\mathcal{M}}(\sigma^*(t), t), \forall t \in T$.*

[8] In some literature of international conflicts, this approach is called *game-free* analysis, as in Banks(1990) and Fey & Ramsey (2009, 2011)

4.2 Setups, Incentive Compatibility and Individual Rationality

As noted in the previous section, the baseline model has a substantial limitation: The results and implications derived above might depend on the specific form and assumptions in the settings. To obtain results that are robust to model specifications, we now consider a general class of models that describe the nature of nuclear proliferation. As it can be seen from the model in the previous section, as well the existing literature, the essence of the nuclear proliferation is that there is a proliferator attempting to develop nuclear weapons and a counter-proliferator who can either tolerate, possibly together with sanctions and/or granting economic incentives, with the hope that the proliferator would capitulate, or initiate a preventive strike, in order to eliminate the possibility of further developments. As a result, a nuclear proliferation game can be characterized by a combination of outcomes consisting of the (ex-post) probability of conflicts, the (ex-post) probability of a successful developments, and the cost/gain to the counter-proliferator/proliferator for different outcomes that also depends on the duration of the process. Two critical assumptions we have here is that the game ends once the preventive strike is initiated or the proliferator capitulates and that whenever there is a preventive strike, the proliferator cannot obtain nuclear weapons successfully. Thus, in any such model, the outcome can be partitioned into three scenarios—"preventive strike", "successful development", and "capitulate".

To formally setup a general model, let A_1, A_2 be some abstract (non-empty) action spaces for the players. For technical concerns, for each $i \in \{1, 2\}$ we also let $\mathcal{A}_i$ be a sigma-algebra on A_i and μ_i be a measure on $(A_i, \mathcal{A}_i)$. This allows us to capture any specifications about the strategic choices for the players, including the choices of tolerance and initiating a preventive strike, the choices for the development strategy, as well as all the possible actions that can be taken in bargaining or negotiating processes.[9] For instance, in the baseline model $A_1 = A_2 = \mathbb{R}_+$ are the sets of time chosen by players, $\mathcal{A}_1, \mathcal{A}_2$ are then the *Borel algebra* on $\mathbb{R}_+$ and μ_1, μ_2 are both Lebesgue measures. One can also think of A_i as $\mathbb{R}_+ \times B_i$, where for any $(t_i, b_i) \in A_i$, t_i denotes the time chosen by player i and b_i denotes the combination of all other actions adopted by player i in a bargaining or negotiating process, for $i \in \{1, 2\}$. We do not impose any further assumptions on the set $A := A_1 \times A_2$. However, we maintain the specifications about the information structure. That is, the type spaces of

[9] Furthermore, by considering only an abstract action space, we allow the model to be an extensive form game, by reducing the extended form into strategic form with action spaces A_1, A_2.

players are $T_1 = T_2 = \mathbb{R}_+$, where $c \in T_1$ denotes the private information for player 1 and $\tau \in T_2$ denotes the private information for player 2, c, τ are independent and that c is drawn from F_1 and τ is drawn from F_2. We also maintain Assumption 1 as basic assumptions on F_1, F_2.

On the other hand, consider three functions $\pi_H : A \times T_2 \to \mathbb{R}, \pi_L : A \times T_2 \to \mathbb{R}$ and $\boldsymbol{g} : A \times T_2 \to \Delta(\{1,2,3\})$, where $\Delta(\{1,2,3\})$ denotes the all probability measures on the finite set $\{1,2,3\}$ that stands for the three possible outcomes:"capitulate", "successful development" and "preventive strike", respectively. It is sometimes useful to write $\boldsymbol{g} = (g_1, g_2, g_3) \in [0,1]^3$, with $g_1 + g_2 + g_3 = 1$. The outcome space of this class of nuclear proliferation games is then $\mathbb{R}^2 \times \Delta(\{1,2,3\})$ and $(\pi_H, \pi_L, \boldsymbol{g})$ is the outcome function. With these notations, π_H, π_L then denote the amount of transfers and g_1, g_2, g_3 denote the probability of "capitulate", "successful development" and "preventive strike", respectively. Several remarks for this setting are noteworthy. First, by allowing the range of $\boldsymbol{g}$ to be a probability distribution, we allow the possibility that the outcomes are randomly selected by a distribution, given a combination of players' actions, rather than being deterministic. Second, by allowing the range of π_H, π_L being real numbers, both "sanctions" (negative) and "economic incentives" (positive) are captured. Third, by allowing the functions to depend on actions, it allows the amount of transfers, as well as the probability distribution of possible outcomes to be endogenously determined by the interactions between players and hence include all possible models with negotiations, bargaining and choices between sanctions and economic incentives. Finally, although it will not change the results derived below if the outcome function depends also on c, for simplicity we only allow $\pi_H, \pi_L, \boldsymbol{g}$ to be dependent only on T_2. To rule out some pathological cases, we impose the following mild condition on the functions π_L and π_H.

Assumption 4.

1. For each $a_1 \in A_1$, $\int_{A_2} \int_0^\infty |\pi_j(a_1, a_2, \tau)| dF_2(\tau) d\mu_2(a_2) < \infty$ for $j \in \{H, L\}$.

2. For each $a_2 \in A_2$ and $\tau \geq 0$, $\int_{A_1} |\pi_j(a_1, a_2, \tau)| d\mu_1(a_1) < \infty$ for $j \in \{H, L\}$.

Additionally, due to the nature of nuclear proliferation process, it is not reasonable to rule out the possibility that the counter-proliferator can always choose to initiate a preventive strike unilaterally, and that the proliferator can always choose not to start developing. To capture this nature, we impose the following assumptions:

Assumption 5.

1. There exists $\tilde{a_1}$ such that, $\pi_L(\tilde{a_1}, a_2, \tau) = 0$ and $g_3(\tilde{a_1}, a_2, \tau) = 1$, $\forall a_2 \in A_2$, $\tau \in T_2$.[10]

2. There exists $\tilde{a_2}$ such that $\pi_L(a_1, \tilde{a_2}, \tau) = 0$ and $g_1(a_1, \tilde{a_2}, \tau) = 1$, $\forall a_1 \in A_1$, $\tau \in T_2$.

We also maintain the structure of payoffs in the baseline model in previous section. That is, when the outcome is "capitulate", the counter-proliferator gets $-\pi_H$ and the proliferator gets π_H; when the outcome is "successful development", the counter-proliferator gets $-\pi_H - w$ and the proliferator gets $\pi_H + \nu$; when the outcome is "preventive strike", the counter-proliferator gets $-\pi_L - c$ and the proliferator gets $\pi_L - s$, where $w, s, \nu > 0$. With the notations above, the payoffs can be written as

$$u_1(a_1, a_2, c, \tau) = -\left[(\pi_L(a_1, a_2, \tau) + c)g_3(a_1, a_2, \tau) + \pi_H(a_1, a_2, \tau)(1 - g_3(a_1, a_2, \tau)) + wg_2(a_1, a_2, \tau)\right],$$

$$u_2(a_1, a_2, \tau) = (\pi_L(a_1, a_2, \tau) - s)g_3(a_1, a_2, \tau) + \pi_H(a_1, a_2, \tau)(1 - g_3(a_1, a_2, \tau)) + \nu g_2(a_1, a_2, \tau)$$

To sum up, with the notations introduced above, a nuclear proliferation game can be written as $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$, satisfying assumptions 4 and 5. Hereafter, we refer a nuclear proliferation game $\mathcal{P}$ as a tuple $(A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$ that satisfies assumptions 4 and 5. It is noteworthy that the baseline model in section 3, as well as the brinkmanship model (i.e. contest games with private values) in Fearon (1994) and Powell (2003) can all be written as nuclear proliferation game $\mathcal{P}$.

As defined above, a direct mechanism in this environment is given by $\mathcal{D} = (\pi_H, \pi_L, \boldsymbol{g})$, with $\pi_H : T \times T_2 \to \mathbb{R}$, $\pi_L : T \times T_2 \to \mathbb{R}$ and $\boldsymbol{g} : T \times T_2 \to \Delta(\{1, 2, 3\})$. In the analyses below, given any nuclear proliferation game $\mathcal{P}$ and a Bayes Nash equilibrium σ^* in $\mathcal{P}$, it is useful to consider the *interim* expected payoffs. That is, $U_1^{\mathcal{P}}(c|\sigma^*) := \mathbb{E}_{F_2}[u_1(\sigma_1^*(c), \sigma_2^*(\tau), c, \tau)]$ and $U_2^{\mathcal{P}}(\tau|\sigma^*) := \mathbb{E}_{F_1}[u_2(\sigma_1^*(c), \sigma_2^*(\tau), \tau)]$. Under Assumption 5, in any Bayes Nash equilibrium, both players should have interim expected payoffs no less than the payoff they can obtain unilaterally—always initiate a preventive strikes or always not for counter-proliferator and not to develop at all for the proliferator. This property, which is often referred as *individual rationality*, is defined formally below.

Definition 1. A nuclear proliferation game $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$, together with a Bayes

[10] Although the setups in the baseline model in section 3 does not exhibit this property, it can be modified by extending the action space of player 1 to extended real numbers so that ∞ can be player 1's choice.

Nash equilibrium σ^*, is said to be individually rational if:

$$U_1^{\mathcal{P}}(c|\sigma^*) \geq -c, \forall c \geq 0,$$

and

$$U_2^{\mathcal{P}}(\tau|\sigma^*) \geq 0, \forall \tau \geq 0.$$

Moreover, as mentioned above, a direct mechanism is *incentive compatible* if the identity function is an equilibrium strategy for each player. That is, each player would have no incentive to misreport their types given that the other is telling the truth. This is also defined formally as follows:

Definition 2. A direct mechanism $\mathcal{D} = (\pi_H, \pi_L, \boldsymbol{g})$ is incentive compatible if

$$\mathbb{E}_{F_2}[u_1(c, c, \tau, \tau)] \geq \mathbb{E}_{F_2}[u_1(c', c, \tau, \tau)], \forall c', c \geq 0,$$

and

$$\mathbb{E}_{F_1}[u_2(c, \tau, \tau)] \geq \mathbb{E}_{F_1}[u_2(c, \tau', \tau)], \forall \tau', \tau \geq 0.$$

Furthermore, in the analyses below, it is useful to consider the interim expected payoffs $U_1(c, c') := \mathbb{E}_{F_2}[u_1(c', c, \tau, \tau)]$ and $U_2(\tau', \tau) := \mathbb{E}_{F_1}[u_2(c, \tau', \tau)]$, which denotes the interim expected payoffs of reporting their types as c', τ' when their actual types are c, τ in the direct mechanism, respectively. We also write $U_1^*(c) := U_1(c, c)$ and $U_2^*(\tau) := U_2(\tau, \tau)$, which is the interim expected payoffs when reporting their types truthfully, given that the other is reporting truthfully. With these notations, a direct mechanism $\mathcal{D} = (\pi_H, \pi_L, \boldsymbol{g})$ is incentive compatible and individually rational if and only if

$$U_1^*(c) \geq U_1(c', c), \quad U_1^*(c) \geq -c, , \forall c, c' \geq 0$$

and

$$U_2^*(\tau) \geq U_2(\tau', \tau), \quad U_2^*(\tau) \geq 0, \forall \tau, \tau' \geq 0.$$

We now begin analyzing the equilibria in the class of nuclear proliferation games satisfying assumptions 4 and 5.

