Rewriting the Rules of the NPT

A Game Theoretic Approach to Nuclear Proliferation

by
Dianne R. Pfundstein

Ashok Rai, Advisor

A thesis submitted in partial fulfillment
of the requirements for the
Degree of Bachelor of Arts with Honors
in Economics

WILLIAMS COLLEGE

Williamstown, Massachusetts

9 May 2006

Abstract

This paper employs the tools of game theory to analyze the efficacy of current frameworks designed to prevent nuclear weapons proliferation. Signaling models are constructed to analyze the inability of the current nonproliferation regime to detect and halt nuclear proliferation. The models demonstrate that only specifically targeted sanctions may be successful in preventing proliferation under the structure of the Treaty on the Nonproliferation of Nuclear Weapons (NPT). Such targeted sanctions, which may limit a suspect country's importation of sensitive materials and technologies and are often difficult to enforce, must severely increase costs for a state that develops a nuclear weapons program while remaining within the structure of the NPT. The models also reveal that the Additional Protocol (AP), which was designed to remedy some of the faults of the NPT, is similarly weak in detecting violations and enforcing nonproliferation. Only by offering positive incentives, such as trade agreements, to those countries that sign the AP, combined with an ability to detect nuclear proliferation, can the AP framework separate potential proliferators from their peaceful counterparts. The current NPT-AP structure fails to prevent proliferation because states that do not sign the AP do not have to withdraw from the NPT and lose the benefits of nuclear technology transfer that are provided under the structure of the treaty and that enable states to develop nuclear weapons. Finally, a normative model of the nonproliferation regime is presented and suggests that penalizing states that fail to sign the AP with ejection from the NPT may be effective in separating proliferators and thereby forestalling nuclear proliferation.

Acknowledgements

First and foremost I would like to thank my advisor, Ashok Rai, for his invaluable help and support during the writing of this thesis. He trusted my ideas in the spring of 2005 when I was a student in his game theory class and had faith in my project even when I had only a limited idea of what my thesis would become. I sincerely appreciate his challenging questions, which forced me to hone my intuition and refine my models. I would also like to thank John Bakija for his insightful comments and questions throughout the course of the year, and Steve Sheppard for his comments on an early draft. I am also grateful to the other economics thesis students and to the members of Ashok Rai's game theory class in Spring 2005 for their queries and suggestions throughout the stages of my investigation. Finally I thank my family, friends, and Brad for their support throughout this intensely challenging and rewarding process.

Introduction

Since the end of the Cold War and demise of the former Soviet Union, the greatest threats the world faces from nuclear weapons are not those arising from conflicts between evenly matched powers, but rather from weak states acquiring small arsenals of nuclear weapons. The dissemination of nuclear technologies and materials, often facilitated by the nonproliferation regime, makes the acquisition of weapons by even more countries a very real possibility. During the Cold War, both the United States and its enemy possessed arsenals large enough to literally wipe their opponents off the map, and proliferation theory focused on this ability of the two states to destroy each other. In the current era of proliferation, states that seek to acquire nuclear weapons cannot hope to attain such a destructive capability in the near future. Traditional deterrence theory would suggest that the reasons for holding nuclear weapons therefore become less clear, given that a new proliferator could not hope to inflict a level of damage on a nuclear-armed, superpower opponent equal to what its opponent could inflict on the small state. New theories are thus needed to understand nuclear proliferation in the twenty-first century.

Before the events of September 11th temporarily shifted focus away from nuclear weapons proliferation, “efforts to prevent the proliferation of nuclear weapons [had topped] America’s national security agenda” (Doyle and Engstrom 1998, 39). Recently, North Korea has received a great deal of media attention in the United States and aroused the interest of the international community. The state expelled inspectors from the International Atomic Energy Agency (IAEA) in December 2002 and announced that it would resume its reprocessing of spent nuclear fuel. It then withdrew from the Treaty on the Nonproliferation of Nuclear Weapons (NPT) in January of 2003 (International Institute for Strategic Studies [IISS] 2004, 26). Recent information indicates that North Korea may possess, at most, six to twelve nuclear weapons

(IISS 2004, 48). In six-party talks in fall 2005, North Korea did express a willingness to abandon its nuclear program altogether, although negotiations have stalled somewhat since that announcement (CNN.com 19 Sep 2005). Iran is also making headlines. In February 2006 the board of the IAEA voted to report Iran to the UN Security Council for failures to demonstrate that the country is not pursuing an illicit weapons program (Sciolino 2006). The stalling of negotiations with North Korea and the continued defiance of Iran suggest that there is room for further insight into the ability of the international community to curtail nuclear weapons proliferation.

My research draws largely on the political science literature that addresses nuclear proliferation and that presents various theories about the reasons for which states choose to develop nuclear weapons, and how they may be deterred from using their weapons when they acquire them. Traditional theories of nuclear proliferation were largely developed during the Cold War, when two major powers, the United States and Soviet Union, were engaged in a competition for nuclear supremacy. In the post-Cold War world, when the standoff between two world powers has been replaced by a single actor dominating the international system, the ways of thinking about proliferation by small rogue states must adapt to this new framework.

Prior to the breakup of the Soviet Union, the world was characterized by an arms race between two powerful nations who each sought supremacy in nuclear weapons capability. Today's new proliferators are attempting to acquire nuclear weapons in a world in which the United States (and, to an extent, the former Soviet Union) already possess overwhelming nuclear arsenals with which the new proliferators' nuclear programs could not hope to compete in the near future (IISS 2004). Attempts to model nuclear proliferation in the twenty-first century must stem from an understanding of the motivations behind the acquisition of nuclear weapons in a world in which the new proliferator cannot hope to compete with the nuclear arsenals of the powerful states of the international community.[1]

Should the United States care about nuclear proliferation by such small states, and if so, how can it hope to halt or prevent such proliferation? Given the fact that the United States is in no way prepared to eliminate its own nuclear capabilities, it would likely be difficult for the country to simply convince other states to give up their own weapons. In fact, it could be argued that the very fact that some actors in the international system possess nuclear weapons actually creates an incentive for a smaller state to acquire them—particularly if the smaller state does not support the status quo prevailing in the anarchic international system. On the other hand, many states choose to develop nuclear weapons for reasons unrelated to the structure of the international system, but rather because of domestic politics or the presence of a regional enemy. To prevent states from acquiring nuclear weapons, the international community must develop a framework for understanding how its actions affect a state's ability and intention to acquire nuclear weapons. A successful strategy for combating proliferation by small states must stem from an understanding of the incentives potential proliferators face to develop nuclear weapons. Given that a proliferator is unlikely to announce its intentions to a hostile world community, a state or a coalition of states hoping to prevent or halt proliferation must formulate realistic beliefs and expectations about the proliferator's intentions and motivations, and then use such information to construct strategies to combat nuclear proliferation.

The enactment of the Treaty on the Nonproliferation of Nuclear Weapons (NPT) was the greatest success for nuclear diplomacy during the Cold War and defines the current framework of nuclear arms control. Under the treaty, completed in 1968 and signed by more than 180 nations, the United States, Russia, China, the United Kingdom and France are permitted to

[1] An in-depth analysis of the factors motivating a state to seek nuclear weapons can be found in Appendix A.

possess nuclear weapons; all other signatories agree to renounce the option of developing them (Schell 2000). Non-signatories include Israel, India, Pakistan (all of which have developed nuclear weapons), and as of its withdrawal in 2003, North Korea. North Korea's nuclear program was therefore in violation of the NPT when it restarted its reprocessing activities, and remained in violation until the country's withdrawal in 2003. The structure of the NPT grants non-nuclear weapons states the right to receive peaceful nuclear energy technologies from those states that have nuclear weapons capabilities, in exchange for renouncing the ability to develop a nuclear weapons program.[2] Yet this transfer of nuclear energy technologies actually facilitates the development of a nuclear weapons program, because the materials and technological knowledge necessary for running a nuclear energy reactor may be employed in the development of a nuclear weapons program. It is precisely this element of the current nonproliferation regime that countries such as Iraq and Iran have exploited in their own nuclear weapons programs and that I will model in this paper.

Although many states have signed the NPT, not all have adhered to its conditions. North Korea, Iraq, Iran and other states have all pursued nuclear weapons programs while signatories to the NPT. Iran, which is believed to be pursuing nuclear weapons, has also signed the Additional Protocol, which was designed to remedy some of the limitations of the NPT. The questions I seek to answer are the following: why is the current nonproliferation regime unsuccessful in preventing states that sign nonproliferation treaties from developing nuclear weapons? How may the current treaty structure be modified to better prevent, detect, and curtail nuclear proliferation? To answer these questions, I employ game theory to model the current nonproliferation regime and to explain why efforts to prevent nuclear weapons proliferation have been unsuccessful. I

[2] The specific structure of both the NPT and the Additional Protocol is examined in greater detail in the following section of this paper.

also employ these models to determine conditions under which the international community may halt or prevent nuclear weapons proliferation.

I will now describe the basic structure of the signaling games that I develop to illustrate the conditions of the current nonproliferation regime and to suggest a means of correcting one of the fundamental flaws of this regime. These games will rely on the concept of Perfect Bayesian Equilibrium (PBE), in which a player of unknown type plays a strategy that sends a signal to the other player. The second player revises his belief about the type of player he is facing, based on the action taken by the first player, and then plays his best response to that strategy.[3] Each game will have two players: player B (the big state or world hegemon, i.e. the United States) and player M (a small state whose intentions regarding nuclear weapons proliferation is unknown). In all games, player M may be of two types: a rogue state (R), that is actively pursuing a nuclear weapons program, or a peaceful state (P), that does not have any intention to pursue a nuclear weapons program. These are signaling games of incomplete information because player M's type is unknown to player B, and player M makes the first move in all games. After player M has moved in all games, player B has the option of either sanctioning player M or not.

The games reveal that the structure of the NPT fails to separate rogue states from peaceful states, unless targeted sanctions are employed that sufficiently diminish the benefits that a state receives from being a signatory of the NPT. Such targeted sanctions could control the materials and technologies entering a country suspected of developing nuclear weapons, or entail the physical elimination of suspected nuclear sites. Such targeted sanctions would thereby significantly reduce a state's ability to develop a nuclear weapons program. The NPT's failure to separate rogue from peaceful states is consistent with the evidence that Iraq, Iran and other states

[3] The motivation for this basic model was first conceived in a final paper that I wrote for Professor Ashok Rai's ECON 385 class in Spring 2005 and draws on Dutta's *Strategies and Games* (1999).

signed the NPT and proceeded with the development of nuclear weapons program. The treaty is unsuccessful because the rewards associated with membership are so high and the impact of inspections and sanctions so low that states seeking nuclear weapons technologies derive huge benefits from belonging to the NPT—not the least of which is the ability to convince the rest of the world that they are not pursuing any illicit weapons. My models reveal that, under the original structure of the NPT, only highly costly, targeted sanctions can cause potential proliferators to separate themselves from nonproliferators.

I also develop a model of the Additional Protocol, which was crafted to strengthen the ability of the NPT to detect nuclear weapons proliferation. I find that this structure also fails to separate rogue states from peaceful states, unless the probability of detecting a state's rogue status is significantly high, in which case the rogue state is unlikely to sign the AP at all. The AP structure is still fairly new; thus countries are still in the process of signing and it is difficult to know whether a state that has not yet signed the AP has deliberately chosen not to do so or has not yet made a decision. However, Iran, a state that is believed to be pursuing a nuclear weapons program, signed the AP in 2003. Because the AP was designed to increase the ability of inspections to detect weapons proliferation, it would seem that a state seeking nuclear weapons technology would not wish to sign the AP. According to the model that I construct, Iran must believe that the probability of detection under the AP structure is still low—a conclusion that does not bode well for the authority of the international nonproliferation regime.