4.3 Main Results and Implications

As noted above, although it seems difficult to examine the properties of equilibria in a large class of games due to their complexity and generality, the revelation principle can be

applied and it is without loss to restrict our attentions to the class of incentive compatible direct mechanisms. Formally, for any nuclear proliferation game $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$ and any Bayes Nash equilibrium σ^* in $\mathcal{P}$, by the revelation principle, there exists an incentive compatible direct mechanism $\mathcal{D} = (\pi_H, \pi_L, \boldsymbol{g})$ such that

$$\pi_H^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau) = \pi_H(c, \tau, \tau)$$
$$\pi_L^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau) = \pi_L(c, \tau, \tau)$$
$$\boldsymbol{g}^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau) = \boldsymbol{g}(c, \tau, \tau),$$

for all $c, \tau \geq 0$. It then follows that $U_1^{\mathcal{P}}(c|\sigma^*) = U_1^*(c)$ and $U_2^{\mathcal{P}}(\tau|\sigma^*) = U_2^*(\tau)$, $\forall c, \tau \geq 0$ and hence $(\mathcal{P}, \sigma^*)$ is individually rational if and only if the corresponding incentive compatible direct mechanism $\mathcal{D}$ is individually rational. Therefore, to study the class of nuclear proliferation games $\mathcal{P}$ and its equilibria σ^* is equivalent to study the class of direct mechanisms $\mathcal{D}$ that are incentive compatible and individually rational. This observation will be useful for the analyses below.

The first result for the general model of nuclear proliferation considers the property of equilibirum payoff of the counter-proliferator and the probability of preventive strike as well as the relation between them. The following lemma is the key to these results.

Lemma 1. *A direct mechanism $\mathcal{D} = (\pi_H, \pi_L, \boldsymbol{g})$ is incentive compatible if and only if*

1. *$U_1^*(c) = U_1^*(0) - \int_0^c \mathbb{E}_{F_2}[g_3(y, \tau, \tau)]dy$ for all $c \geq 0$.*
2. *$\mathbb{E}_{F_2}[g_3(c, \tau, \tau)]$ is nonincreasing in in c.*

From Lemma 1 and the revelation principle, we are then able to characterize the equilibrium payoff of the counter-proliferator in any nuclear proliferation games. In addition, together with individual rationality, it can be shown that the interim probability of preventive strike exhibits certain properties that coincide with the results in the baseline model.

Proposition 2. *Let $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$ be any nuclear proliferation game and σ^* be a Bayes Nash equilibrium in $\mathcal{P}$. Then,*

1. *$U_1^{\mathcal{P}}(c|\sigma^*) = U_1^{\mathcal{P}}(0|\sigma^*) - \int_0^c \mathbb{E}_{F_2}[g_3^{\mathcal{P}}(\sigma_1^*(y), \sigma_2^*(\tau), \tau)]dy, \forall c \geq 0$. In particular, equilibrium payoff of player 1, $U_1^{\mathcal{P}}(c|\sigma^*)$, depends only on the interim probability of preventive strike $\mathbb{E}_{F_2}[g_3^{\mathcal{P}}]$ in equilibrium up to a constant.*

2. $\mathbb{E}_{F_2}[g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)]$ *is nonincreasing in* c.

3. $\lim_{z\to\infty} \int_z^\infty \mathbb{E}_{F_2}[g_3^{\mathcal{P}}(\sigma_1^*(y), \sigma_2^*(\tau), \tau)]dy = 0$.

In the previous section, we showed that under assumptions 1, 2 and 3, the interim probability of preventive strike is nonincreasing in the cost of counter-proliferator c and is approaching to 0 as c increases (in particular, is 0 whenever $c \geq \bar{c}$). This result echos with the common observation in literature that the more asymmetric the military capability between the two is, the more likely there would be preventive strikes and hence the relationship is more unstable. Proposition 2 generalized this result and indicates that it is robust to all the Bayes Nash equilibria in all nuclear proliferation games. Specifically, assertion 2 implies that in *any* Bayes Nash equilibrium of *any* nuclear proliferation game, the interim probability of preventive strike, as a function of c, is always decreasing. Moreover, assertion 3 implies that this interim probability is not only decreasing, but is approaching to zero quite "rapidly" in *any* Bayes Nash equilibrium of *any* nuclear proliferation game, in the sense that $\lim_{z\to\infty} \int_z^\infty \mathbb{E}_{F_2}[g_3^{\mathcal{P}}(\sigma_1^*(y), \sigma_2^*(\tau), \tau)]dy = 0$ implies $\lim_{c\to\infty} \mathbb{E}_{F_2}[g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)] = 0$.[11]

Furthermore, assertion 1 of Proposition 2 provides a characterization of the equilibrium payoffs of the counter-proliferator in any nuclear proliferation game. In particular, in *any* Bayes Nash equilibrium of *any* nuclear proliferation game, the interim equilibrium payoff of the counter-proliferator is always nonincreasing in c. It also implies that the interim equilibrium payoff depends only on the interim probability of preventive strike up to a constant. Substantively, this suggests that no matter what the bargaining and negotiating processes, as well as the degree of sanctions and economic aids are, counter-proliferator will always get the same interim expected payoff up to a constant, provided that the interim probability of preventive strike is the same in such equilibrium.

In addition to the equilibrium payoff of the counter-proliferator and the property of the interim probability of preventive strike, in any Bayes Nash equilibrium of any nuclear proliferation game, there are other crucial outcomes that can be investigated, including the costs and the probability of preventive strike as well as the cost/gain and the probability of a successful development more widely. These are presented in the next proposition.

Proposition 3. *Let* $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$ *be any nuclear proliferation game and* σ^* *be a*

[11] So that the result in baseline model that $g_3(c, \tau, \tau) = 0$ whenever $c \geq \bar{c}$ is a special case of assertion 3.

Bayes Nash equilibrium in $\mathcal{P}$. Then:

$$\begin{aligned}
&\mathbb{E}[(c+s)g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)+(w-\nu)g_2^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)]\\
&\quad\leq -U_1^{\mathcal{P}}(0|\sigma^*)+\mathbb{E}\left[\left(\frac{1-F_1(c)}{f_1(c)}\right)g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)\right]\\
&\quad\leq \mathbb{E}[c],
\end{aligned}$$

where the expectation is taken under the joint distribution $F=F_1F_2$ and $U_1^{\mathcal{P}}(0|\sigma^*)\geq 0$. In particular,

$$\mathrm{Cov}\left(c-\frac{1-F_1(c)}{f_1(c)},g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)\right)\leq \mathbb{E}[(w-\nu)g_2^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)-sg_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)].$$

Proposition 3 gives another property of equilibrium outcomes of any nuclear proliferation game satisfying assumptions 4 and 5. The first implication can be seen from the inequality

$$\mathbb{E}[(c+s)g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)+(w-\nu)g_2^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)]\leq \mathbb{E}[c]<\mathbb{E}[c+s],$$

which states that the expected *total cost* in such nuclear proliferation game is always less than the expected total cost of preventive strike. This implies that by allowing the possibility of a successful development and the bargaining process, the nuclear proliferation game yields a more efficient outcome than the counter-proliferator initiating a preventive strike alone. Namely, it is more beneficial overall for the two sides to bargain rather than the counter-proliferator initiating a preventive strike unilaterally. Moreover, such efficiency gains are uniform in the sense that not only the expected total cost is less than overall total cost of preventive strike, it is always no greater than the expected cost of preventive strike for the counter-proliferator alone. Secondly, the inequalities impose some constraints on $g_3^{\mathcal{P}}\circ\sigma^*$ and $g_2^{\mathcal{P}}\circ\sigma^*$, given values of w,ν,s and the form of F_1. Substantively, in addition to the property given by Proposition 2 that the interim probability of preventive strike, as a function of c, must converge to zero "rapidly" as c tends to infinity, the constraints imposed by the above inequality further indicate that the ex-ante probability of preventive strike must also not be "too large". Moreover, from the inequality

$$\mathbb{E}[(c+s)g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)+(w-\nu)g_2^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)]\leq \mathbb{E}[c],$$

this restriction is more stringent as s increases and $w-\nu$ decreases. In addition, from the inequality

$$-U_1^{\mathcal{P}}(0|\sigma^*)+\mathbb{E}\left[\left(\frac{1-F_1(c)}{f_1(c)}\right)g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)\right]\leq \mathbb{E}[c],$$

this restriction depends also on the distribution of c. Specifically, the larger the expected *hazard rate* of c is, the more strigent the restriction would be. Finally, the inequality

$$\text{Cov}\left(c-\frac{1-F_1(c)}{f_1(c)}, g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)\right) \leq \mathbb{E}[(w-\nu)g_2^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau) - s g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)]$$

implies that, in addition to the property that the interim probability of preventive strike is nonincreasing in c in any nuclear proliferation game, the (linear) relationship between such probability and the *virtual value*, $c-\frac{1-F_1(c)}{f_1(c)}$ is also bounded and, specifically, is negative when $w-\nu$ is relatively smaller. Moreover, if $c-\frac{1-F_1(c)}{f_1(c)}$ is increasing in c (as commonly assumed is most of the classic literature in mechanism design) and $w-\nu$ is relatively small, this inequality can be regarded as another form of monotonicity of $g_3^{\mathcal{P}} \circ \sigma^*$. Further implications of these results and their connections with the literature will be discussed in the next section.

While Proposition 3 considers the equilibrium payoff of counter-proliferator as well as the interim probability of preventive strike in any Bayes Nash equilibrium of any nuclear proliferation game, it is also crucial to understand whether one can achieve an outcome where either preventive strike and/or successful development is impossible. The study about the probability of preventive strike is crucial since it directly relates to the likelihood of conflicts and hence to the debate between optimists and pessimists. Likewise, the study about the probability of a successful development is substantively relevant as it is not only related to the interest of counter-proliferator, but also to the possibility of encouraging or discouraging non-nuclear states to acquire nuclear powers.

The two results below will characterize the necessary and sufficient conditions for the (ex-post) probability of successful development and preventive strike to be zero respectively and examine what is needed (and what is not needed) to obtain such outcomes in any nuclear proliferation game.

Proposition 4. *Let $\mathcal{P}=(A,\pi_H^{\mathcal{P}},\pi_L^{\mathcal{P}},\boldsymbol{g}^{\mathcal{P}})$ and σ^* be any pair of nuclear proliferation game and its Bayes Nash equilibrium. If $g_3^{\mathcal{P}}\equiv 0$, then*

$$w \leq -\frac{\mathbb{E}[\pi_H^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)]}{\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau))]} \leq \nu$$

when $\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)]>0$ and

$$\mathbb{E}[\pi_H^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)]=0$$

when $\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)] = 0$*, where the expectation operator is taken under the joint distribution* $F = F_1 F_2$*.*

Proposition 4 gives a necessary condition for a Bayes Nash equilibria in any nuclear proliferation game to have zero probability of preventive strike. This enables us to examine the essence of the proliferation process whenever there is an equilibrium under which preventive strikes are impossible and, in the same way, identifies all the specifications (including bargaining process, payoff structures, outcomes etc.) in any nuclear proliferation games that are impossible to yield an equilibrium with zero probability of preventive strikes. To see this, consider first a Bayes Nash equilibrium σ^* in any nuclear proliferation game $\mathcal{P}$ with $\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau] > 0$. Proposition 4 then requires

$$w \leq -\frac{\mathbb{E}[\pi_H^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)]}{\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)} \leq \nu.$$

Therefore, in any Bayes Nash equilibrium with zero probability of preventive strike, it must be that $w \leq \nu$. That is, in *any* nuclear proliferation game with $w > \nu$, there cannot be any Bayes Nash equilibrium that has zero probability of preventive strike. Substantively, this result means that whenever the lost of the counter-proliferator is greater than the gain of the proliferator after the proliferator acquires nuclear weapons, no matter what the bargaining and negotiating process, as well the choices available to the two players, there must be some cases (some region on the $c - \tau$ space) that a preventive strike occurs.

In addition, the above inequality also implies that

$$-\nu\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)] \leq \mathbb{E}[\pi_H^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)] \leq -w\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)].$$

In particular, since $\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)] > 0$ and $w, \nu > 0$ by definition, this implies $\mathbb{E}[\pi_H^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)] < 0$. Substantively, this means that to achiecve an equilibrium with zero probability of preventive strike, the expected transfer throughout the proliferation process must be negative and must be bounded by two values, $-\nu\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)]$ and $-w\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)]$. As we can interpret the value of $\pi_H^{\mathcal{P}}$ as the net value of economic incentives given by counter-proliferator combined with the amount of sanctions imposed on the proliferator, it then follows that in any equilibrium where preventive strikes never occur, sanctions must be conducted and must be on average more salient than all the economic aids to the proliferator in equilibrium. Therefore, for the counter-proliferator, to achieve such equilibrium outcomes, sanction must be adopted and must exceed certain degree. Using economic incentives alone can never reach equilibria without preventive strikes.

On the other hand, for any Bayes Nash equilibrium σ^* in any nuclear proliferation game $\mathcal{P}$ with $\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau] = 0$, it must be that $\mathbb{E}^{\mathcal{P}}[\pi_H(\sigma_1^*(c), \sigma_2^*(\tau), \tau)] = 0$. Since $g_2^{\mathcal{P}} \geq 0$ and hence $\mathbb{E}[g_2^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau] = 0$ if and only if $g_2^{\mathcal{P}} \circ \sigma^* = 0$ (almost everywhere), in any Bayes Nash equilibrium of nuclear proliferation game with both the probability of preventive strike and successful development being zero, it must be that the expected gain/loss during the process is zero, which can be regarded as a degenerate case since neither player gets any net benefits nor do they incur and costs.

In fact, if we notice that in any direct mechanism, the incentive constraint for player 2,

$$\begin{aligned} U_2^*(\tau) &= \mathbb{E}_{F_1}[(\pi_L(c,\tau,\tau) - s)g_3(c,\tau,\tau) + \pi_H(c,\tau,\tau)(1 - g_3(c,\tau,\tau)) + \nu g_2(c,\tau,\tau)] \\ &\geq \mathbb{E}_{F_1}[(\pi_L(c,\tau,\tau') - s)g_3(c,\tau,\tau') + \pi_H(c,\tau,\tau')(1 - g_3(c,\tau,\tau')) + \nu g_2(c,\tau,\tau')] \\ &= U_2(\tau,\tau'), \forall \tau, \tau' \geq 0 \end{aligned}$$

has one free parameter up to the choice in defining a direct mechanism because reporting τ' and his type τ affect his payoff only through the mechanism, it then follows that the condition in proposition 3 is a necessary and sufficient condition for existence of equilibrium with zero probability of preventive strike, as stated and proved below.

Corollary 1. *There exists a nuclear proliferation game and a Bayes Nash equilibrium that yields zero ex-post probability of preventive strike and non-zero (with positive measure) ex-post probability of a successful development if and only if* $0 < w \leq \nu$.

In addition to the probability of preventive strike, it is also crucial to examine the probability of a successful development in equilibria of nuclear proliferation games. While the study of the probability of preventive strike is related to the *desirability* about the impact of nuclear proliferation on stability, the study of the probability of a successful development concerns the *feasibility* that whether the proliferator can be discouraged from obtaining nuclear weapons in equilibrium. As Proposition 4, the proposition below gives a necessary condition for any Bayes Nash equilibrium in any nuclear proliferation game that exhibits zero probability of a successful development as its outcome.

Proposition 5. *Let* $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$ *and* σ^* *be any pair of nuclear proliferation game*

its Bayes Nash equilibrium. If $g_2^{\mathcal{P}} \equiv 0$*, then:*

$$\begin{aligned}
&\mathbb{E}[sg_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)] \\
\leq &\mathbb{E}[\pi_L^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau) g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau) + \pi_H^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)(1 - g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau))] \\
\leq &\mathbb{E}[c(1 - g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau))],
\end{aligned}$$

where the expectation operator is taken under the joint distribution $F \equiv F_1 F_2$*.*

To investigate the implications of Proposition 5, first notice that the term

$$\pi_L^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau) g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau) + \pi_H^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)(1 - g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau))$$

can be interpreted as the ex-post average transfer throughout the bargaining process. Therefore, the first inequality in Proposition 5 implies that, in any Bayes Nash equilibrium of any nuclear proliferation game with zero probability of a successful development, the expected transfer must be nonnegative. Contrary to the condition in Proposition 4, this inequality implies that to achieve an equilibrium with zero probability of a successful development, the economic aids given by the counter-proliferator must outweigh the sanctions in expectation and therefore using sanctions alone could never achieve equillibria with zero probability of a successful development. In fact, combining Proposition 4 and Proposition 5, we can see an essential trade-off between two counter-proliferation strategies—sanctions versus economic aids. The results above imply that, to achieve an always peaceful outcome, sanctions must prevail. In contrast, economic aids must prevail in order to discourage successful development.

Furthermore, aside from being positive, the inequality given in this proposition shows that the expected average transfer is bounded by

$$\mathbb{E}[sg_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)]$$

and

$$\mathbb{E}[c(1 - g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau))].$$

As the lower bound represents the proliferator's expected loss of being attacked, it follows that, if the counter-proliferator wishes to discourage the proliferator from acquiring nuclear weapons successfully, unlike the conventional wisdom which claims that the economic aids must exceed the potential benefit ν, the relevant benchmark for economic aids should be the expected loss for being attacked instead.

Finally, the inequality

$$\mathbb{E}[sg_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)] \leq \mathbb{E}[c(1-g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau))]$$

also gives a constraint on the probability of preventive strike. Qualitatively, this means that in any Bayes Nash equlilibrium with zero probability of a successful development, the probability of preventive strike must also be small enough on average.

On the other hand, for any Bayes Nash equilibrium σ^* in any nuclear proliferation game $\mathcal{P}$ with $g_2^{\mathcal{P}} \circ \sigma^* \equiv 0$, Proposition 3 can be simplified as follows:

Corollary 2. *Let $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$ and σ^* be any pair nuclear proliferation game and its Bayes Nash equilibrium. If $g_2^{\mathcal{P}} \equiv 0$, then*

$$\mathbb{E}[(c+s)g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)] \leq -U_1^{\mathcal{P}}(0|\sigma^*) + \mathbb{E}\left[\left(\frac{1-F_1(c)}{f_1(c)}\right) g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)\right] \leq \mathbb{E}[c],$$

where the expectation operator is taken under the joint distribution $F = F_1F_2$ and $U_1^{\mathcal{P}}(0|\sigma^*) \geq 0$. In particular,

$$\mathrm{Cov}\left(c - \frac{1-F_1(c)}{f_1(c)}, g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)\right) \leq -s\mathbb{E}[g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)].$$

Proof. It follows directly from proposition 3 by taking $g_2^{\mathcal{P}} \circ \sigma^* \equiv 0$. ■

Corollary 2 specializes the constraints imposed by Proposition 3 in equilibria with zero probability of a successful development. This further leads to several necessary conditions for such equilibria. First,

$$\mathbb{E}[(c+s)g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)] \leq -U_1^{\mathcal{P}}(0|\sigma^*) + \mathbb{E}\left[\left(\frac{1-F_1(c)}{f_1(c)}\right) g_3^{\mathcal{P}}(\sigma_1^*(c),\sigma_2^*(\tau),\tau)\right]$$

imposes a condition on the probability of preventive strike in equilibrium that depends on the reciprocal of hazard rate of c. In particular, this implies that under certain distributions, it is not possible to achieve an equilibrium with zero probability of successful development but non-zero probability of preventive strike.[12] Second, from both inequalities, there is an implicit upper bound on the probability of preventive strike in such equilibria. This property indicates that, although it seems that the probability of preventive strike and the probability of a successful development are substitutes to each other, as preventive strikes

[12]For example, consider the case where c is exponentially distributed with mean $\frac{1}{\lambda} < s$. Then $\frac{1-F_1(c)}{f_1(c)} = \frac{1}{\lambda}$ and hence the first inequality cannot be satisfied for any function $g_3 \geq 0$, since $U_1^*(0) \geq 0$.

are used to hinder successful developments, they are sometimes positively correlated in the sense that only when the probability of preventive strike is small enough could zero probability of a successful development be possible in equilibrium. Finally, the inequality

$$\text{Cov}\left(c - \frac{1-F_1(c)}{f_1(c)}, g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)\right) \leq -s\mathbb{E}[g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)]$$

shows that aside from being decreasing as a function of c in equilibrium, the probability of preventive strike must be negatively correlated with the virtural value $c - \frac{1-F_1(c)}{f_1(c)}$. Further implications of these results and its connection with the literature will be discussed in the next section.