I finally assert that the current NPT-AP structure fails to either prevent proliferation or separate proliferators from nonproliferators because of a fatal design flaw: under the current NPT-AP structure, states that do not sign the AP may still remain within the NPT and receive nuclear energy technologies. Thus a state that wishes to develop nuclear weapons and that believes that there is a significant probability of IAEA inspections detecting its illicit activity can simply choose not to sign the AP and retain the benefits of peaceful nuclear energy transfer. This is why the AP model that I construct is able to separate proliferators from nonproliferators only when the probability of the inspections' success is very low. I finally construct a normative model of the NPT-AP structure in which a state must choose whether to sign the AP or to leave the NPT and lose all the benefits of peaceful energy transfer. Structuring the nonproliferation regime in this manner, and imposing a deadline by which states must sign, yields only two possible equilibria: both the proliferator and nonproliferator will sign the AP, despite a significant probability of detecting the proliferator's status; or the proliferator will choose to leave the NPT if the probability of being detected is too high, thereby signaling his status as a proliferator to the international community, since there are no other possible equilibria under which any small state would choose not to sign the AP.

Scholars have pointed out the inefficacy and inefficiency of the NPT from its inception. Even in the 1970s, shortly after the treaty's inception, Albert Wohlstetter predicted the ability of countries to legally exploit the provisions of the treaty to develop nuclear weapons (1976-77). Richard Betts went so far as to predict that the NPT would not serve to prevent proliferation, but merely act as a symbolic commitment among the international community (1977). During the 1990s, in the wake of the end of the Cold War, some advocated the abandonment of the treaty and the need for a new framework for preventing proliferation (Carpenter 1994, Ollapally and Ramanna 1995). But by the turn of the twentieth century, the treaty is viewed by some as a flawed but still useful facet of a world nonproliferation regime (Walker 2000, Ozga 2000).

My suggestion for restructuring the NPT is unique among those solutions produced by critics of the NPT. Betts asserts that a commitment to nonproliferation requires a flexible policy for managing potential proliferators (1977), while Carpenter stresses the need for the United States to lower its own arsenal of nuclear weapons and promote nuclear-free zones as a means of curtailing proliferation (1994). Manning also emphasizes the need for a reduction in arsenals, as well as a heightened control over the production of plutonium, which could be used in a nuclear weapon (1997-98). Finally, those that support the NPT as a means of ensuring world security stress the need for a consistent and clear interpretation of the treaty's provisions (Ozga 2000), a strengthening of current enforcement and inspection mechanisms in order to curb proliferation, and a movement towards regional peace agreements (Walker 2000). The AP was designed to increase the enforcement and inspections provisions of the NPT, but the normative model that I construct presents a new and better method for curbing proliferation: the model forces states to choose whether to sign the AP and submit to additional inspections, or forfeit membership in the NPT. Such a structure can cause proliferators to separate from nonproliferators and thereby help curtail the development of nuclear weapons.

Many previous models of nuclear proliferation have examined competition in the nuclear arena as arms races, often structured as prisoner's dilemmas. Classic nuclear strategist Thomas Schelling characterizes an arms race as an interactive buildup of weapons by two actors, wherein each side responds to the other side's weapons buildup (Schelling and Halperin 1961, 34). Many theorists have conceived of arms races as variations on prisoner's dilemmas. Downs and Rocke (1990) provide an excellent review of Cold War-era prisoner's dilemmas modeling arms races and arms control arrangements. Brams and Kilgour's arms race model details the way in which both players in an arms race end up in a highly proliferated, socially inferior state because neither side can credibly commit to a lower level of armament (Brams and Kilgour 1988, 18). In a different model of de-escalation of conflict between two nuclear states, they find that compromise is most likely when each side believes its opponent is highly likely to retaliate (Brams and Kilgour 1988, 27), which is similar to the result expected from deterrence theory. Yet their series of models is based on a world in which two powers are racing at levels of relative parity in capabilities—not the case in which the United States hopes to influence the much weaker North Korea.

The large bulk of game theoretic literature addressing nuclear weapons therefore analyzes nuclear proliferation in terms of arms races and negotiations of arms-limiting treaties. Many of these models address optimal levels of weapons attainment between two powers capable of reaching parity in nuclear weapons, while others focus on the utility the two actors have from their weapons stocks and the associated level of domestic spending. My research, however, focuses on an entirely different element of arms accumulation. I construct models of the current structure of the world nonproliferation regime to demonstrate why current frameworks for nonproliferation not only fail to prevent proliferation but may actually facilitate it. I show that the current structure of the NPT-AP regime fails to separate players that plan to pursue nuclear weapons from those that do not, and thus grants similar benefits to both types of state.

In sum, the problem of small, marginalized states acquiring nuclear weapons is a very real challenge with which the world community is currently grappling. North Korea may have as many as six to a dozen nuclear warheads in its possession, and as a state branded a rogue actor in the international system, it would seem to have very little investment in acquiescing in the current international order. Iran is also believed to be developing a nuclear program of its own, despite protests from the international community. The games that I construct reveal the fundamental flaws of current efforts to prevent nuclear proliferation and present a new design for a nonproliferation framework that can better separate rogue states from peaceful states and thereby help to prevent further proliferation. The remainder of this paper is organized as follows. Section 1 presents models of the structure of the NPT and demonstrates why this treaty has been unsuccessful in preventing proliferation by some small states. In section 2, I model the Additional Protocol, which was enacted to fix some of the weaknesses of the NPT but is also a flawed framework for preventing nuclear weapons proliferation. Section 3 presents a normative model of the NPT-AP that is better able to separate rogue from peaceful actors and thereby better able to prevent proliferation. I conclude in section 4 and offer suggestions for future research.

Related Literature

Traditional, Cold War analyses of nuclear proliferation envisioned the competition between proliferators as an arms race between two actors. Arms races are often modeled as prisoner's dilemmas, in which each country increases its stock of weapons because it cannot be sure of its opponent's intentions. The two countries thus end up at a socially sub-optimal outcome relative to what would have been achieved had the two sides agreed to lower absolute levels of weapons. Any agreement to maintain lower levels of weapons would be inherently unstable, given the advantages to be gained by having more weapons than one's opponent (Brams and Kilgour 1988).

There are many theories to explain why states choose to pursue nuclear weapons programs, even in the face of hostility from the international community. Although a desire for national security is often a strong factor motivating the pursuit of nuclear weapons (Pierre and Moyne 1976), a state may pursue nuclear weapons for reasons seemingly unrelated to its national security. The question of whether nuclear weapons provide a state with additional security is by no means uncontested.[4] For some states, the pursuit of nuclear weapons is a way to build their

[4] For a more in-depth analysis of the reasons why states pursue nuclear weapons programs, and of the efficacy of nuclear weapons in providing a state with greater security, see Appendix A.

own prestige or status, both within the international community and in the eyes of their own citizens (Bracken 2003). For other states, the pursuit of nuclear weapons may be motivated by domestic political conditions. A state may use such a nuclear weapons program to distract attention from domestic problems or as a means of promoting nationalism (Pierre and Moyne 1976).

A state that develops a nuclear weapons program is likely to be motivated by a combination of these and other factors. An analysis conducted by Sonali Singh and Christopher R. Way (2004) examines a variety of factors that may contribute to a state's decision to pursue a nuclear weapons program. The pair constructed an event history model (also known as a hazard model), which predicts the probability of the occurrence of an event—in this case, the likelihood of a state being at one of four stages of nuclear weapons development. The dependent variable in this study was thus one of four different phases of nuclear weapons acquisition, the most advanced of which was the testing of a nuclear weapon, and the least of which was no interest in pursuing nuclear weapons. The data set followed 154 countries from 1945 to 2000; each country was classified as belonging to one of the four proliferation categories in each year. The authors examined three different categories of explanatory variables: technological determinants (e.g. industrial capacity), external determinants (e.g. the presence of a hostile neighbor), and internal determinants (e.g. the domestic political environment). The authors find that a state's overall level of industrial development and the existence of an enduring rivalry with another country are strongly and positively associated with an increasing likelihood to explore or acquire nuclear weapons. They also find that increasing per-capita GDP is a particularly strong predictor of nuclear weapon ambition for countries with low absolute levels of development, suggesting that a country must meet a certain base level of economic development in order to pursue nuclear technology. They also find that the absence of a security guarantee from a great power is also associated with the decision to "go nuclear," suggesting that states that feel ostracized by the world superpowers are more likely to engage in the development of nuclear weapons technology; however, this coefficient was not statistically significant. Finally, the authors find that a low level of integration in the world economy is also associated with the pursuit of a nuclear program. The authors argue that any actions that increase the threat environment that a state faces—including actions taken to disarm a state—will only encourage nuclear weapons proliferation. The results of Singh and Way's study largely support the political science theories of nuclear proliferation: that states are motivated by a hostile threat environment compounded by a lack of great-power security guarantee.

Regardless of the reasons that a state chooses to develop nuclear weapons, it is important to consider the strategic considerations that govern their use—specifically, their ability to deter the use of nuclear weapons against one's own state by another power. The study of deterrence, like many of the theories that surround the proliferation of nuclear weapons and arms races, developed in the context of the Cold War arms race between the United States and the Soviet Union. Theorists offer different perspectives on the efficacy of nuclear weapons in preventing an attack against one's state, and the conditions under which deterrence is likely to be successful. Such theories are recently being extended to include situations in which two different states have vastly different levels of nuclear weapons—i.e. in a standoff between the United States, with an extensive nuclear arsenal, and North Korea, with only a handful of weapons. Analyzing the efficacy of nuclear weapons for achieving security is important for any consideration of forced disarmament, as it is always possible that a conflict with a nuclear-armed state could lead to the use of the weapons whose elimination is sought. Indeed, the possibility that a small state might face a nuclear weapons attack may actually increase the incentives for such a state to develop nuclear weapons, so that it may hit its enemy with the most destructive force possible (Snyder 2003).

The crux of deterrence, as understood during the Cold War, lay in the possession of a secure second-strike capability. This meant that, in order to convince the opponent not to launch its nuclear weapons, a state needed to possess enough weapons such that it could both survive an initial attack by its opponent, and inflict a devastating attack on the opponent that would invite retaliation and thereby lock the two sides in a nuclear war. Presumably, such a war would effectively destroy both states and their populations. Under such a structure, when both sides possess this second-strike capability, launching a strike against one's enemy is irrational and foolish because it essentially invites one's own destruction. Thus, each country renders its weapons unusable by making the cost of its opponent's attack unbearably high—nuclear weapons thereby become essentially useless except as a means of deterring one's opponent from using them (Freedman 1986, 753-55). Such an equilibrium in which both sides have the ability to inflict devastating damage on the other side, even after sustaining a nuclear attack, is known as Mutual Assured Destruction (MAD), a term that came into use in 1964 (Freedman 1986, 757). In the twenty-first century, new proliferators do not possess the ability to inflict massive damage on an opponent, and opinions differ about whether nuclear weapons are successful in inducing caution in opponents.[5]

There have been a number of different ways of modeling nuclear proliferation. As mentioned above, one of the most common ways of modeling nuclear proliferation has been to

[5] For a more detailed description of deterrence theory and its applicability to twenty-first century proliferation, see Appendix A. In Appendix B, I construct a series of simple, sequential games to illustrate the principles of deterrence theory and the realities of deterrence in situations in which two states possess vastly different nuclear weapons capabilities.

evaluate arms races based on prisoner's dilemma models. Downs and Rocke develop their own models for examining arms control and cooperation. Their basic model is based on the utility that a state attaches to the levels of weapons held both by their own country and by the opposing country. They find that arms control is most likely to be successful when each side in the agreement is able both to evaluate its own utility and to understand how the other side assesses its utility function, since in their game the two opponents may have different preferences and utility functions (Downs and Rocke 1990). Of course, in real life there is no way of guaranteeing that you know exactly what your opponent's utility function may be, particularly if the opponent exhibits some of the "irrational" behavior posited by Payne (2003). Arms control agreements are fundamentally difficult to negotiate because they address state assets that are highly important, highly secretive, and highly technical. Both sides to an arms agreement will be inherently uncomfortable with divulging secrets so closely linked to their national security, and the diplomats who actually handle such arms control negotiations may not have enough technical knowledge in an ever-changing field to be able to execute the best agreements (Schelling and Halperin 1961, 82-83).

The idea that both players must know and understand both their own and their opponent's utility functions and calculations is similar to a result obtained by van der Ploeg and de Zeeuw (1990). Their Nash Equilibrium model of arms accumulation incorporates the tradeoffs that a nation must make between arms (guns) and other national goods (butter) when choosing the level of weapons it will produce. In an examination of cooperation in arms control agreements, they find that the best bilateral treaties are those that allow monitoring by each side of its opponent's weapons stock. Such transparency directly conflicts with Schelling's statement that treaties are difficult to negotiate precisely because they involve sensitive secrets about a nation's security; on the other hand, complete transparency in weapons accumulation could serve as a proxy for the transparency of motives and utilities mandated by Downs and Rock, if we assume that a willingness to accept weapons monitoring would signify peaceful intentions and that there would be a way to monitor "cheating" on the treaty itself. The vast literature on nuclear proliferation is explored in greater detail in Appendices A and B.

1. Failures of the NPT Regime

The following games are designed to illustrate the conditions under which countries such as Iraq, Iran, North Korea and Libya pursued nuclear weapons technologies while choosing to remain within the framework of the NPT. The models therefore illustrate why the incentive structure of the NPT fails to deter states from pursuing nuclear weapons and offer insights into how the system may be improved to prevent proliferation.

The Treaty on the Non-Proliferation of Nuclear Weapons (NPT) was completed and offered for signature in 1968. The treaty grew out of the fears that accompanied the introduction of the atomic bomb at the end of WWII, and out of the recognition that nuclear energy technologies could be converted into weapons production capabilities. In 1953, Eisenhower launched his "Atoms for Peace" proposal, which ultimately prompted the creation in 1957 of the International Atomic Energy Agency (IAEA), designed for both the "promotion and control of nuclear technology" (UN, Background Information 2005). As of 2005, 188 states were signatories of the NPT; notably, Israel, India and Pakistan have never signed the treaty and North Korea withdrew in December of 2003 (UN, BI 2005). All of these states are believed to possess or to be in the process of attempting to acquire nuclear weapons (Cohen and Graham 2004).

The express goals of the treaty were to prevent the "wider dissemination of nuclear weapons," and to grant states access to the "benefits of peaceful applications of nuclear technology" (UN, NPT). The ultimate goal of the treaty, which remains unmet, is the eventual elimination of nuclear arms around the world (Lewis 2004, 246). The treaty recognizes the five states that had tested weapons by 1 January 1967 as Nuclear Weapons States (NWS), all of which signed the treaty and include the United States, France, the United Kingdom, China and the Soviet Union. Under Article I of the treaty, NWS agree not to transfer nuclear weapons to non-nuclear weapons states (NNWS), and not to assist such states in the development of nuclear weapons. Accordingly, the NNWS agree not to solicit or accept any such offers of assistance, nor to seek to develop nuclear weapons on their own. The treaty provides for the transfer of peaceful, nuclear energy technologies from the NWS to the NNWS, but those who receive such technology transfers must agree to inspections by the IAEA to ensure that such states are not violating the terms of the treaty (UN, NPT 2000). The framers of the treaty had in mind that states might seek to "[divert] nuclear energy from peaceful uses to nuclear weapons or other nuclear explosive devices" (UN, NPT 2005). Accordingly, countries that sign the NPT agree to allow inspectors to examine their nuclear facilities and agree to place their nuclear fuel under safeguards, to ensure that such materials are not being diverted to military applications (Lewis 2004, 247).

The treaty's key weaknesses lie in its powers of enforcement (or lack thereof). Although the treaty does provide for inspections by the IAEA, Article III stipulates that such inspections must, "avoid hampering the economic or technological development" of the state under investigation (UN, NPT 2000). The inspections regime generated by the NPT has been manipulated by many countries, including Iraq, which deliberately deceived inspections teams while pursuing clandestine nuclear weapons programs. Since the IAEA was only required to visit those sites declared as nuclear facilities by the nation under inspection, the Iraqis merely limited scheduled IAEA access to areas of their nuclear facility that were not involved in illicit activities (Kay 1995). Similarly, the treaty provides no explicit punishment mechanisms for nations that violate the terms of the treaty. Finally, the treaty permits states to withdraw, as North Korea did in late 2003, when "extraordinary events…have jeopardized the supreme interests of its country" (UN, NPT).

States that pursued nuclear weapons while members of the NPT manipulated this membership to their advantage. Interviews with Iraqi scientists following the First Gulf War indicated that Iraq deliberately chose to remain within the NPT because its leaders thought that doing so would help to hide Iraq's pursuit of nuclear weapons and suggest to the rest of the world that the country was not engaged in illicit research (Kay 1995). Iraq and other proliferators that chose to remain within the NPT used their status as NPT signatories to attempt to signal to the rest of the world that they were not involved in any nuclear weapons research.[6]

The incentive structure of the NPT has therefore failed not only to prevent potential proliferators from being welcomed into the international nonproliferation regime, but it has also failed to prevent states within the NPT from developing nuclear weapons programs. In fact, membership in the NPT has actually *facilitated* the development of weapons programs: under the treaty, states receive nuclear energy technologies and reactors that could be used to develop a nuclear weapons program. A state or group of actors intending to construct a nuclear weapon needs a source of nuclear fuel for the explosive device (either plutonium or highly enriched uranium); strong chemical explosives to help "trigger" the weapon; a source of neutrons to start the nuclear reaction inside the bomb (often a mixture of deuterium and tritium gases); the necessary other materials for the construction of the weapon's physical structure; and trained personnel capable of working with the nuclear materials and constructing the weapon.[7] The NPT regime and other technology-sharing endeavors have facilitated the dissemination of nuclear technologies around the world and the training of nuclear scientists. Additionally, the breakup of the Soviet Union and demise of its nuclear weapons programs has spawned a large

[6] For a more detailed discussion of proliferation by members of the NPT, including Iraq, see Appendix C.
[7] Information on how to construct a nuclear weapon is from Garwin, Richard and Georges Charpak, *Megawatts and Megatons: A Turning Point in the Nuclear Age?* (NY: Alfred A Knopf, 2001). This text also gives a good overview of the science underlying nuclear reactions and the use of nuclear energy.

body of unemployed nuclear scientists—thus the knowledge and expertise necessary for building a nuclear weapon is not very difficult to obtain in the twenty-first century. Much of the training that countries receive in peaceful nuclear energy technologies under the NPT may also be employed for the development of a nuclear weapons program.

The most central requirement for building a nuclear weapon and the component that is generally the most difficult to obtain is the nuclear fuel source. There are a few different designs employed in nuclear weapons, ranging in complexity and destructive capability. But a nuclear weapon requires either highly enriched uranium-235 (a specific isotope of uranium occurring at low prevalence in natural, mined uranium) or plutonium. The specific isotope of uranium necessary for a nuclear reactor or a nuclear weapon must be refined and extracted from naturally occurring, mined uranium. The enriched uranium used in nuclear reactors may be at a much lower level of refinement, but the facilities that refine uranium for use in a nuclear energy reactor may be employed to produce more highly-enriched uranium for use in a nuclear weapon. Plutonium is produced as a by-product of some nuclear reactions harnessed for the production of energy. Thus, the possession of a functioning nuclear energy reactor may serve as a means of producing fuel for a nuclear weapon. Many efforts to control nuclear proliferation therefore focus on recovering the products produced in nuclear power reactors—either the plutonium produced as a by-product of the energy reaction or any leftover uranium that may be reprocessed to produce highly enriched uranium suitable for the construction of a nuclear weapon. A state or group of individuals seeking to develop a nuclear weapons program is greatly aided by the possession of a functioning nuclear energy reactor that may help furnish the fuel for a nuclear weapon; additionally, states such as South Africa that possessed both a source of uranium and a uranium processing facility had a further advantage in the quest for nuclear weapons (Fig 1999).

Thus states that have nuclear energy reactors already in place—reactors such as those provided to states under the provisions of the NPT—have a head start in developing a nuclear weapon, because they have the materials and expertise that could be diverted to the development of a nuclear weapon. Such states also have the trained scientists necessary to run a nuclear reactor and that could divert such knowledge to the development of a nuclear weapon. It is much easier for a state to develop a nuclear weapons program when it has a nuclear power reactor, and it is much easier for a non-nuclear state to obtain a nuclear power reactor by belonging to the NPT. For these reasons, I assume in the following models that it is extremely difficult for a country to develop a nuclear weapon without belonging to the NPT, and that belonging to the NPT confers large benefits on those states that seek to develop nuclear weapons.

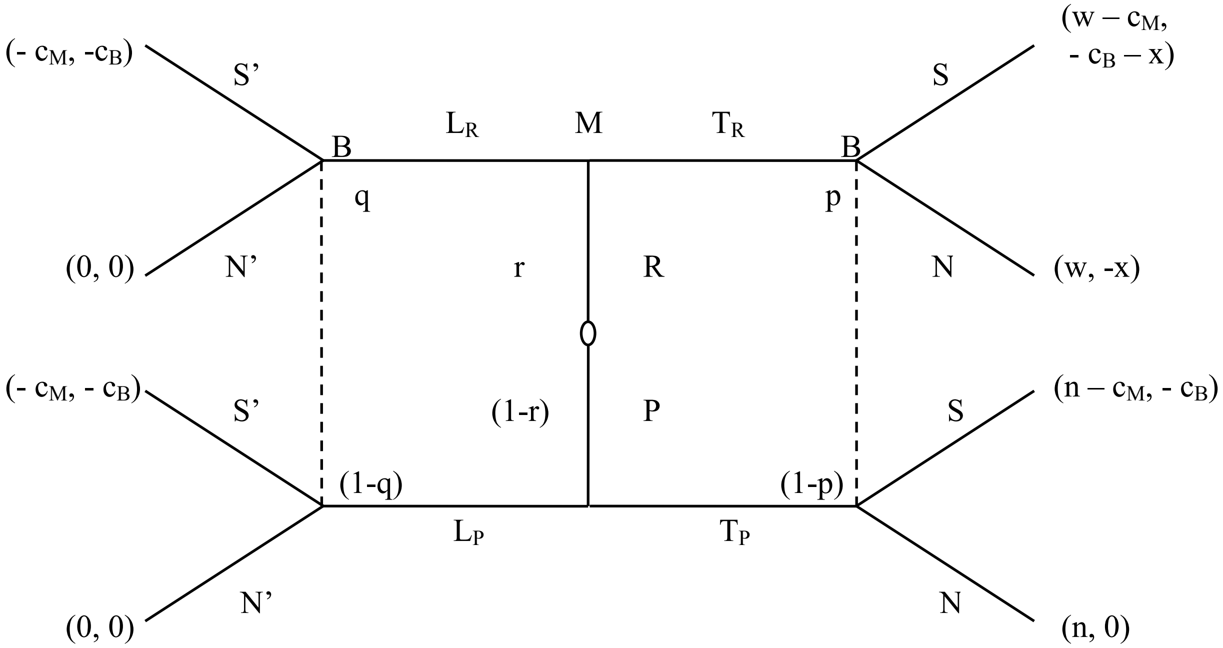
The following models are designed to illustrate the failures of the NPT incentive structure to separate potential proliferators from non-proliferators, precisely because of the fact that membership in the NPT generates benefits that aid countries seeking to develop nuclear weapons programs rather than effectively discouraging nuclear weapons proliferation.

Proposition A: Untargeted sanctions under the NPT structure fail to separate proliferators from non-proliferators.

This game has two players: the small country (M) that is already a member of the NPT, and the big country (B). Player M may be of two types: rogue (R) or peaceful (P). A rogue player will develop nuclear weapons capabilities while a peaceful state will not. This game is played at the point at which a country is already a member of the NPT—the rogue country has already made the decision to develop weapons and only chooses whether or not to remain a member of the NPT. Thus both types of player M have two strategies: they may either stay within the NPT (strategy T) or leave the NPT (strategy L). This game is one of incomplete information because player B is unsure about what type of player M he is facing; player M knows his type and moves first, thus sending a signal to player B that causes him to update his beliefs about player M's type.