Similar to Proposition 4, since the incentive constraints for player 2 has one free parameter, the above characterizations are also in fact sufficient for existence of a nuclear proliferation game with an individually rational Bayes Nash equilibrium such that the probability of preventive strike is zero and the probability of preventive strike is non-zero.

Corollary 3. *There exists a nuclear proliferation game and a Bayes Nash equilibrium with zero ex-post probability of a successful development and non-zero (with positive measure) ex-post probability of preventive strike in equilibrium if and only if there exists $\kappa \leq 0$, and a non-zero (with positive measure), decreasing function $\beta : \mathbb{R}_+ \to [0,1]$ such that*

$$\mathbb{E}_{F_1}[(c+s)\beta(c)] \leq \kappa + \mathbb{E}_{F_1}\left[\left(\frac{1-F_1(c)}{f_1(c)}\right)\beta(c)\right]$$

and

$$\kappa + \int_0^c \beta(y)dy \leq c, \forall c \geq 0$$

We close this section with a remark that although the conditions in corollaries 2 and 3 might not be satisfied in some cases, the "degenerate" game where $g_2^{\mathcal{P}} \equiv g_3^{\mathcal{P}} \equiv \pi_H^{\mathcal{P}} \equiv 0$ can always achieved the equilibrium outcome with both the probability of preventive strike and the probability of a successful development being zero. We also notice that comparing to Corollary 2, the sufficient condition for existence in Corollary 3 is stronger. This is because nonzero probability of preventive strike would have to further satisfy the condition in Proposition 2 that characterizes the incentive constraint of player 1, whereas nonzero probability of a successful development needs not to. With the results in the above two sections, we now turn to discuss their further substantive meanings and their connections to the literature.

5 Discussions

In this section, we further examine some substantive implications of the results above and draw some connections to the literature. As discussed above, in the baseline model, it is possible that a preventive strike is unnecessary. That is, it is possible that proliferator is not aiming for a successful development but the counter-proliferator still initiates a preventive strike. This exhibits a form of inefficiency regarding the nuclear proliferation process, as posited by the pessimists of nuclear proliferation. Indeed, in the baseline model, whenever $t^* > 0$ (which will always hold if $\pi_H, \pi_L \geq 0$, with strict inequality holds for some t and $\pi_H(0) = \pi_L(0) = 0$), this result would be inevitable since the proliferator can always gain from the proliferation process by "blakmailing" the counter-proliferator for economic aids and, on the other hand, the counter-proliferator cannot identify whether the proliferator is effectively aiming for successful development or is simply "blackmailing". This mechanism is similar to what Baliga & Sjostrom (2008) referred as *deterrence of ambiguity* and it further implies that although such ambiguity may be beneficial for the proliferator, it would damage international stability and create some efficiency loss. Furthermore, if Assumption 3 is added, we can see from Figure 2 and Figure 3 that the probability of preventive strike is decreasing as the cost of preventive strike c is increasing, and will reach zero when such cost exceeds certain value ($\bar{c}$). This result is reminiscent of the findings of Feaver & Niou (1996) as well as Benson & Wen (2011). Moreover, as it is reasonable to suppose that the more asymmetric the military capabilities between the proliferator and the counter-proliferator are, the less it costs the counter-proliferator to initiate a preventive strikes, this result also confirms a conventional wisdom in the literature about the effect of asymmetry (Golestein, 2006). On the other hand, however, the effects of changes in the cost of a successful development to the counter-proliferator (w), the cost of being attacked (s), and the gain of a successful development to the proliferator (ν) remain uncertain and depend on further specifications of the functions π_H, π_L and F_1, F_2.

As pointed out in the previous sections, a main concern of the baseline model or all the "specific" models for the nuclear proliferation process is the robustenss of their results—that whether the results will remain the same after changing some assumptions and specifications of a particular model. Indeed, as seen above, some results derived in other models in the literature are similar to the results of the baseline model in this paper whereas some others

do not. The results we obtained in section 4 in a detail-free fashion addresses this issue. Specifically, according to Proposition 2, the (interim) probability of preventive strike as a function of cost is decreasing and its covariance with the virtual value is bounded above and these properties are robust to specific details of the models. As a result, we can robustly verify the common conclusion in the literature that preventives strikes are more likely when military capabilities are more asymmetric. On the other hand, from Proposition 3, we can see that in any models of nuclear proliferation within this class, the change in values of w, ν and s will affect the probability of preventive strike through the bounds of expected total cost of these games. From these inequalities, as w and s increase and as ν decreases, the constraint on the probability of preventive strike *and* the probability of a successful development becomes more stringent. It is, however, unclear that whether these probabilities will decrease certainly, nor is it clear that which probability will respond to such changes. For instance, when the gain of a successful development to the proliferator ν is larger, in expectation, either smaller probability of a successful development or smaller probability of preventive strike or both could be a possible result. Therefore, in a more general sense, the common conclusions in the literature that involves w, s and ν are only partially valid. Overall, using a detail-free approach, among the common conclusions in the literature, the claim that the probability of preventive strike decreases when cost of preventive strike rises is confirmed, while the effects of changes in the cost of a successful development to the counter-proliferator, cost of being attacked, and the gain of a successful development of the propliferator, or other factors are uncertain.

Thus, these results can provide some further insights to the debate between the pessimists and the optimists of nuclear proliferation. The argument of pessimists that nuclear proliferation might cause preventive strikes and hence cause international instability is generally true. However, such instability occurs only when it is not very costly for the counter-proliferator to attack. As the cost becomes higher, the probability of preventive strike must approach to zero rapidly. On the other hand, despite the possibility of "unnecessary attacks", Proposition 3 implies that, when the expected cost of preventive strike decreases, the upper bound of total cost of the entire proliferation process will decrease. Therefore, from another perspective, the nuclear proliferation process is efficient in a sense that the total costs must be smaller than the case when there is no such bargaining processes and the expected cost of this process is less when the military capabilities of the two are more asym-

metric. Consequently, although the concerns of pessimists that nuclear proliferation might cause instability are grounded, they might not be as serious as the pessimists perceived. Conversely, the observation that the total cost of the proliferation process is bounded by the counter-proliferator's expected cost of preventive strike could be regarded as a rebuttal for optimists against pessimists on the process of nuclear proliferation.

The results above also provide some insights to the effect of sanctions and economic aids as measures to discourage proliferation. As seen in Proposition 4, in any equilibrium of any proliferation game that preventive strikes do not occur for sure, the expected transfers must fall into a particular region and in particular, must be negative. That is, whenever economic aids are granted and outweigh the amount of sanctions, preventive strike cannot be ruled out as a last resort. On the other hand, to achieve an outcome where a successful development is impossible, Proposition 5 indicates that the expected transfer must be in a region and in particular, must be positive. Therefore, to completely rule out the possibility of a successful development, economic aids must outweigh sanctions. Thus, for the counter-proliferator, to be always successful in discouraging proliferation, economic aids must be granted and preventive strike must not be rule out as a resort in equilibrium. The only possible scenario that both probabilities are zero is the degenerate case where the expected transfer is zero. These observations therefore sheds lights on the conventional wisdom about economic aids. Specifically, Mazzar's (1995) claim that economic aids are more effective than sanctions is consistent with the implication above, provided that preventive strike is not ruled out as an option, whereas the argument by Montgomery (2005) that economic aids are more effective than preventive strikes is somewhat incomplete in that it did not investigate whether economic economic aids alone can be successful without possibility of preventive strike.

Although the baseline model, as well as the generalized results presented above could provide some insights to the literature and connects to the existing theories for nuclear proliferation process, there are still some limitations that could be addressed in future studies. First, although Proposition 1 shows the existence of equilibrium in the baseline model, there are no further results about the uniqueness. If there are multiple equilibria, equilibrium selection will then be an issue when eliciting implications from the model. Second, despite the generality of the results above, there are still some specializations. Namely, the quasi-linearity of both players' payoffs, the deterministic nature of costs and

gains (i.e., w, s and ν), as well as the zero-sum feature of transfers (i.e. no loss due to sanctions or grants). More general version of the models should have less specifications for these parameters and functional forms. Third, the equilibrium concept we adopted here is Bayes Nash equilibrium and hence is an opened-loop solution concept where the players decide a complete course of action before playing the game. To further capture the dynamic nature of proliferation that one can learn about the other's private information as time passes by given that the development is not yet successful nor has the preventive strike occurred, a closed-loop equilibrium concept with belief-updating mechanisms might provide further insights. Finally, the number of players can also be generalized to more than two, in order to study multilateral bargaining and regulation scenarios (e.g., the six-party talk on the Korean Peninsula).

6 Conclusion

The proliferation of nuclear weapons has been one of the major topics of studies of international and global security. Apart from the policy and specific case studies, theoretical and empirical studies have been focusing on causes and consequences of nuclear proliferation. However, theories from the short-run perspective, especially the process of nuclear proliferation where the counter-proliferators and the proliferators interact strategically, with trade-offs among tolerance, sanctions, and preventive strikes are relatively less abundant. In this paper, we first model the process of nuclear proliferation as a contest game with interdependent values, characterized its equilibrium and examined some of its property and implications. Then, we introduced a detail-free analysis by avoiding further specifications of the model and derived several results. From these results, we find that the probability of preventive strike is decreasing as the counter-proliferator's cost of striking rises regardless of the game form or equilibrium selection. On the other hand, we also showed that the effects of other factors (including the cost of a successful development to the counter-proliferator, the cost of being attacked, and the gain of a successful development to the proliferator) on the probability of preventive strike are generally uncertain. We also characterized several restrictions on the probabilities of preventive strike and successful development in any equilibrium, which provide some insights to the trade-offs among sanctions, preventive strike, and granting economic aids as resorts to discourage proliferations, as well as the debate between the pessimism and the optimism on the issue of proliferation processes.