While there was an inspections framework in place during the initial phase of the NPT, these inspections were extremely weak. States that signed the NPT were supposed to declare and to grant IAEA inspectors access to all nuclear facilities; however, the inspections process was ineffective and failed to detect the Iraqi nuclear weapons program that was housed at a facility examined by IAEA inspectors (Kay 1995). In this model, I therefore assume that the big country has no ability to detect what type of player M he is facing, and thus player M is rogue with probability (r) and peaceful with probability (1-r). Player B has the choice of enacting sanctions against M (strategies S and S') or not (strategies N and N'). These economic sanctions are purely punitive measures that do not have any impact on a country's ability to develop a nuclear weapons program. For example, the big country might choose to suspend trade with the small country, but this would not have a significant impact on the ability of the small country to continue with its nuclear weapons program. Obviously, the large country would not be seeking to punish all countries within the NPT, so this model and the rest of the models in this paper apply only to countries such as North Korea that are suspected of pursuing a nuclear weapons program. Player B would prefer sanctioning a rogue state over sanctioning a peaceful state, when such states are members of the NPT.

When a peaceful player M chooses to remain in the NPT, he receives a payoff of (n), representing the transfer of peaceful nuclear energy technology under the NPT framework. The rogue state receives a payoff of (w) when he remains within the NPT, representing both the transfer of nuclear technology that he would receive under the NPT, plus the ability to build a nuclear weapon that he receives because of the transfer of technology and ($w > n$). This model therefore assumes that the ability to develop a nuclear weapons program is strongly enhanced by belonging to the NPT. Sanctions impose a cost of (c_M) on player M, while sanctioning player M costs player B (c_B). In some cases, the value of (c_B) is likely to be quite small, particularly in relation to (c_M). On the other hand, there may be cases in which the cost of imposing sanctions on player M might be quite high for player B—for example, if player B chooses to punish player M by discontinuing the purchase of country M's oil. Such a strategy was pursued against Libya in the 1980s (Simons 2003, 132-33). When the rogue player M chooses to remain within the NPT framework, Player B incurs a disutility of (x), since the rogue player will be able to develop a nuclear weapons program and ($x > c_B$). The game is structured as follows, with (p) and (q) representing player B's belief that he is at a particular node, and payoffs (M, B):



(Figure A.1: Basic NPT Model)

NPT Model A: Solution

The model is solved based on the following assumptions: ($c_B > 0$) and ($w > n > c_M$). Under this assumption, the benefits of being a member of the NPT exceed the costs of sanctions for both types of player M. For any strategy pursued by player M, player B's best response is to play (N', N). For a separating equilibrium to exist in which the rogue player leaves the treaty and the peaceful player remains within the structure of the NPT, it must be the case that the payoff to the rogue player from leaving the NPT is greater than or equal to that for remaining within the NPT, and that by remaining in the NPT the peaceful player receives a payoff that meets or exceeds the payoff from leaving, given the strategy pursued by player B (N', N). In this model, given player B's strategy, the best response of both types of player M is to remain in the NPT. The model thus yields the following equilibrium:

$$\{(T_R, T_p), (N', N), 0 \leq q \leq 1, 0 \leq p \leq 1, r = p\} \qquad \text{(a.1)}$$

I will use this notation to describe equilibria throughout the remainder of this paper. The first set of letters in parentheses, in this case (T_R, T_p), describes player M's equilibrium strategy. The letters in subscript denote the type of player M pursuing a given strategy. Thus in this equilibrium, both the rogue (R) and peaceful (P) player M are playing strategy (T) and remaining within the NPT. The second set of letters in parentheses describes player B's strategy in equilibrium—in this case, player B plays (N', N) and never imposes sanctions. The rest of the equilibrium is described by a unique set of constraints that are placed on other variables in the model and that specify the conditions under which the equilibrium strategies hold true. In this case, the equilibrium specifies parameters for (q), (p) and (r), all of which are probabilities. In later models, the equilibria will also be described by parameters on different variables in the models that yield unique equilibria.

In this model, the only solution is a pooling equilibrium in which both the rogue and the peaceful player choose to remain within the NPT framework and player B never sanctions player M. Because of the costs associated with the imposition of sanctions, and because he cannot tell what type of player he is facing, player B never sanctions either type of player M. The positive incentives for joining the NPT encourage both the rogue and the peaceful player to sign the NPT and thus there is no separation equilibrium.

Solving the model under either the assumption that ($w > c_M > n$) or that ($c_M > w > n$) yields the same equilibrium as solving the model under the condition above. Because player B never imposes sanctions, the value of (c_M) never factors into player M's calculation of the best response to player B's strategy.

Note that the failure of the NPT to separate proliferators from non-proliferators holds true even when the cost to player B of imposing sanctions is zero, and the benefits of being a member of the NPT exceed the costs of sanctions for both types of player M ($w > n > c_M$). When ($w > c_M > n$), and ($c_B = 0$), there are three pooling equilibria in which both types of player remain within the NPT. There is also one separating equilibrium, but not the one that the international community would hope to enforce: when player B plays (N', S) and sanctions those players that remain within the NPT, then the peaceful state leaves the treaty and the rogue state remains a member receiving the full benefits of energy transfer. When ($c_M > w > n$), there are the same three pooling equilibria, plus a fourth in which both types of player M leave the NPT when player B plays (N', S). For a more detailed examination of the equilibria that arise when ($c_B = 0$), see Appendix D.

An obvious objection to the above model is the fact that, since player B seems to prefer that player M not remain within the NPT (except when player M is peaceful and player B does not impose sanctions, in which case player B is indifferent between the peaceful player M's status in the NPT), player B should seem to be better off if he could eject players from the NPT. In the above model, player B fares just as well or better when player M chooses to leave the NPT as when he chooses to stay in the NPT—that is, player B would receive a higher utility from sanctioning the rogue player M that chooses to leave the NPT than he would by sanctioning a rogue player M that remains in the NPT, and so on. Thus simply ejecting players from the NPT would seem to give player B higher utility. This would obviously solve the problem of players that manipulate their NPT status to produce a nuclear weapon.

So why does player B not simply eject players from the NPT? And why did the international community construct such a treaty structure in the first place? The answer to the latter question will become clearer later in this paper. The drafters of the NPT likely believed that they were constructing a mechanism that would be effective in curbing and monitoring nuclear weapons proliferation—proliferation that they may have believed would have occurred without any treaty structure in place. The framers likely believed that they had constructed a system similar to that which I present in the third section of this paper and that suggests a simple way of designing an NPT structure that can be effective in identifying and curbing proliferation (see section 3). The international community would not have adopted something that it believed would not work, even though many in the academic community have long been denouncing the treaty as a failure (Betts 1977, Carpenter 1994, Ollapally and Ramanna 1995 etc).

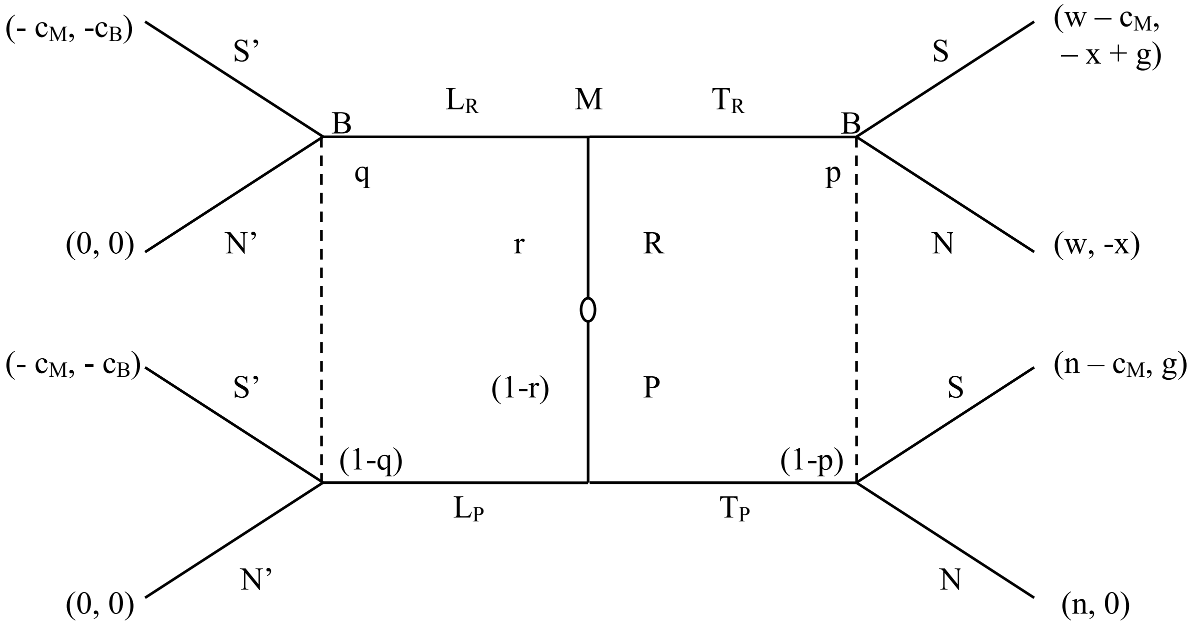
The answer to the question of why the "big state" or the international community as a whole does not simply eject states from the NPT is somewhat more complicated. First of all, the NPT has no explicit provision for ejecting a state that is suspected or even discovered to be developing nuclear weapons. In fact, there is no means of enforcement specified anywhere within the text of the treaty, and a state has the option to "[exercise] its national sovereignty" and voluntarily withdraw from the NPT (UN NPT 2000). The method of "enforcing" NPT infractions has been to refer proliferators to the IAEA and ultimately to the UN Security Council, as has recently occurred with Iran (Sciolino 2006). During the Cold War, a vote to expel a country from the NPT would have required the cooperation of the United States and the Soviet Union on a question of nuclear weapons proliferation, a sensitive subject for two countries engaged in their own nuclear arms race. No country has ever been expelled from the treaty, setting a precedent that becomes increasingly difficult to surmount as time passes. And the backlash that would result from the expulsion of a peaceful country from the NPT would likely be severe, especially given the fact that the treaty aims to spread peaceful nuclear energy technologies to growing nations in a period of dwindling world oil reserves. The United States would likely face an exceptionally high burden of proof in cases of suspected proliferation after the failure to discover weapons of mass destruction in Iraq, making any effort to chastise a potential proliferator difficult. Most pertinent for my models, the act of a "pre-emptive" ejection of a player from the NPT would mean that player B would presumably act without receiving any information from the small state—that is, without the informed player (M) making a move that sends a signal to player B about its type.

Finally, as mentioned above, the models in this paper do not apply to every country that is a member of the NPT but only to those that are suspected of pursuing a nuclear weapons program. Player B and the international community as a whole would not seek to eject all players from the NPT because there are likely very large benefits for the international community when a state is a member of the NPT. A state such as Switzerland that is not suspected of pursuing a nuclear weapon likely does not grant player B any disutility by belonging to the NPT—in fact, player B may believe it is in its interest for a state like Switzerland to belong to the NPT.

In sum, a variety of factors prevent the expulsion of a player from the NPT as a way to prevent the spread of nuclear weapons. The NPT treaty itself provides no explicit enforcement mechanism, and states may leave the treaty voluntarily when they feel that it is in their interest to do so. The traditional method of addressing states that seem to violate the terms of the NPT has been to refer such states to the UN Security Council, which then has the ability to draft a declaration, implement sanctions etc. International pressures make it difficult to expel states from the NPT without any hard evidence that a state is pursuing a nuclear weapons program, particularly given the failure to find WMDs in Iraq. Finally, the models in this paper are only applicable to those states that are believed to be in pursuit of nuclear weapons, and there are likely to be many positive utilities for both players M and B associated with membership in the NPT. Since the international regime does not currently expel states from the NPT, the models in this paper will assume that such a mechanism for curtailing proliferation is not a strategy choice available to player B.

Proposition B: When the large state receives a small net gain in utility from imposing sanctions on players that remain within the NPT, and the cost of sanctions to player M are sufficiently low, the NPT fails to separate proliferators from non-proliferators.

In the following model, player B receives a small net gain in utility from sanctioning a player that chooses to remain within the NPT. When a small state chooses to remain within the NPT, the large state may feel that sanctioning the small state will impact the state's ability to develop nuclear weapons—even though that may not actually be the case because the sanctions are not specifically directed at preventing nuclear weapons development. Even though the large state does incur a small cost from imposing the sanctions, player B believes that he receives a positive utility that offsets the cost of imposing sanctions. The following model therefore examines the equilibria that occur when the big state believes that it gains utility by sanctioning states that choose to remain within the NPT, because it believes that sanctioning such states will prevent them from developing nuclear weapons. In the following model, the net gain in utility that player B receives from sanctioning states that choose to remain within the NPT is represented by (g). Player B still receives the disutility of (-x) when the rogue player B remains within the NPT, because the sanctions that player B implements cannot directly impact the ability of player B to pursue nuclear weapons, and ($x > g$). When player M does not sign the NPT, player B again incurs the disutility of ($- c_B$) when he imposes sanctions and ($c_B > 0$). The game is structured as follows, with (p) and (q) again representing player B's belief that he is at a particular node:



(Figure B.1: "Bonus" Sanctions Model)

NPT Model B: Solutions

The model is first solved under the assumption that ($w > n > c_M$). Now, for any strategy pursued by player M, player B's best response is to sanction those players that remain in the NPT and to not sanction those that leave the NPT (N', S). For a separating equilibrium to exist in which the peaceful player remains in the NPT and the rogue player leaves the NPT, it must be the case that the rogue player M receives an equal or higher payoff when he leaves the NPT than when he stays, and that the peaceful player receives an equal or higher payoff by remaining in the NPT than leaving, given player B's strategy (N', S). Because the benefits for both types of player M of remaining in the NPT exceed the costs of sanctions, the best response of both types of player M is to remain in the NPT. Thus the model yields the following equilibrium:

$$\{(T_R, T_p), (N', S), p = r, 0 \leq q \leq 1\} \qquad \text{(b.1)}$$

Because player B now receives a small gain in utility when he sanctions a player that chooses to remain within the NPT, in equilibrium player B chooses to sanction players that stay within the NPT and to not sanction those that leave the NPT. Because the cost of the sanctions is less than the benefit both types of player M receive from remaining within the treaty, both types of player M still choose to remain within the NPT, even though doing so will result in sanctions. The benefits conveyed by the NPT on its signatories are likely to be so large that sanctions would have to be extremely costly to exceed the benefits that even a peaceful state receives from signing the NPT. The fact that countries such as Iraq and North Korea chose to remain within the NPT while developing nuclear weapons programs suggest that the assumption that ($w > n > c_M$) best reflects reality.

Nevertheless, it is still useful to solve the model under alternate assumptions. Solving under the assumption that ($w > c_M > n$) yields a different equilibrium. Player B still pursues a strategy of (N’, S), but now, for the peaceful player, the costs of sanctions exceed the benefits of remaining in the NPT. Since he will not face sanctions if he leaves the NPT, thereby receiving a payoff of (0), the peaceful player M’s best response is now to leave the NPT. Solving under the assumption that ($w > c_M > n$) thus yields the following equilibrium:

$$\{(T_R, L_p), (N', S), q = 0, p = 1\} \tag{b.2}$$

Thus the incentive structure has caused a separation of players, but it has driven the peaceful player M out of the NPT rather than allowing him to receive the benefits of the NPT structure. This is probably not an equilibrium that the international community would wish to achieve, since the peaceful state is being driven away from the NPT while the proliferator remains a signatory to reap its benefits.

Finally, the model is solved under the condition that ($c_M > w > n$). Player B’s best response to all strategies pursued by player M is still to sanction those players that stay within the NPT while not sanctioning those that leave the NPT (N’, S). Yet now the cost of sanctions exceeds the benefits of remaining in the NPT for both types of player M. Given the strategy that player B is pursuing, the best response for both types of players is now to leave to the NPT. Solving under the assumption that ($c_M > w > n$) thus yields the following equilibrium:

$$\{(L_R, L_p), (N', S), q = r, 0 \leq p \leq 1\} \tag{b.3}$$

Now, because the cost of sanctions exceeds the benefits of the treaty for both types of player M, neither type of player chooses to remain within the NPT.

In this model, player B believes that sanctioning a player that remains within the NPT will give it a small gain in utility, because he believes that sanctions will have an impact on player M’s ability to acquire nuclear weapons. Player B therefore chooses to sanction the players that remain within the NPT and not sanction those that leave, in all three cases. But, since they are not effectively targeted against proliferation, the sanctions do not have the ability to impact weapons production and may even drive the peaceful players out of the NPT if the cost of sanctions is too high (equilibria b.2 and b.3). Thus the relative values of the benefits from remaining within the treaty and the disutility incurred from sanctions determines whether player M chooses to remain within the treaty. If the goal of the big state is not to drive the peaceful states out of the NPT, then the cost of sanctions must be significantly low (equilibrium b.1), and the NPT again fails to separate proliferators from non-proliferators.

Both North Korea and Libya have been subjected to a variety of untargeted sanctions and thus NPT Models A and B are appropriate for examining these countries. Libya was a member of the NPT and remained a member while pursuing its nuclear weapons research (Bahgat 2004). The country relied on outside help for the development of its weapons program and faced heavy sanctions that likely limited its ability to develop a weapon; however, the country never chose to leave the NPT. These sanctions were implemented because of Libya's sponsor of terrorist activity, and thus were not targeted specifically to limit the ability to develop a nuclear weapons program. Thus the impact of sanctions likely did not diminish the benefits of the NPT enough that Libya chose to withdraw from the treaty. These sanctions likely generated a bonus for the countries implementing the sanctions, because they were designed to punish Libya for its terrorist activities. Thus NPT Model B, and equilibria (b.1) and (b.2) likely reflect Libya's participation in the NPT.

North Korea has also withstood a series of general sanctions. Until 2003, the country was a member of the NPT, but has since withdrawn from the treaty and presumably continued with its aspiration to develop nuclear weapons (IISS 2004). Until January 2003, however, the country had chosen to remain within the framework of the NPT to pursue its nuclear weapons program. It is of course unclear how the US views its payoffs in sanctioning North Korea, but it may be possible that the US worries about the impact of the sanctions on the people of North Korea and thus NPT Model B may not be appropriate. The pooling equilibria of NPT Model A may therefore be most appropriate for considering North Korea prior to 2003. Since 2003, North Korea may have felt that the cost of remaining within the NPT and being the subject of targeted sanctions and inspections have been raised, and may more closely fit the following model of targeted sanctions, in which the cost of developing nuclear weapons is significantly increased.

A note on the assumptions underlying the variables (w) and (g)

In the preceding model, player B receives a small net gain in utility when it sanctioned player M because it believes that doing so will eliminate the ability of player M to produce nuclear weapons. In later models, player B's disutility associated with player M's ability to build a nuclear weapon (-x) will be eliminated under certain circumstances. In all models, the rogue player M receives a large benefit, (w) from membership in the NPT, associated with an increased ability to develop nuclear weapons. This assumption is based on the premise that membership in the NPT facilitates the ability of player M to develop a weapon, unless carefully targeted sanctions are implemented—in which case the ability of player M to build a weapon is, in theory, greatly limited. This is why the variable (w) only appears when the rogue player M is a member of the NPT and why it is later eliminated under certain conditions. When player B believes it is implementing sanctions that limit the ability of player M to develop a nuclear weapon, he may receive the positive utility of (g), and/or the elimination of the disutility of (-x), depending on the model in question. This is the case because player B believes that membership in the NPT facilitates the small state's development of a nuclear weapon, and therefore that it is much more difficult for a state to build a nuclear weapon when it is not a member of the NPT (or is the recipient of carefully targeted sanctions). If the opposite were true—that it is easier to build a nuclear weapon outside the framework of the NPT than within it—then the payoff structures of these models would not apply. In such a case, player B should be focusing its energies on those states that are not members of the NPT, and player B would receive the disutility of (-x) when a rogue state *does not* belong to the NPT rather than when the rogue state does belong to the NPT. If it is easier for a state to build a weapon when it does not belong to the NPT, then player B should receive a bonus, (g) when it sanctions those states that do not belong to the NPT, since it would be assumed that such sanctions would hinder a state's ability to acquire a nuclear weapon.

Although games could be constructed to reflect such a situation, I believe that my models best reflect the realities of the international system. If all countries adhere to the NPT, including those that are permitted to possess weapons but prohibited from aiding other countries in acquiring such technologies, then it is generally much more difficult for a state to develop and acquire a nuclear weapon when it is *not* a member of the NPT. Thus the international system is on the right track in policing those states that remain members of the NPT, since membership conveys such large benefits to potential proliferators. As explained in an earlier section of this paper, the rules of the NPT facilitate the transfer of the technologies, materials and skills that would be useful in the construction of a nuclear weapon, and allow states to freely leave the treaty, as North Korea did in 2003, after it had received the benefit of years of technological assistance from nuclear weapons states. Iraq, Iran and Libya pursued nuclear weapons while members of the NPT, with somewhat mixed results.[8]

[8] More detailed histories of the nuclear weapons aspirations of North Korea and other countries can be found in Appendix C.

There are a few obvious examples of states that have developed nuclear weapons without belonging in the NPT—or are there? South Africa was able to develop a nuclear weapons program, before it became a member of the NPT in 1991. Yet South Africa received a substantial amount of assistance in the development of a uranium enrichment program and nuclear energy programs from the nuclear weapons states themselves, particularly the US and the UK, because of its large uranium deposits. South Africa also received substantial assistance from France and other nations. These acts of assistance violated the provisions of the NPT: had the weapons states more closely adhered to their own agreement, South Africa would have had a much more difficult time developing a nuclear weapons program that achieved only a low level of sophistication, largely due to targeted sanctions implemented in the 1980s (Fig 1999). Similarly, Libya and Iran both received help from outside sources in the pursuit of its weapons program, also in violation of the terms of the NPT (Bahgat 2004, Bidwai and Vanaik 2000, 73-5).

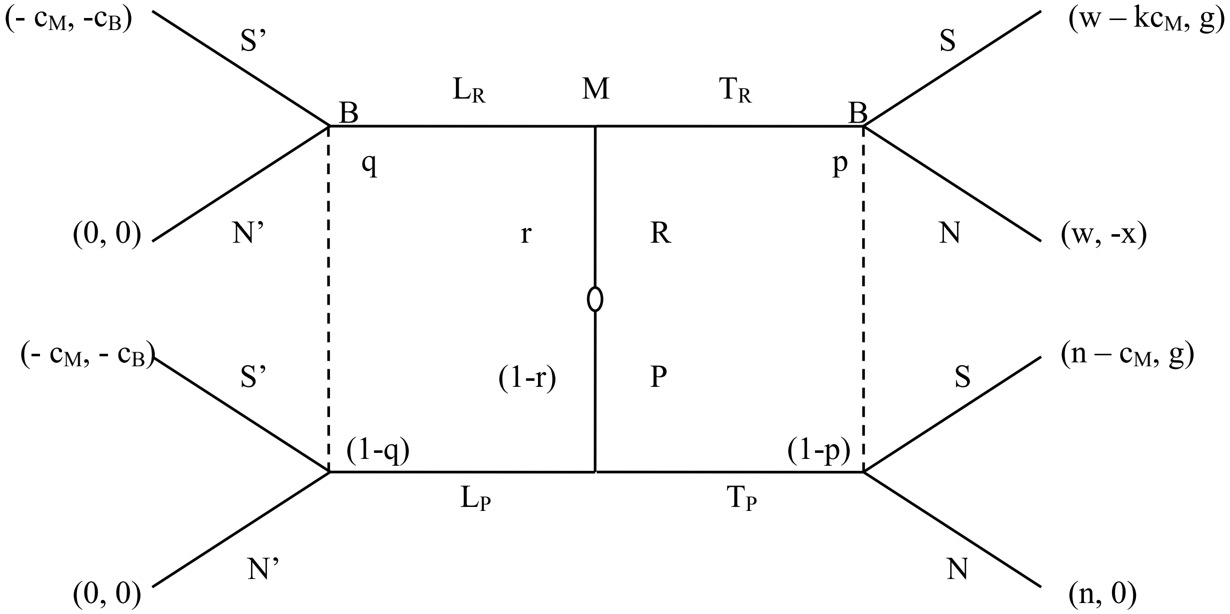
Other noteworthy examples of proliferation from outside the NPT include India, Pakistan and Israel. India and Pakistan have developed nuclear weapon programs without becoming members of the NPT, but their work has also been aided partly by outside actors that belong to the NPT, some of which occurred before the creation of the NPT and some of which violated the terms of the treaty (Bidwai and Vanaik, 61-5). Israel is probably the only country that has developed a nuclear weapons program without membership in the NPT and without substantial outside assistance. However, its program was already well under way by the time the NPT was implemented, and received help from France in the wake of the Suez crisis of 1956 (Cohen 1998, 53-9). Israel's program has been and continues to remain shrouded in secrecy and is likely not a useful comparison because it developed nuclear weapons prior to the existence of the NPT, and it is impossible to know if its program would have been facilitated by membership in the NPT.