References

Asal, Victor & Kyle Beardsley. 2007. "Proliferation and International Crisis Behavior." *Journal of Peace Research.* 44(2): 139-155.

Baliga, Sandeep & Tomas Sjostrom. 2008. "Strategic Ambiguity and Arms Proliferation." *Journal of Political Economy* 116(6): 1023-1057.

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

Benson, Brett V. & Quan Wen. 2011. "A Bargaining Model of Nuclear Weapons." In *Causes and Consequences of Nuclear Proliferation,* eds. Robert Rauchhaus, Matthew Kroenig, and Erik Gartzke. New York: Routledge, 111-137.

Berkowitz, Bruce D. 1985. "Proliferation, Deterrence, and the Likelihood of Nuclear War." *Journal of Conflict Resolution* 29(1): 112-136.

Bueno de Mesquita, Bruce & William H. Riker. 1982. "An Assessment of the Merits of Selective Nuclear Proliferation." *Journal of Conflict Resolution* 26(2): 263-302.

Conway, Paul D. 2003. *PSanctions or Engagement? Designing United States Diplomatic Policy Tools to Confront Nuclear Proliferation in Iran, North Korea, India and Pakistan.* Doctoral dissertation, Department of Political Science, Brandeis University, Waltham, MA.

Dunn, Lewis A. 1982. *Controlling the Bomb: Nuclear Proliferation in the 1980s.* New Haven: Yale University Press.

Dunn, Lewis A. 1991. *Containing Nuclear Proliferation.* London: IISS.

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

Feaver, Peter D. 1993. "Proliferation Optimism and Theories of Nuclear Operations." *Security Studies* 2(3/4): 159-191.

Feaver, Peter D. & Emerson M.S. Niou. 1996. "Managing Nuclear Proliferation: Condemn, Strike, or Assist?" *International Studies Quarterly* 40(2): 209-233.

Fey, Mark & Kristopher W. Ramsey. 2009. "Mechanism Design Goes to War: Peaceful Outcomes with Interdependent and Correlated Types." *Review of Economic Design.* 13(3): 233-250.

Fey, Mark & Kristopher W. Ramsey. 2011. "Uncertainty and Incentives in Crisis Bargain-

ing: Game-Free Analysis of International Conflict." *American Journal of Political Science.* 55(1):149-169.

Forland, Astrid. 2007. "Preventive War as an Alternative to Treaty-Based Nuclear Non-Proliferation." In *Nuclear Proliferation and International Security*, eds. Morten Bremer Marli & Sverre Lodgaard. New York: Routledge, 30-49.

Gartzke, Eric & Dong-Joon Jo. 2011. "Bargaining, Nuclear Proliferation, and Interstate Disputes." In *Causes and Consequences of Nuclear Proliferation,* ed. Robert Rauchhaus, Matthew Kroenig and Erik Gartzke. New York: Routledge, 155-182.

Goldblat, Jozef. 2007. "Ban on Nuclear-Weapon Proliferation in Light of International Law." In *Nuclear Proliferation and International Security,* eds. Morten Bremer Marli & Sverre Lodgaard. New York: Routledge, 9-29.

Goldstein, Lyle J. 2006. *Preventive Attack and Weapons of Mass Destruction.* Stanford: Stanford University Press.

James, Carolyn C. 2000. "Nuclear Arsenal Games: Coping with Proliferation in a World of Changing Rivalries." *Canadian Journal of Political Science.* 33(4): 723-746.

Jo, Dong-Joon & Erik Gartzke. 2007. "Determinants of Nuclear Weapons Proliferation." *Journal of Conflict Resolution* 51(1): 167-194.

Kaiser, Karl. 1989. "Non-Proliferation and Nuclear Deterrence." *Survival* 31(2): 123-136.

Karl, David J. 1996. "Proliferation Pessimism and Emerging Nuclear Powers." *International Security* 21(3): 87-119.

Martel, William C. 2001. "Proliferation and Pragmatism: Opportunities and Risks in the Twenty-First Century." In *Deterrence and Nuclear Proliferation in the Twenty-First Century,* ed. Stephen J. Cimbala. Westport: Praeger Publishers, 103-118.

Mazarr, Michael J. 1995. "Going Just a Little Nuclear: NonProliferation Lessons from North Korea." *International Security* 20(2): 92-122.

Mearsheimer, John J. 1993. "The Case for a Ukrainian Nuclear Deterrent." *Foreign Affairs* 73(3): 50-66.

Montgomery, Alexander. 2005. "Ringing in Proliferation: How to Dismantle and Atomic Bomb Network." *International Security* 30(2): 153-187.

Moriarty, Tom. 2004. "Entering the Valley of Uncertainty: The Future of Preemptive Attack." *World Affairs* 167(2): 71-77. Sagan, Scott D. & Kenneth N. Waltz. 1995. *The Spread of Nuclear Weapons: A Debate.* New York: W.W. Norton & Company.

Powell, Robert. 2003. "Nuclear Deterrence Theory, Nuclear Proliferation, and National Missile Defense." *International Security* 27(4): 86-118.

Sagan, Scott D. 2003. "More will be Worse." In *The Spread of Nuclear Weapons: A Debate Renewed,* eds. Scott D. Sagan & Kenneth N. Waltz. New York: W.W. Norton & Company, 46-87.

Schneider, Barry R. 1994. "Nuclear Proliferation and Counter-Proliferation: Policy Issues and Debates." *Mershon International Studies Review* 38(2): 209-234.

Simpson, John. 1994. "Nuclear Non-Proliferation in the Post-Cold War Era." *International Affairs* 17(1): 17-39.

Singh, Sonali & Christopher R. Way. 2004. "The Correlates of Nuclear Proliferation: A Quantitative Test." *Journal of Conflict Resolution* 48(6): 859-885.

Waltz, Kenneth N. 1990. "Nuclear Myths and Political Realities." *American Political Science Review* 84(3): 731-745.

Waltz, Kenneth N. 2003. "More may be Better." In *The Spread of Nuclear Weapons: A Debate Renewed,* eds. Scott D. Sagan & Kenneth N. Waltz. New York: W. W. Norton & Company, 3-45.

Woods, Matthew. 2002. "Reflections on Nuclear Optimism: Waltz, Burke and Proliferation." *Review of International Studies* 28(1): 163-189.

Appendix

A.1 Proof of Proposition 1

Proof of Proposition 1. To begin with, fix $T > 0$, given any $0 \leq t^* \leq \hat{t} < T$, consider first the best response of player 1 when player 2 is following the strategy of form

$$t_2(\tau) = \begin{cases} \tau & \text{if} \quad \tau \leq \hat{t} \\ t^* & \text{if} \quad \tau > \hat{t} \end{cases}. \tag{5}$$

Let $U_1(t_1, c|t^*, \hat{t}) := \mathbb{E}_{F_2}[u_1(t_1, t_2(\tau), c, \tau)]$ denote the interim expected payoff of player 1 with type c when choosing t_1. Then,

$$U_1(t_1, c|t^*, \hat{t}) =$$
$$\begin{cases} (-c - \pi_L(t_1))(1 - F_2(t_1)) + \int_0^{t_1}(-\pi_L(\tau) - w)dF_2(\tau) & \text{if} \quad t_1 < t^* \\ (-\pi_H(t^*))(1 - F_2(\hat{t})) + (-c - \pi_L(t_1))(F_2(\hat{t}) - F_2(t_1)) + \int_0^{t_1}(-\pi_L(\tau) - w)dF_2(\tau) & \text{if} \quad t^* \leq t_1 \leq \hat{t} \\ -\pi_H(t^*)(1 - F_2(\hat{t})) + \int_0^{\hat{t}}(-\pi_L(\tau) - w)dF_2(\tau) & \text{if} \quad t_1 > \hat{t} \end{cases}.$$

Let $\underline{v}$, $\bar{v}$, v^* be the value functions of the three segments and $\underline{\gamma}$, $\bar{\gamma}$, γ^* be the associated optimal choices. That is:

$$\underline{v}(c|t^*, \hat{t}) := \max_{0 \leq t_1 \leq t^*} \left[(-c - \pi_L(t_1))(1 - F_2(t_1)) + \int_0^{t_1}(-\pi_L(\tau) - w)dF_2(\tau)\right]$$

$$\bar{v}(c|t^*, \hat{t}) := \max_{t^* < t_1 \leq \hat{t}} \Bigg[(-\pi_H(t^*))(1 - F_2(\hat{t})) + (-c - \pi_L(t_1))(F_2(\hat{t}) - F_2(t_1)) + \int_0^{t_1}(-\pi_L(\tau) - w)dF_2(\tau)\Bigg]$$

$$v^*(t^*, \hat{t}) := -\pi_H(t^*)(1 - F_2(\hat{t})) + \int_0^{\hat{t}}(-\pi_L(\tau) - w)dF_2(\tau),$$

and

$$\underline{\gamma}(c|t^*, \hat{t}) \in \operatorname*{argmax}_{0 \leq t_1 \leq t^*} \left[(-c - \pi_L(t_1))(1 - F_2(t_1)) + \int_0^{t_1}(-\pi_L(\tau) - w)dF_2(\tau)\right]$$

$$\bar{\gamma}(c|t^*, \hat{t}) \in \operatorname*{argmax}_{t^* < t_1 \leq \hat{t}} \Bigg[(-\pi_H(t^*))(1 - F_2(\hat{t})) + (-c - \pi_L(t_1))(F_2(\hat{t}) - F_2(t_1))) + \int_0^{t_1}(-\pi_L(\tau) - w)dF_2(\tau)\Bigg].$$

Thus, the maximization problem

$$\max_{t_1 \geq 0} U_1(t_1, c|t^*, \hat{t})$$

is then equivalent to

$$\max\left\{\underline{v}(c|t^*, \hat{t}), \bar{v}(c|t^*, \hat{t}), v^*(t^*, \hat{t})\right\}.$$

Under assumptions 1 and 2, the objective functions are differentiable and by compactness of $[0, t^*]$ $[t^*, \hat{t}]$, $\underline{v}, \bar{v}, v^*$ and $\underline{\gamma}, \bar{\gamma}$ are well defined. Moreover, $\underline{v}$ and $\bar{v}$ are differentiable almost everywhere and

$$\underline{v}'(c|t^*, \hat{t}) = -[1 - F_2(\underline{\gamma}(c|t^*, \hat{t}))]$$
$$\bar{v}'(c|t^*, \hat{t}) = -[F_2(\hat{t}) - F_2(\bar{\gamma}(c|t^*, \hat{t}))],$$

by the Envelope Theorem and hence $\underline{v}'(c|t^*, \hat{t}) < \bar{v}'(c|t^*, \hat{t}) < 0, \forall c \geq 0$.