Of states that developed nuclear weapons programs, only one likely did it without substantial aid from the direct benefits conferred by the NPT, or illicit help from those who received the benefits of the treaty. The evidence suggests that membership in the NPT and the technology transfers provided by the treaty facilitate the development of a nuclear weapons program. If all members of the NPT, especially all that possess nuclear weapons technologies, adhere to the provisions of the treaty and do not give help to states seeking to develop weapons programs, then it seems to be much easier for a state to develop a nuclear weapon by signing the NPT and receiving the benefits of nuclear energy technology transfers—and receiving the IAEA's stamp of approval for nuclear activities. I do not claim that it is impossible to develop nuclear weapons outside the NPT, nor that any of the strategies presented herein will completely eliminate a state's ability to develop nuclear weapons, but rather that the costs of developing a nuclear weapons program are much higher for a state that is outside the NPT, or under a regime of targeted sanctions, than for a state that is unsanctioned in the NPT. The implications of this assumption will become clearer in later models, in which the ability to impose certain sanctions eliminates the disutility that player B receives when player M is able to develop nuclear weapons (-x), and the rogue player M's increased ability to develop a weapon under the NPT (w) is similarly eliminated.

Proposition C: Under the framework of the NPT, targeted sanctions that sufficiently increase the costs of developing nuclear weapons can separate proliferators from non-proliferators.

In the following model, player B may impose sanctions that directly impact a state's ability to develop nuclear weapons. Such sanctions may be either economic or military and are specifically targeted to limit the ability of player M to develop weapons. Such sanctions might ban the import of certain types of mechanical parts that may be used to reprocess spent nuclear fuel from a reactor and thereby produce fuel for a nuclear weapon; or such sanctions might entail a bombing campaign that physically eliminates a suspected nuclear weapons facility. Such targeted sanctions may only be implemented when the small country is a member of the NPT—the big country does not have the authority under the treaty to administer the type of monitoring and inspections necessary to implement such targeted sanctions when the small country does not remain within the NPT. Additionally, the ability to make targeted import controls effective will require the cooperation of many different countries, which will be facilitated and legitimated by working through and within the international framework of the NPT. The targeted sanctions are designed to increase the costs of developing a nuclear weapon and thereby diminish the benefits that the small rogue state receives from remaining within the treaty. Because the large state cannot tell what type of state he is facing, he would impose the targeted sanctions on both types of player M that choose to remain within the NPT framework; yet the impact of these targeted sanctions would not be the same for both types of player M, since they are designed to directly impact activities associated with the construction of a nuclear weapon. Presumably, then, the cost of the targeted sanctions would be higher for the rogue player M than for the peaceful player M.

When player B implements the targeted sanctions, he assumes that he is eliminating the ability of player M to develop weapons and thus his disutility of (-x) is eliminated when he imposes targeted sanctions on the rogue player M. The additional cost of such targeted sanctions for the rogue player M is represented by (k) in the following game, where the cost to the rogue player M from receiving targeted sanctions is now ($-kc_M$), and ($k > 1$). As in NPT Model B, player B receives a small gain in utility from sanctioning the players that remain within the NPT (g), again because he believes that the imposition of sanctions will have a negative impact on the ability of the state to develop nuclear weapons, and ($x > g$). The strategies remain the same as those in the previous games, and the variables representing payoffs are the same as those for NPT Model A. When player B imposes targeted sanctions on players that choose to remain within the NPT, he is not guaranteed to eliminate the ability of the small state to develop nuclear weapons, but rather increases the costs of doing so. Thus the rogue player M retains the (w) payoff associated with remaining in the NPT but incurs a cost of ($-kc_M$). The peaceful player M does not incur the extra cost represented by (k) because he is not trying to develop weapons and thus the targeted sanctions do not impact his peaceful activities. For example, if the targeted sanctions involve a ban on the sale of certain machine parts needed to produce a weapon, such a ban would not place an extra cost on player M because he is not trying to import such machine parts for the construction of a weapon. The model is structured as follows:



(Figure C.1: NPT Targeted Sanctions Model)

NPT Model C: Solutions

The model is solved under the assumption that ($w > n > c_M$). Player B's best response to all strategies pursued by player B is to sanction those players that remain in the NPT and not sanction those that leave the NPT (N', S) because of the presence of (g). For a separating equilibrium to exist in which the rogue player leaves the NPT and the peaceful player remains in the treaty, it must be the case that the rogue state's payoff from leaving the treaty meets or exceeds that for being a member of the treaty, and that the peaceful state receives an equal or higher payoff by remaining in the treaty rather than leaving, given the strategy pursued by player B. Given the relative values of (n) and (c_M), the peaceful player M's best response to player B's strategy is to remain in the NPT. The rogue player M's best response to player B's strategy now depends on the value of (k). If the benefits of NPT membership under sanctions exceed the payoff that the rogue player M receives from leaving the NPT, then he will remain in the treaty; otherwise, he will leave the treaty. Thus the rogue player M will remain in the treaty when $[(w - kc_M) \geq 0]$, and will leave the treaty when $[(w - kc_M) \leq 0]$. Solving for (k) and under the assumption that $(w > n > c_M)$, the model yields the following equilibria:

$$\{(L_R, T_p), (N', S), q = 1, p = 0, k \geq (w/c_M)\} \qquad (c.1)$$

$$\{(T_R, T_p), (N', S), r = p, 0 \leq q \leq 1, k \leq (w/c_M)\} \qquad (c.2)$$

Equilibrium (c.2) is similar to those observed in the other models, in which both types of player M are induced to remain within the NPT because of the benefits from doing so. But in equilibrium (c.1), the two types of player M separate. When $[k \geq (w/c_M)]$, the rogue player M leaves the NPT, and the peaceful player M remains within the NPT. When the extra cost of the targeted sanctions is large enough and player B is playing (N', S), the rogue player M's best response is to leave the NPT. Note that according to these conditions, the larger the benefit from remaining within the NPT (w), the larger the value of (k) needed to drive the rogue player M out of the NPT. Because the targeted sanctions do not impact the peaceful player M, this player chooses to remain within the NPT. By raising the cost of targeted sanctions sufficiently high, the big state can cause the rogue player M to leave the NPT.

Now, Model C is solved under the assumption that $(w > c_M > n)$. Player B's best response to all strategies played by player M is still (N', S). But now the costs of sanctions exceed the benefits of NPT membership for the peaceful player M, and thus his best response to player B's strategy is to leave the NPT. The rogue player M's best response to player B's strategy again depends on the value of (k): he will remain in the treaty when $[(w - kc_M) \geq 0]$, and will leave the treaty when $[(w - kc_M) \leq 0]$. Solving for (k), under the assumption that $(w > c_M > n)$, yields the following equilibria:

$\{(T_R, L_p), (N', S), q = 0, p = 1, k \leq (w/c_M)\}$ (c.3)

$\{(L_R, L_p), (N', S), q = r, 0 \leq p \leq 1, k \geq (w/c_M)\}$ (c.4)

The value of (c_M) now drives the peaceful player M out of the NPT in both equilibria. As long as the value of (k) is sufficiently low, as in equilibrium (c.3), the rogue player M still remains within the NPT; when it is too high, both players are driven out of the treaty. The international community would not want to pursue equilibrium (c.3), since in this case the rogue player M retains the benefits of NPT membership while the peaceful player is driven out of the treaty.

Finally, the model is solved under the assumption that ($c_M > w > n$). Once again, player B's best response to player M's possible strategies is (N', S). The cost to the peaceful player M of sanctions still exceeds the benefits he receives from membership in the NPT, and thus his best response is to leave the treaty. Now, the "basic" cost of sanctions (c_M) exceeds the benefits of treaty membership for the rogue player as well. Since the value of (k) is strictly greater than one, the rogue player's best response to player B's strategy, given that ($c_M > w > n$), is to leave the treaty. The model now yields the following equilibrium:

$\{(L_R, L_p), (N', S), q = r, 0 \leq p \leq 1\}$ (c.5)

As in Models A and B, when the cost of sanctions is too high, both types of player M are driven out of the NPT.

Assuming that the basic costs of sanctions (c_M) are not higher than the benefits conveyed by remaining in the treaty (w, n), and that the additional cost of targeted sanctions for the rogue player M (k) is high enough, then the NPT has the ability to separate rogue from peaceful states. The rogue player will leave the treaty and the peaceful player will remain within the treaty when the additional cost of targeted sanctions (k) is sufficiently high (equilibrium c.1). By leaving the treaty, the rogue player cuts himself off from the technological advantages conveyed by the NPT and thus faces a much more difficult task in developing a nuclear weapon without outside assistance.

Targeted sanctions were employed against both South Africa and Iraq in efforts to halt the production of nuclear weapons in these countries. South Africa never signed the NPT and therefore does not fit these models exactly; but this targeted sanctions model could fit South Africa's case if the strategy options from which South Africa chooses are not whether to remain within the NPT or not, but rather whether to sign the NPT or not. If this is the case, then the country may fit the equilibria in which the rogue player decides not to be a member of the NPT in its pursuit of its nuclear weapons (equilibria c.1, c.4 and c.5). South Africa was forced to pursue its nuclear weapons campaign without the aid of officially permitted technology transfers under the NPT, and it is likely this fact, combined with targeted and general sanctions implemented against the country, that helped limit the sophistication of the country's weapons program, even though the sanctions were largely imposed too late to effectively halt weapons production (Fig 1999). South Africa is a unique case, however, because it possessed valuable uranium deposits and received a large amount of technical assistance in developing a nuclear program, despite the fact that it was not a member of the NPT.[9] For South Africa, these benefits made remaining outside the framework the preferable choice.

Iraq also faced a program of targeted sanctions that had a more direct impact than the sanctions imposed on South Africa. Iraq, however, decided to remain within the framework of the NPT while pursuing its weapons program, rendering its behavior more like equilibria (c.2) and (c.3). Although the targeted sanctions certainly increased the costs of developing a nuclear weapon, they were not high enough that Iraq chose to withdraw from the treaty and lose the payoff of (w), or to prevent the country from pursuing an advanced program that was only

[9] For more specific information on South Africa's nuclear program, see Appendix C.

months away from producing a functioning weapon at the time of the Gulf War in 1991 (Kay 1995).

Note: A "Nervous" player B implementing targeted sanctions fails to separate proliferators from non-proliferators.

If player B believes that it gains utility (g) only when it correctly sanctions a rogue player M that is a member of the NPT, then it no longer receives the bonus (g) from sanctioning the peaceful player that signs the NPT. Player B may feel nervous about incurring the outcry of the international community for sanctioning a player that is not pursuing nuclear weapons. Under these conditions, and when ($w > n > c_M$), then the model yields only two equilibria, in which both types of player M choose to remain within the NPT (and player B pursues two different strategies). Because player B is nervous about imposing sanctions on a peaceful state, he is unable to separate the rogue players from the peaceful players. A more detailed evaluation of the equilibria that arise when player B "nervously" imposes targeted sanctions is included in Appendix D.

This model of a nervous player B implementing targeted sanctions may be somewhat useful in considering the United States' current position on nuclear proliferation. After the failure to find WMD in Iraq, the United States has come under harsh scrutiny from the world community for invading a country that did not possess any illicit weaponry. This may lead the United States to be cautious in its implementation of sanctions against countries that it believes are pursuing illicit weapons program in the future. The reluctance to sanction small states whose weapons activities are unclear may continue to lead to equilibria in which proliferators choose to remain within the NPT.

Failures of the NPT: Summary

Under the current structure of the NPT, the implementation of untargeted, general sanctions against states suspected of pursuing nuclear weapons fails to separate the proliferators from the non-proliferators. Model A yields only pooling equilibria in which both types of player M choose to remain in the NPT. Even when player B receives a small net gain in utility from sanctioning those players that sign the NPT (g), when the costs of untargeted sanctions are less than the benefits of NPT membership, the structure fails to separate rogue players from peaceful players. Implementing targeted sanctions that raise the cost of constructing a nuclear weapon can be effective in separating rogue players from peaceful players. If the additional cost to the rogue player M of the targeted sanctions (k) is sufficiently high, the rogue player will reveal its type by leaving the NPT, while the peaceful player remains within the NPT structure. In reality, the NPT has done a poor job of detecting and punishing nuclear weapons proliferation. In the following section I model the Additional Protocol, which was designed to help correct some of the shortcomings of the NPT.

2. The Additional Protocol: A Misguided Attempt to Correct the Failures of the NPT

The IAEA's inability to discover Iraq's nuclear weapons program prior to the Persian Gulf War, during which time Iraq remained a party to the NPT, prompted a decision by the international community to try to enhance the powers of the nonproliferation regime. The Additional Protocol (AP) was designed with the intention of strengthening the IAEA inspection teams' abilities to detect illicit nuclear weapons activity. As mentioned in the previous section of this paper, the original NPT limited the IAEA to inspect only a narrow range of sites *declared* by the state under investigation. Inspections had to be announced cleared in advance with the state's government (Hirsch 1995, 142-3). Under the Additional Protocol, non nuclear weapons states (NNWS) are required to make broader disclosures about their nuclear facilities: a greater range of activities potentially involved in the production of a nuclear weapon must be declared to the IAEA. Thus facilities that were previously outside the range of required declaration, such as those that could manufacture parts needed for the enrichment of uranium, must be declared under the AP.

The most important element of the Additional Protocol is the fact that it grants the IAEA "complementary access" (i.e. "special inspection[s]") to sites not declared by the state to conduct environmental sampling (Hirsch 1995, 144-147). The hope for the AP is that it will grant IAEA inspectors access to a much wider range of facilities, without necessarily granting the state under inspection advance notice, and that the AP will therefore grant inspectors a greater ability to detect illicit weapons programs. By the end of November 2004, 61 states had signed and implemented the Additional Protocol, while 26 more had signed the protocol but not yet implemented it (Hirsch 1995, 161). Among the signatories of the Additional Protocol is Iran (Bowen and Kidd 2004, 257), about whose nuclear program the international community is currently very worried, and Libya, which was attempting to develop a nuclear weapons program from the 1980s to 2003 (Bahgat 2004, 389).

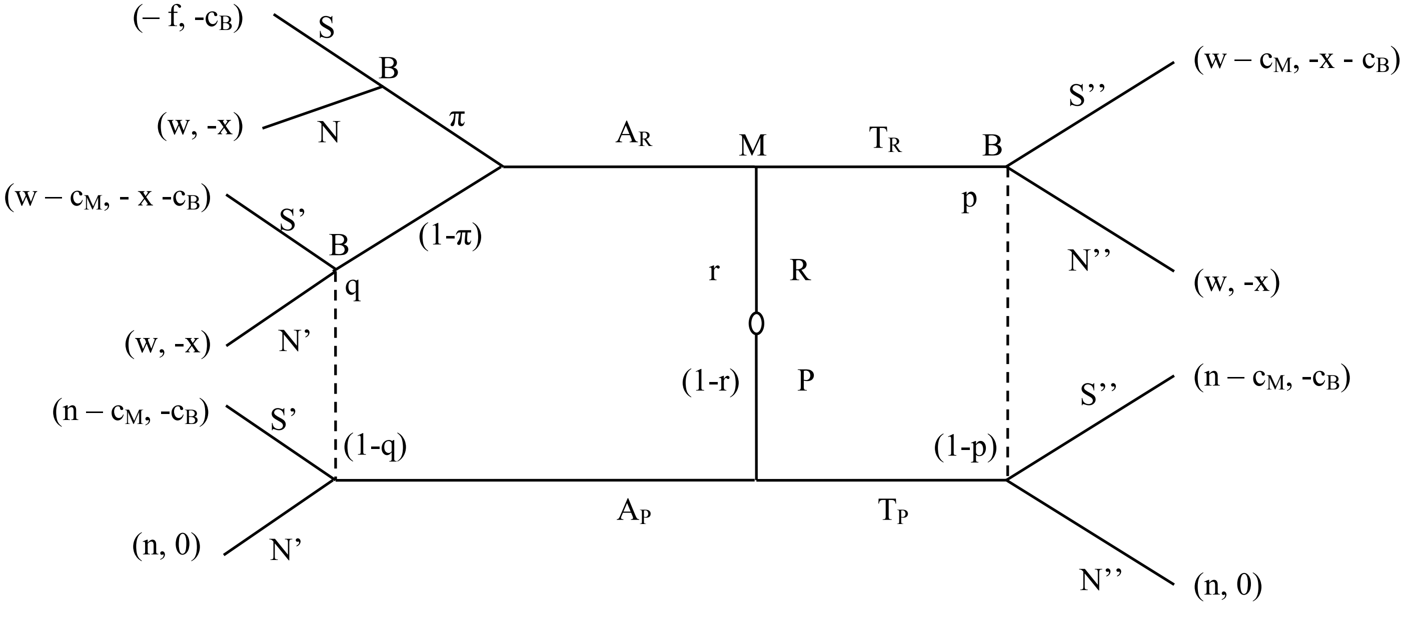
The fact that the AP provides the IAEA with a greater ability to detect illicit nuclear weapons programs should presumably prevent states that seek to develop nuclear weapons from signing the AP. The provisions of the AP should cause countries with intentions to build nuclear weapons to separate themselves from peaceful states by failing to sign the AP. For the purposes of adapting the NPT signaling game to reflect the conditions imposed by the AP, the following model assumes that the inspections regime has a chance to detect illicit nuclear weapons activity. The first model reflects the technical structure of the AP as it exists today.

Proposition D: The current structure of the AP is not an effective means of separating proliferators from non-proliferators.

Once again, the game involves two players: a small state, M, that is suspected of pursuing a nuclear weapons program and that may be either rogue (R) with probability (r) or peaceful (P) with probability (1-r); and a big state, B, that chooses whether to impose sanctions on the small country (strategies S, S', S'') or not (strategies N, N', N''). Unlike in the past game, the large state has a limited ability to detect what type of player M he is facing when player M chooses to sign the AP. In this game, the probability that inspections will be successful in detecting a rogue state's weapons program is represented by (π). Player B would like to punish a rogue player that he is able to detect. When player M does not sign the AP, the large state has no ability to detect what type of player M he is facing, as was the case in the NPT model. I assume that, while the inspections would have the ability to detect the positive presence of a nuclear weapons program, they could never *detect* the absence of a nuclear weapons program—that is, they could never determine with certainty that a player was peaceful, even though they could determine with certainty that a player was rogue. In this game, the only way to prevent a player from producing nuclear weapons is by detecting the state's true type and imposing sanctions—i.e., the sanctions imposed against a state of unknown type are not effective in curtailing a weapons program but are merely punitive.

In the following game, player M chooses whether to sign the AP (strategy A) or refrain from signing the AP and remain within the old framework of the NPT (strategy T). Under the current framework of the Additional Protocol, countries that sign do not receive any material bonus from doing so, unlike the NPT, which provides tangible transfers of nuclear technologies and materials to those non-nuclear weapons states that belong to the treaty. Singing the AP does not provide a country with a similar bonus, but rather permits the IAEA to conduct more extensive inspections on a country's nuclear facilities and requires the country to disclose a wider range of its activities to the agency. Thus countries submit themselves to a higher level of inspections without receiving any tangible benefits in return. Note that the right half of the game is identical to the right half of NPT Model A, in which player M chooses whether or not to remain a member of the NPT.

The payoff variables remain the same as those from the NPT games, with one addition. When the small country is detected as rogue and the large country decides to impose sanctions, the small rogue country receives a punishment or fine of (f), where ($f > c_M$). In this case, the big country is able to effectively eliminate a weapons program only when player M is detected to be a rogue state—under other conditions, the sanctions that the big state implements are only able to punish the state. Thus for player B, the payoff associated with player M's ability to develop nuclear weapons (- x) is eliminated only when player M is detected to be rogue and is sanctioned. This model also assumes that player B does incur a cost from implementing sanctions against player M ($c_B > 0$). Because there is no positive incentive to sign the AP, note that when the peaceful player M faces a strategy of (S', S'') or (N', N'') by player B, he is indifferent between signing the AP and remaining within the framework of the NPT. The peaceful country has nothing to hide by submitting to additional inspections, but since it does not receive any positive incentive to sign the AP, he is indifferent between signing and not signing. Similarly, the rogue player receives the same payoff when player B plays (S') as when he plays (S''), and from (N') and (N''). The rogue player M still faces, however, the possibility of being detected as rogue and thereby incurring the fine of (-f). The game is structured as follows, with (π) representing the probability that player B detects that player M is rogue, and (p) and (q) representing player B's belief that he is at a particular node:



(Figure D.1: Basic Additional Protocol Model)

AP Model A: Solutions

The model is solved under the assumption that ($w > n > c_M$). Player B's best response to any combination of strategies pursued by player M is to sanction the rogue player M that he is able to detect, and not to impose sanctions under any other circumstances. His best response is thus to play (S, N', N''). For a separating equilibrium to exist in which the rogue player M decides not to sign the AP and the peaceful player M does sign the AP, it must be the case that the rogue player receives an equal or higher payoff by leaving the NPT than he would receive by signing the AP, and that the peaceful player M must receive a payoff from signing the AP that is greater than or equal to that he would receive by not signing, given player B's strategy. Given that player B will only sanction the detected rogue player M, the peaceful player M is indifferent between signing the AP and not signing the AP under all circumstances, since he receives the same payoff (n) for both strategy choices. Thus the peaceful player M may either sign the AP (strategy A) or not (strategy T) in equilibrium.

The calculation for the rogue player M is slightly more complicated. Given that player B will sanction the rogue player M if he is detected, and that the rogue player would thereby incur the payoff of (-f), and given that player B would not impose sanctions in any other case, then the rogue player M would choose not to sign the AP if there were any chance that he would be detected as rogue. Thus the rogue player M will sign the AP only when the probability of being detected as rogue (π) is exactly zero, since in this case the rogue player would receive the payoff of (w) because player B is only imposing sanctions on those rogue players that he is able to detect. For any value of (π) greater than or equal to zero, then, the rogue player M will choose not to sign the AP (strategy T), and will sign the AP (strategy A) when the probability of detection (π) is exactly zero. Because the peaceful player M is indifferent between signing the

AP or not in equilibrium, and given player B's strategy, there are four possible equilibria in this game that encompass all possible strategies for player M:

$\{(A_R, T_p), (S, N', N''), p = 0, q = 1, \pi = 0\}$ (d.1)

$\{(T_R, A_p), (S, N', N''), q = 0, p = 1, \pi \geq 0\}$ (d.2)

$\{(A_R, A_p), (S, N', N''), 0 \leq p \leq 1, r = q, \pi = 0\}$ (d.3)

$\{(T_R, T_p), (S, N', N''), 0 \leq q \leq 1, r = p, \pi \geq 0\}$ (d.4)

The most critical aspect of this model is the fact that the rogue player M signs the AP (strategy A_R) only when the probability of being detected as rogue is exactly zero (equilibria d.1 and d.3). Because there is no bonus for signing the AP, there is no reason that the rogue player M would risk being detected by signing the AP when he can do just as well by not signing and remaining within the structure of the NPT. When the probability of being detected as rogue is greater than zero (equilibria d.2 and d.4), then the rogue player chooses not to sign the AP. Thus this model does yield the desired separating equilibrium, in which the rogue player chooses not to sign the AP and the peaceful player does (equilibrium d.2), but only because the inspections actually have a chance of detecting a player's rogue status. If the rogue player does not sign the AP, then the increased inspections capabilities provided by the agreement cannot be applied to the state. The peaceful player M is always indifferent between signing the AP and not signing when player B plays (S, N', N''), and thus the model yields all four possible combinations of strategies for player M, including a separating equilibrium in which the peaceful player chooses not to sign the AP and the rogue player does (equilibrium d.1).