Therefore, there exist a unique $\bar{c}(t^*, \hat{t})$ such that $\bar{v}(\bar{c}(t^*, \hat{t})|t^*, \hat{t}) = v^*(t, \hat{t})$ and a unique $\underline{c}(t^*, \hat{t})$ such that $\underline{v}(\underline{c}(t^*, \hat{t})|t^*, \hat{t}) = \bar{v}(\underline{c}(t^*, \hat{t})|t^*, \hat{t})$, provided that $\underline{v}(0|t^*, \hat{t}) \geq \bar{v}(0|t^*, \hat{t})$. It then follows that :

$$t_1(c|t^*, \hat{t}) = \begin{cases} \underline{\gamma}(c|t^*, \hat{t}) & \text{if} \quad 0 \leq c \leq \underline{c}(t^*, \hat{t}) \\ \bar{\gamma}(c|t^*, \hat{t}) & \text{if} \quad \underline{c}(t^*, \hat{t}) < c \leq \bar{c}(t^*, \hat{t}) \\ \hat{t} & \text{if} \quad c > \bar{c}(t^*, \hat{t}) \end{cases}$$

is player 1's best response against $t_2(\tau|t^*, \hat{t})$.

On the other hand, for player 2, given $t^*, \hat{t}$ and the strategy $t_1(c|t^*, \hat{t})$ obtained above, let $U_2(t_2, \tau|t^*, \hat{t}) := \mathbb{E}[u_2(t_1(c|t^*, \hat{t}), t_2, \tau)]$ denote the interim expected payoff. Then:

$$U_2(t_2, \tau|t^*, \hat{t})$$
$$= \begin{cases} \pi_H(t_2)P(t_2 \leq t_1(c|t^*, \hat{t})) + \int_{\{c|t_2 > t_1(c|t^*, \hat{t})\}} [\pi_L(t_1(c|t^*, \hat{t})) - s] dF_1(c) & \text{if} \quad t_2 \leq \tau \\ (\pi_H(\tau) + \nu)P(\tau \leq t_1(c|t^*, \hat{t})) + \int_{\{c|\tau > t_1(c|t^*, \hat{t})\}} [\pi_L(t_1(c|t^*, \hat{t})) - s] dF_1(c) & \text{if} \quad t_2 > \tau \end{cases}.$$

Also, define

$$\varphi(t|t^*, \hat{t}) := \pi_H(t)P(t \leq t_1(c|t^*, \hat{t})) + \int_{\{c|t > t_1(c|t^*, \hat{t})\}} [\pi_L(t_1(c|t^*, \hat{t})) - s] dF_1(c),$$
$$\psi(t|t^*, \hat{t}) := (\pi_H(t) + \nu)P(t \leq t_1(c|t^*, \hat{t})) + \int_{\{c|t > t_1(c|t^*, \hat{t})\}} [\pi_L(t_1(c|t^*, \hat{t})) - s] dF_1(c),$$

Then

$$\max_{t_2 \geq 0} U_2(t_2, \tau|t^*, \hat{t})$$

is equivalent to

$$\max\{\max_{t\leq\tau}\varphi(t|t^*,\hat{t}),\psi(\tau|t^*,\hat{t})\}.$$

Notice that $\arg\max_{t\in[0,T]}\varphi(t|t^*,\hat{t})$ and is nonempty by continuity of φ and compactness of $[0,T]$. Let $x^*(t^*,\hat{t})$ be selected from $\operatorname{argmax}_{t\geq 0}\varphi(t|t^*,\hat{t})$. It then follows that $x^*(t^*,\hat{t})$ is continuous in $(t^*,\hat{t})$ by the theorem of maximum and $x^*\leq T$.

In addition, define $\hat{x}(t^*,\hat{t})$ as follows:

$$\begin{cases} \varphi(x^*(t^*,\hat{t}|t^*,\hat{t}))=\psi(\hat{x}(t^*,\hat{t}|t^*,\hat{t})) \quad \hat{x}(t^*,\hat{t})\geq x^*(t^*,\hat{t}) \quad \text{if} \quad \psi(\hat{t}|t^*,\hat{t})\leq\varphi(x^*(t^*,\hat{t}|t^*,\hat{t})) \\ \hat{x}(t^*,\hat{t}):=\hat{t} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

By existence of $x^*(t^*,\hat{t})$ and continuity of φ and ψ, $\hat{x}$ is well-defined, continuous in $(t^*,\hat{t})$ and $x^*\leq\hat{x}\leq\hat{t}$. Notice that by construction, $t_2(\tau|t^*,\hat{t})\in\operatorname{argmax}_{t_2\geq 0}U_2(t_2,\tau|t^*,\hat{t})$ if $t_2(\tau|t^*,\hat{t})$ is of form:

$$t_2(\tau|t^*,\hat{t})=\begin{cases} x^*(t^*,\hat{t}) & \text{if} \quad \tau>\hat{x}(t^*,\hat{t}) \\ \tau & \text{if} \quad \tau\leq\hat{x}(t^*,\hat{t}) \end{cases}.$$

Finally, since the set $\{(t^*,\hat{t})|0\leq t^*\leq\hat{t}\leq T\}$ is non-empty, compact and convex. Also, as shown above, $(x^*,\hat{x})$ is a continuous self-map on $\{(t^*,\hat{t})|0\leq t^*\leq\hat{t}\leq T\}$. Thus, by Brouwer's Fixed Point Theorem, there exist $(t^*,\hat{t})$ such that $x^*(t^*,\hat{t})=t^*,\hat{x}(t^*,\hat{t})=\hat{t}$. Such fixed point $(t^*,\hat{t})$ and the associated $\underline{\gamma}(c|t^*,\hat{t}),\bar{\gamma}(c|t^*,\hat{t})$ and $\underline{c}(t^*,\hat{t}),\bar{c}(t^*,\hat{t})$ then defines a Bayes Nash equilibrium (t_1^*,t_2^*) ■

A.2 Proof of Lemma 1

Proof of Lemma 1. For sufficiency, let $\mathcal{D}=(\pi_H,\pi_L,\boldsymbol{g})$ be an incentive compatible direct mechanism. Then for any $c'>c\geq 0$, from incentive compatibility,

$$\begin{aligned} U_1^*(c) &\geq U_1(c,c') \\ &= -\mathbb{E}_{F_2}\left[\pi_L(c',\tau,\tau)g_3(c',\tau,\tau)+\pi_H(c',\tau,\tau)(1-g_3(c',\tau,\tau))+wg_2(c',\tau,\tau)+cg_3(c',\tau,\tau)\right] \\ &= U_1^*(c')-\mathbb{E}_{F_2}[g_3(c',\tau,\tau)](c-c'). \end{aligned}$$

Since $c'>c$, it follows that

$$\frac{U_1^*(c')-U_1^*(c)}{c'-c}\leq-\mathbb{E}_{F_2}[g_3(c',\tau,\tau)].$$

On the other hand, incentive compatibility also implies:

$$\begin{aligned}
U_1^*(c') &\geq U_1(c', c) \\
&= -\mathbb{E}_{F_2}\left[\pi_L(c,\tau,\tau)g_3(c,\tau,\tau) + \pi_H(c,\tau,\tau)(1 - g_3(c,\tau,\tau)) + wg_2(c,\tau,\tau) + c'g_3(c,\tau,\tau)\right] \\
&= U_1^*(c) - \mathbb{E}_{F_2}[g_3(c,\tau,\tau)](c' - c),
\end{aligned}$$

and hence

$$\frac{U_1^*(c') - U_1^*(c)}{c' - c} \geq -\mathbb{E}_{F_2}[g_3(c,\tau,\tau)].$$

Together, we have:

$$-\mathbb{E}_{F_2}[g_3(c,\tau,\tau)] \leq \frac{U_1^*(c') - U_1^*(c)}{c' - c} \leq -\mathbb{E}_{F_2}[g_3(c',\tau,\tau)] \tag{6}$$

for any $c' > c \geq 0$ and therefore $\mathbb{E}_{F_2}[g_3(c,\tau,\tau)]$ is nonincreasing in c. Furthermore, since $g_3 \in [0,1]$, equation (6) then implies that U_1^* is Lipschitz continuous, in particular, is absolutely continuous and hence is differentiable almost everywhere, with $U_1^{*'}(c) = -\mathbb{E}_{F_2}[g_3(c,\tau,\tau)]$ for any c at which U_1^* is differentiable by the envelope theorem. The fundamental theorem of calculus for Lebesgue integral then gives:

$$U_1^*(c) = U_1^*(0) - \int_0^c \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy, \forall c \geq 0$$

Conversely, let $\mathcal{D} = (\pi_H, \pi_L, \boldsymbol{g})$ be a direct mechanism satisfying 1 and 2 in the assertion. Then for any $c \geq 0$ and any $c' > c$, as above,

$$\begin{aligned}
U_1(c, c') &= U_1^*(c') + \mathbb{E}_{F_2}[g_3(c',\tau,\tau)](c' - c) \\
&= U_1^*(0) - \int_0^{c'} \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy + \int_c^{c'} \mathbb{E}_{F_2}[g_3(c',\tau,\tau)]dy \\
&= U_1^*(c) + \int_c^{c'} \left(\mathbb{E}_{F_2}[g_3(c',\tau,\tau)] - \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]\right) dy \\
&\leq U_1^*(c),
\end{aligned}$$

where the last inequality follows from 2. Since c, c' are arbitrary, the above inequality also implies that $U_1(c, c') \leq U_1^*(c)$ for any $c' < c$ whenever $c > 0$. This completes the proof. ■

A.3 Proof of Proposition 2

Proof of Proposition 2. Let $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$ and σ^* be any pair of individually rational nuclear proliferation game and its Bayes Nash equilibrium. By the revelation principle, there

exists an incentive compatible and individually rational direct mechanism $\mathcal{D} = (\pi_H, \pi_L, \boldsymbol{g})$ such that $\pi_H(c,\tau,\tau) = \pi_H^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$, $\pi_L(c,\tau,\tau) = \pi_L^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$ and $\boldsymbol{g}(c,\tau,\tau) = \boldsymbol{g}^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$, for any $c, \tau \geq 0$ and hence $U_1^*(c) = U_1^{\mathcal{P}}(c|\sigma^*)$ for any $c \geq 0$. From lemma 1, since $\mathcal{D}$ is incentive compatible,

$$U_1^*(c) = U_1^*(0) - \int_0^c \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy, \forall c \geq 0.$$

Assertion 1 then follows. Also, from lemma 1 and incentive compatibility of $\mathcal{D}$, $\mathbb{E}_{F_2}[g_3(c,\tau,\tau)]$ is nonincreasing in c and hence assertion 2 follows. For assertion 3, since σ^* is a Bayes Nash equilibrium, $U_1^{\mathcal{P}}(c|\sigma^*) \geq \mathbb{E}_{F_2}[u_1^{\mathcal{P}}(a_1, \sigma_2^*(\tau), c, \tau)]$ for any $a_1 \in A_1$, in particular, by Assumption 5, let $\tilde{a_1}$ be such that $g_3^{\mathcal{P}}(\tilde{a_1}, a_2, \tau) = 0$ for any $a_2 \in A_2, \tau \geq 0$,

$$U_1^{\mathcal{P}}(c|\sigma^*) \geq -\mathbb{E}_{F_2}[\pi_H(\tilde{a_1}, \sigma_2^*(\tau), \tau) + wg_2(\tilde{a_1}, \sigma_2^*(\tau), \tau)], \forall c \geq 0.$$

Under Assumption 4, since σ_2^* is $\mathcal{A}_2$ measurable, for $\bar{w} := \mathbb{E}_{F_2}[\pi_H^{\mathcal{P}}(\tilde{a_1}, \sigma_2^*(\tau), \tau) + wg_2^{\mathcal{P}}(\tilde{a_1}, \sigma_2^*(\tau), \tau)]$, $|\bar{w}| < \infty$. By the revelation principle and assertion 1, we have:

$$-U_1^*(0) + \int_0^c \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy \leq \bar{w}, \forall c \geq 0.$$

Hence,

$$\int_0^\infty \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy \leq U_1^*(0) + \bar{w} < \infty.$$

Since $g_3 \geq 0$, it then follows that

$$\lim_{z\to\infty} \int_z^\infty \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy = 0.$$

Assertion 3 then follows by the revelation principle. ■

A.4 Proof of Proposition 3

Proof of Proposition 3. Let $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$ and σ^* be any pair of individually rational nuclear proliferation game and its Bayes Nash equilibrium with $g_2^{\mathcal{P}} \equiv 0$. By the revelation principle, there exists an incentive compatible and individually rational direct mechanism $\mathcal{D} = (\pi_H, \pi_L, \boldsymbol{g})$ such that $\pi_H(c,\tau,\tau) = \pi_H^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$, $\pi_L(c,\tau,\tau) = \pi_L^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$ and $\boldsymbol{g}(c,\tau,\tau) = \boldsymbol{g}^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$, for any $c, \tau \geq 0$ and hence $U_1^*(c) = U_1^{\mathcal{P}}(c|\sigma^*)$ for any $c \geq 0$, $U_2^*(\tau) = U_2^{\mathcal{P}}(\tau|\sigma^*)$ for any $\tau \geq 0$. For conciseness of notations, write $\Pi(c,\tau) := \pi_L(c,\tau,\tau)g_3(c,\tau,\tau) + \pi_H(c,\tau,\tau)(1 - g_3(c,\tau,\tau))$. We may then rewirte the interim expected

payoff of both players as:

$$U_1^*(c) = -\mathbb{E}_{F_2}[\Pi(c,\tau) + cg_3(c,\tau,\tau) + wg_2(c,\tau,\tau)]$$
$$U_2^*(\tau) = \mathbb{E}_{F_1}[\Pi(c,\tau) - sg_3(c,\tau,\tau) + \nu g_2(c,\tau,\tau)]$$

Also, incentive compatibility and lemma 1 implies:

$$U_1^*(c) = U_1^*(0) - \int_0^c \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy, \forall c \geq 0.$$

Together, we have

$$\mathbb{E}_{F_2}[\Pi(c,\tau)] = -U_1^*(0) + \int_0^c \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy - c\mathbb{E}_{F_2}[g_3(c,\tau,\tau)] - w\mathbb{E}_{F_2}[g_2(c,\tau,\tau)], \quad (7)$$

for any $c \geq 0$. Additionally,

$$-U_1^*(0) + \int_0^c \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy \leq c, \quad (8)$$

for any $c \geq 0$ by individual rationality. On the other hand, for player 2, individual rationality implies:

$$\mathbb{E}_{F_1}[\Pi(c,\tau) - sg_3(c,\tau,\tau) + \nu g_2(c,\tau,\tau)] \geq 0$$

for any $\tau \geq 0$. By taking expectation with respect to τ,

$$\mathbb{E}[\Pi(c,\tau) - sg_3(c,\tau,\tau) + \nu g_2(c,\tau,\tau)] \geq 0. \quad (9)$$

Notice that

$$\begin{aligned}
\mathbb{E}_{F_1}\left[\int_0^c \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy\right] &= \int_0^\infty \int_0^c \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy f_1(c)dc \\
&= \int_0^\infty \int_y^\infty \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]f_1(c)dcdy \\
&= \int_0^\infty \mathbb{E}_{F_2}[g_3(c,\tau,\tau)]\left(\frac{1-F_1(y)}{f_1(y)}\right)f_1(y)dy \\
&= \mathbb{E}_{F_1}\left[\mathbb{E}_{F_2}[g_3(c,\tau,\tau)]\left(\frac{1-F_1(c)}{f_1(c)}\right)\right] \\
&= \mathbb{E}\left[\left(\frac{1-F_1(c)}{f_1(c)}\right)g_3(c,\tau,\tau)\right],
\end{aligned}$$

by applying the Fubini's Theorem and the fact that c, τ are independent. Therefore,

$$\begin{aligned}
&\mathbb{E}[\Pi(c,\tau) - sg_3(c,\tau,\tau) + \nu g_2(c,\tau,\tau)] \\
&= \mathbb{E}_{F_1}[\mathbb{E}_{F_2}[\Pi(c,\tau)] - s\mathbb{E}_{F_2}[g_3(c,\tau,\tau)] + \nu\mathbb{E}_{F_2}[g_2(c,\tau,\tau)]] \\
&= \mathbb{E}_{F_1}[-U_1^*(0) + \int_0^c \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy - (c+s)\mathbb{E}_{F_2}[g_3(c,\tau,\tau)] + (\nu - w)g_2(c,\tau,\tau)] \\
&= -U_1^*(0) + \mathbb{E}\left[\left(\frac{1-F_1(c)}{f_1(c)} - (c+s)\right)g_3(c,\tau,\tau)\right] + \mathbb{E}[(\nu-w)g_2(c,\tau,\tau)],
\end{aligned}$$

by using the expression in (7). Together with (9),

$$-U_1^*(0) + \mathbb{E}\left[\left(\frac{1-F_1(c)}{f_1(c)}\right) g_3(c,\tau,\tau)\right] \geq \mathbb{E}[(c+s)g_3(c,\tau,\tau) + (w-\nu)g_2(c,\tau,\tau))]. \quad (10)$$

Combining (8) and (10) and take expectation with respect to c on both sides of (8),

$$\mathbb{E}[(c+s)g_3(c,\tau,\tau) + (w-\nu)g_2(c,\tau,\tau)] \leq -U_1^*(0)+ \leq \mathbb{E}\left[\left(\frac{1-F_1(c)}{f_1(c)}\right) g_3(c,\tau,\tau)\right] \leq \mathbb{E}[c].$$

Moreover, under Assumption 1, since the random variable c has finite second moment, $\mathrm{Cov}(c, g_3(c,\tau,\tau))$ is well defined by the Cauchy Schwartz inequality and the fact that $0 \leq g_3 \leq 1$. On the other hand, using integration by part, we observe that

$$\begin{aligned}\mathbb{E}\left[\frac{1-F_1(c)}{f_1(c)}\right] &= \int_0^\infty (1-F_1(c))dc \\ &= c(1-F_1(c))|_0^\infty + \int_0^\infty c f_1(c)dc \\ &= 0 + \mathbb{E}[c],\end{aligned}$$

where the last equality follows from the observation that for any $x \geq 0$,

$$x(1-F_1(x)) \leq \int_x^\infty c dF_1(c)$$

and hence

$$\lim_{x\to\infty} x(1-F_1(x)) \leq \lim_{x\to\infty} \int_x^\infty c dF_1(c) = 0$$

by finiteness of $\mathbb{E}[c]$. Thus, since $g_3 \leq 1$,

$$\mathrm{Cov}\left(\frac{1-F_1(c)}{f_1(c)}, g_3(c,\tau,\tau)\right) \leq \mathbb{E}\left[\left(\frac{1-F_1(c)}{f_1(c)}\right) g_3(c,\tau,\tau)\right] < \infty$$

Therefore, the inequality

$$\mathbb{E}[(c+s)g_3(c,\tau,\tau) + (w-\nu)g_2(c,\tau,\tau)] \leq -U_1^*(0)+ \leq \mathbb{E}\left[\left(\frac{1-F_1(c)}{f_1(c)}\right) g_3(c,\tau,\tau)\right]$$

can be rewritten as

$$\begin{aligned}&\mathrm{Cov}(c, g_3(c,\tau,\tau)) + \mathbb{E}[c+s]\mathbb{E}[g_3(c,\tau,\tau)] + \mathbb{E}[(w-\nu)g_2(c,\tau,\tau)] \\ &\qquad \leq -U_1^*(0) + \mathrm{Cov}\left(\frac{1-F_1(c)}{f_1(c)}, g_3(c,\tau,\tau)\right) + \mathbb{E}[c]\mathbb{E}[g_3(c,\tau,\tau)]\end{aligned}$$

and hence

$$\mathrm{Cov}\left(c - \frac{1-F_1(c)}{f_1(c)}, g_3(c,\tau,\tau)\right) \leq -U_1^*(0) + \mathbb{E}[(w-\nu)g_2(c,\tau,\tau) - s g_3(c,\tau,\tau)].$$

■

A.5 Proof of Proposition 4

Proof of Proposition 4. Let $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$ and σ^* be any pair of individually rational nuclear proliferation game and its Bayes Nash equilibrium with $g_3^{\mathcal{P}} \equiv 0$. By the revelation principle, there exists an incentive compatible and individually rational direct mechanism $\mathcal{D} = (\pi_H, \pi_L, \boldsymbol{g})$ such that $\pi_H(c,\tau,\tau) = \pi_H^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$, $\pi_L(c,\tau,\tau) = \pi_L^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$ and $\boldsymbol{g}(c,\tau,\tau) = \boldsymbol{g}^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$, for any $c, \tau \geq 0$ and hence $U_1^*(c) = U_1^{\mathcal{P}}(c|\sigma^*)$ for any $c \geq 0$, $U_2^*(\tau) = U_2^{\mathcal{P}}(\tau|\sigma^*)$ for any $\tau \geq 0$, and $g_3 \equiv 0$. By lemma 1, for any $c \geq 0$

$$U_1^*(c) = U_1^*(0) + \int_0^c \mathbb{E}_{F_2}[g_3(y,\tau,\tau)]dy = U_1^*(0).$$

Individual rationality of $\mathcal{D}$ then implies that $U_1^*(c) = U_1^*(0) \geq 0$ for any $c \geq 0$. Also, by definition and that $g_3 \equiv 0$,

$$\begin{aligned} U_1^*(c) &= \mathbb{E}_{F_2}[u_1(c,c,\tau,\tau)] \\ &= -\mathbb{E}_{F_2}\left[(\pi_L(c,\tau,\tau) + c)g_3(c,\tau,\tau) + \pi_H(c,\tau,\tau)(1 - g_3(c,\tau,\tau)) + wg_2(c,\tau,\tau)\right] \\ &= -\mathbb{E}_{F_2}[\pi_H(c,\tau,\tau) + wg_2(c,\tau,\tau)]. \end{aligned}$$

Together, we have:

$$\mathbb{E}_{F_2}[\pi_H(c,\tau,\tau) + wg_2(c,\tau,\tau)] \leq 0, \forall c \geq 0.$$

Taking expectation on both sides under F_1 and using the fact that c and τ are independent, we then have:

$$\mathbb{E}[\pi_H(c,\tau,\tau)] \leq -w\mathbb{E}[g_2(c,\tau,\tau)] \tag{11}$$

On the other hand, by definition and $g_3 \equiv 0$,

$$\begin{aligned} U_2^*(\tau) &= \mathbb{E}_{F_1}[u_2(c,\tau,\tau)] \\ &= \mathbb{E}_{F_1}\left[(\pi_L(c,\tau,\tau) - s)g_3(c,\tau,\tau) + \pi_H(c,\tau,\tau)(1 - g_3(c,\tau,\tau)) + \nu g_2(c,\tau,\tau)\right] \\ &= \mathbb{E}_{F_1}[\pi_H(c,\tau,\tau) + \nu g_2(c,\tau,\tau)]. \end{aligned}$$

Individual rationality then implies that:

$$\mathbb{E}_{F_1}[\pi_H(c,\tau,\tau) + \nu g_2(c,\tau,\tau)] \geq 0, \forall \tau \geq 0.$$

Taking expectation with respect to τ and using the fact that c and τ are independent,

$$\mathbb{E}[\pi_H(c,\tau,\tau)] \geq -\nu\mathbb{E}[g_2(c,\tau,\tau)]. \tag{12}$$

Combining equation (7) and (8), it follows that:

$$w \leq -\frac{\mathbb{E}[\pi_H(c,\tau,\tau)]}{\mathbb{E}[g_2(c,\tau,\tau)]} \leq \nu$$

if $\mathbb{E}[g_2(c,\tau,\tau)] = 0$ and that $\mathbb{E}[\pi_H(c,\tau,\tau)] = 0$ if $\mathbb{E}[g_2(c,\tau,\tau)] = 0$. By the revelation principle, the assertion then follows. ■

A.6 Proof of Corollary 1

Proof of Corollary 1. Necessity follows directly from the first case of Proposition 3. For sufficiency, first notice that since a direct mechanism is incentive compatible if and only if identity function is a Bayes Nash equilibrium, it is sufficient to find an incentive compatible and individually rational direct mechanism with $g_3 \equiv 0$. Furthermore, by the converse assertion of lemma 1, if there exists two functions α, β such that $\mathbb{E}[\beta(c,\tau)] > 0$, $\mathbb{E}_{F_2}[\alpha(c,\tau) + w\beta(c,\tau)]$ is independent of c, $\mathbb{E}_{F_2}[\alpha(c,\tau)+w\beta(c,\tau)] \leq 0$ and $\mathbb{E}_{F_1}[\alpha(c,\tau)+\nu\beta(c,\tau)] \geq 0, \forall \tau \geq 0$, then by defining $\pi_H(c,\tau,\tau') := \alpha(c,\tau)$ and $g_2(c,\tau,\tau') := \beta(c,\tau)$ for any $c,\tau,\tau' \geq 0$, any direct mechanism defined by $(\pi_H, \pi_L, \boldsymbol{g})$, with $g_3 \equiv 0$, $g_1 \equiv 1 - g_2$ and any π_L is then incentive compatible and individually rational. We claim that such functions α, β exists by the following example. Let $\beta : \mathbb{R}^2_+ \to [0,1]$ be any function such that $\mathbb{E}[\beta(c,\tau)] > 0$ (for instance, $\beta(c,\tau) := \max\{c\tau, 1\}, \forall c,\tau \geq 0$) and let $\alpha := -w\beta$. Then since $0 < w \leq \nu$, $\mathbb{E}_{F_2}[\alpha(c,\tau) + w\beta(c,\tau)] = 0, \forall c \geq 0$ and $\mathbb{E}_{F_1}[\alpha(c,\tau) + \nu\beta(c,\tau)] \geq 0, \forall \tau \geq 0$ and thus the above conditions are satisfied. This completes the proof. ■

A.7 Proof of Proposition 5

Proof of Proposition 5. Let $\mathcal{P} = (A, \pi_H^{\mathcal{P}}, \pi_L^{\mathcal{P}}, \boldsymbol{g}^{\mathcal{P}})$ and σ^* be any pair of individually rational nuclear proliferation game and its Bayes Nash equilibrium with $g_2^{\mathcal{P}} \equiv 0$. By the revelation principle, there exists an incentive compatible and individually rational direct mechanism $\mathcal{D} = (\pi_H, \pi_L, \boldsymbol{g})$ such that $\pi_H(c,\tau,\tau) = \pi_H^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$, $\pi_L(c,\tau,\tau) = \pi_L^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$ and $\boldsymbol{g}(c,\tau,\tau) = \boldsymbol{g}^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)$, for any $c,\tau \geq 0$ and hence $U_1^*(c) = U_1^{\mathcal{P}}(c|\sigma^*)$ for any $c \geq 0$, $U_2^*(\tau) = U_2^{\mathcal{P}}(\tau|\sigma^*)$ for any $\tau \geq 0$, and $g_2 \equiv 0$. As in the proof of proposition 3, write $\Pi(c,\tau) := \pi_L(c,\tau,\tau)g_3(c,\tau,\tau) + \pi_H(c,\tau,\tau)(1 - g_3(c,\tau,\tau))$. Individual rationality and $g_2 \equiv 0$ then requires:

$$U_1^*(c) = -\mathbb{E}_{F_2}[\Pi(c,\tau) + cg_3(c,\tau,\tau)] \geq -c$$

$$U_2^*(\tau) = \mathbb{E}_{F_1}[\Pi(c,\tau) - sg_3(c,\tau,\tau)] \geq 0.$$

Therefore, by taking expectation with respect to c and τ respectively and rearranging the above inequalities,

$$\mathbb{E}[sg_3(c,\tau,\tau)] \leq \mathbb{E}[\Pi(c,\tau)] \leq \mathbb{E}[c(1-g_3(c,\tau,\tau))].$$

■

A.8 Proof of Corollary 3

Proof of Corollary 3. Necessity follows directly from Corollary 2 by defining

$$\kappa := U_1^{\mathcal{P}}(0|\sigma^*), \quad \beta(c) := \mathbb{E}_{F_2}[g_3^{\mathcal{P}}(\sigma_1^*(c), \sigma_2^*(\tau), \tau)]$$

for any individually rational Bayes Nash equilibrium and nuclear proliferation game $(\sigma^*, \mathcal{P})$ with $g_2^{\mathcal{P}} \circ \sigma^* \equiv 0$.

For sufficiency, as noted in the proof of Corollary 1, a direct mechanism is incentive compatible if and only if identity function is a Bayes Nash equilibrium. Thus, it is sufficient to find an incentive compatible and individually rational direct mechanism with $g_2 \equiv 0$ and $g_3 > 0$ with positive measure. Let $\kappa \leq 0$ and β be such function as in the assertion. Define $\bar{g_3}(c,\tau) := \beta(c)\frac{\tau}{\mathbb{E}_{F_2}[\tau]}$, $\bar{g_2} := 0$, $\bar{g_1} := 1 - \bar{g_3}$ and pick any functions $\bar{\pi_H}, \bar{\pi_L}$ such that $\mathbb{E}_{F_2}[\bar{\pi_L}(c,\tau)\bar{g_3}(c,\tau) + \bar{\pi_H}(c,\tau)(1-\bar{g_3}(c,\tau))] = \kappa + \int_0^c \beta(y)dy - c\beta(c)$. Finally, for any $c, \tau, \tau' \geq 0$ let $\pi_H(c,\tau,\tau') := \bar{\pi_H}(c,\tau)$; $\pi_L(c,\tau,\tau') := \bar{\pi_L}(c,\tau)$ and $\boldsymbol{g}(c,\tau,\tau') := \bar{\boldsymbol{g}}(c,\tau)$. The direct mechanism defined by $(\pi_H, \pi_L, \boldsymbol{g})$ is then incentive compatible and individually rational by lemma 1 and by construction. ■