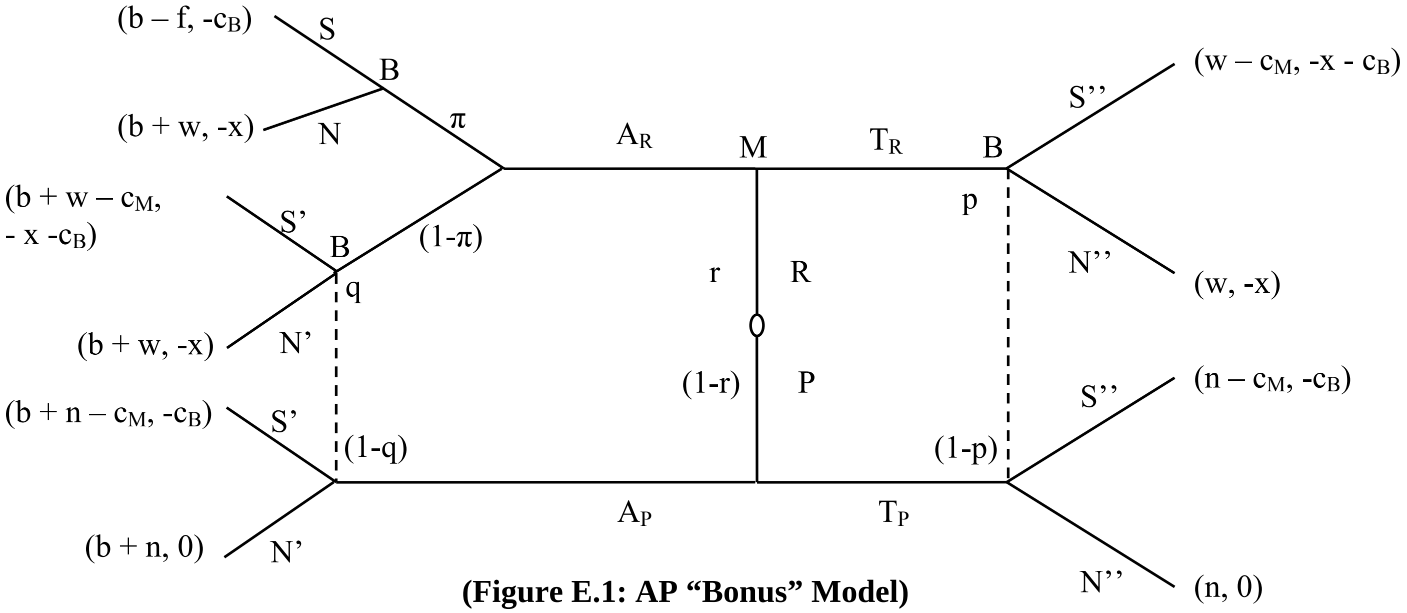
Note that solving the model under the assumption that ($w > c_M > n$), or under the assumption that ($c_M > w > n$), yields the exact same set of equilibria as solving under the condition above. Because player B always plays a strategy of (S, N', N''), the value of (c_M) in relation to (w) and (n) never has any bearing on the solutions to the model.

According to this model, if the goal of the AP is to induce potential rogue states to sign the agreement and thereby subject themselves to greater inspections of their nuclear facilities, then the very ability of those inspections to detect the rogue player's type causes him to not sign the AP and to remain a regular member of the NPT. Attempts to increase the ability to detect weapons proliferation in this case prevent those inspections from being implemented, because they deter the rogue state from signing the AP (when $\pi > 0$). If the goal of the AP is to cause rogue states to reveal themselves by not signing, then the current structure also fails because there is no deadline by which a country must sign. A country's status as a non-signatory may indicate a conscious decision not to sign the AP, or simply that the country has not yet "played the game" and made a decision about signing.

Iran chose to sign the AP in 2003, even though it seems clear that the country is pursuing a nuclear weapons program (or at least confusing the world about its nuclear aspirations). AP Model A suggests that the rogue state would only be willing to sign the AP if it believed that the probability of detection as rogue is zero. This may be true if the country believes it will be able to fool inspectors as Iraq was able to do; or if the country plans to expel inspectors. Iran did announce in February 2006 that it would terminate its relationship with the IAEA in response to the agency's resolution demanding greater transparency from the country (Sciolino 2006). If Iran believed that the chance of being detected as rogue under the AP was zero, then equilibrium (d.3) seems to fit with the behavior of Iran and other countries. On the other hand, the fact that Iran and many other states have signed the AP, even though the probability of detection is likely greater than zero, suggests that there is an aspect of the AP that this model is not capturing. In the next model, I explore the equilibria that arise when there is a positive incentive to sign the AP.

Proposition E: When small states believe that there is a positive incentive to sign the AP, and when the probability of detecting a rogue state is sufficiently high, then the AP may serve as a mechanism for separating proliferators from non-proliferators.

In the following model, states that sign the AP receive a bonus (b). A state may feel that signing the AP produces intangible benefits, such as the goodwill of the international community, or tangible benefits, such as increased trade with the large state. Or, states may feel that signing the AP grants them legitimacy in the eyes of the world community: that a willingness to undergo the increased inspections mandated by the AP suggests that the state is not pursuing an illicit weapons program. Thus both a peaceful and a rogue state may feel that there is a positive incentive to sign the AP. Because the bonus that a state receives from signing the AP may not have any monetary value, the value of this bonus is small: ($b < c_M$). The existence of the bonus thus denotes the small state's preference for being a member of the AP. All other payoff variables remain the same as those in previous models, and ($c_B > 0$). The game is structured as follows:



(Figure E.1: AP "Bonus" Model)

AP Model B: Solutions

The model is solved under the assumption that ($w > n > c_M$). Player B's best response to all possible strategies for player M is to sanction the detected rogue player M, and to not impose sanctions under any other circumstances (S, N', N''). For a separating equilibrium to exist in which the rogue player does not sign the AP and the peaceful player does sign, it must be the case that the rogue player receives a payoff from not signing the AP that is greater than or equal to that which he would receive for signing, and that the peaceful player receives a payoff from signing the AP that is greater than or equal to the payoff that he receives by not signing the AP. Player B will not impose sanctions against the peaceful player M, regardless of the peaceful player's strategy. Because there is now a bonus associated with signing the AP, the peaceful player M's best response to player B's strategy is to sign the AP. There are no longer any equilibria in which the peaceful player M chooses not to sign the AP.

The rogue player M makes a slightly more complicated calculation. When the rogue player M's type is not detected, then the rogue player M will not face sanctions whether he signs the AP or not, because player B is playing (N', N''); yet because there is now a bonus for signing the AP, the rogue player M would prefer to sign the AP when he is not detected. However, the rogue player M knows that by signing the AP, he would risk being detected as rogue under the enhanced inspections provided by the AP and receiving the payoff of (b – f). So, player M must decide what level of "risk" of detection (π) he is willing to accept in order to sign the AP. For the rogue player M to sign the AP, it must be the case that the expected value of the payoff that he receives from signing the AP meets or exceeds the payoff he would receive by not signing, given player B's strategy. The rogue player M will therefore sign the AP when $[(\pi)(b - f) + (1 - \pi)(b + w) \geq w]$, and will not sign when the reverse is true: $[(\pi)(b - f) + (1 - \pi)(b + w) \leq w]$. Solving for ($\pi$), and knowing that the peaceful player M always chooses to sign the AP, given player B's strategy, the model yields the following equilibria:

$\{(T_R, A_p), (S, N', N''), p = 1, q = 0, \pi \geq [b/(f+w)]\}$ (e.1)

$\{(A_R, A_p), (S, N', N''), 0 \leq p \leq 1, r = q, \pi \leq [b/(f+w)]\}$ (e.2)

In this game, there is finally a meaningful equilibrium that separates the rogue players from the peaceful players (e.1). Under these conditions, when the probability of being detected (π) is sufficiently high, the rogue player M does not sign the AP, while the peaceful player M does choose to sign the AP. The chance of being detected as a rogue player and sanctioned accordingly prevents the rogue player from signing the AP while the peaceful player M still chooses to sign. Note, however, that the probability of being detected as rogue is greater than zero, unlike in AP Model A in which the rogue state only signs the AP when the probability of detection as rogue is zero.

Because of the bonus associated with joining the AP, the rogue player is still willing to risk the inspections associated with signing the AP, when the value of (π) is sufficiently low (equilibrium e.2). In this pooling equilibrium, both players choose to sign the AP. The probability of detection (π) is sufficiently low that the rogue player M signs the AP even though there is a chance that he may be detected as rogue and severely sanctioned. As the fine from being detected as rogue and sanctioned (f) increases, so the probability of detection (π) must decrease in order for the rogue player M to sign the AP. Assuming that there is perceived to be a bonus, (b), associated with signing the NPT, then the second equilibrium (e.2) may represent Iran's decision to sign the AP in December 2003. Iran may believe that the probability of being detected is low, or that the fine from being detected and sanctioned is sufficiently low, that it decided to sign the AP.

Note that solving this model under either of the following assumptions yields the same equilibria as solving under the conditions stipulated above: ($w > c_M > n$) and ($c_M > w > n$). Because player B always plays the strategy (S, N', N''), the value of (c_M) never factors into player M's best response function. The model therefore yields the same two equilibria regardless of the value of (c_M) in relation to (w) and (n). Solving the model under the condition that ($c_B = 0$) produces a very similar set of equilibria—there are 3 similar separating equilibria in which player B plays a variety of different strategies, 3 pooling equilibria in which both types of player M sign the AP, and one pooling equilibrium in which both types of player M do not sign the AP. For more information on this variation of NPT Model B, see Appendix E.

Challenges of the AP: Summary

The AP was designed to correct some of the shortcomings of the NPT—mainly that the rules of inspections under the NPT were so weak that states such as Iraq could develop nuclear weapons undetected. AP Model A reveals that, under the literal provisions of the AP, rogue states would only sign the agreement if the probability of being detected as rogue was exactly zero. Since some countries that are considered to be proliferators have signed the AP, this suggests that states may perceive a positive incentive to sign the AP. When states do indeed perceive a benefit associated with signing the AP, as in AP Model B, then there can be a meaningful separation of players. Even when the probability of being detected as rogue is greater than zero, the rogue state may still be willing to sign the AP; or, if the probability of detection is perceived to be too high, the player will not sign.

This raises a question about the true goals of the AP. If the goal of the AP is for rogue states to sign the agreement and thereby subject themselves to inspection, then the agreement as it currently stands and as it is modeled in AP Models A and B is weak. Note that the pooling equilibrium in AP Model B (e.2) occurs only at a fairly low probability of successful detection, as the value of (b) in the numerator of the fraction determining the value of (π) is quite small in relation to (w) and (f). If the goal of the AP is to drive potential proliferators to reveal themselves, then the fact that there is no deadline by which states must sign and no punishments for those who do not sign and merely remain beneficiaries of the NPT, then the AP also fails. The challenge, then, is how to redesign the entire NPT-AP system such that it may be successful in preventing proliferation. I address these issues in the next section of this paper.

3. Restructuring the NPT-AP: A Normative Model

The current structure of the NPT and AP have failed to prevent some states from developing nuclear weapons, and has likely aided some states in acquiring the technologies necessary to build a nuclear weapon. The NPT Models demonstrate the inability of the NPT to separate proliferators from non-proliferators, and the AP models demonstrate the agreement's weak ability to prevent proliferation. By granting inspectors greater access to countries' nuclear facilities, the AP is supposed to grant the IAEA a stronger ability to detect a nuclear weapons program. Yet states that belong to the NPT are not required to sign the AP, and may therefore retain all the benefits of the NPT without facing any increased inspections (Hirsch 1995). Additionally, there is no deadline by which states must choose whether to sign the AP, and thus there is no way to know whether a state that is not currently a member of the AP has deliberately chosen not to sign, or has simply not signed yet due to simple inertia.

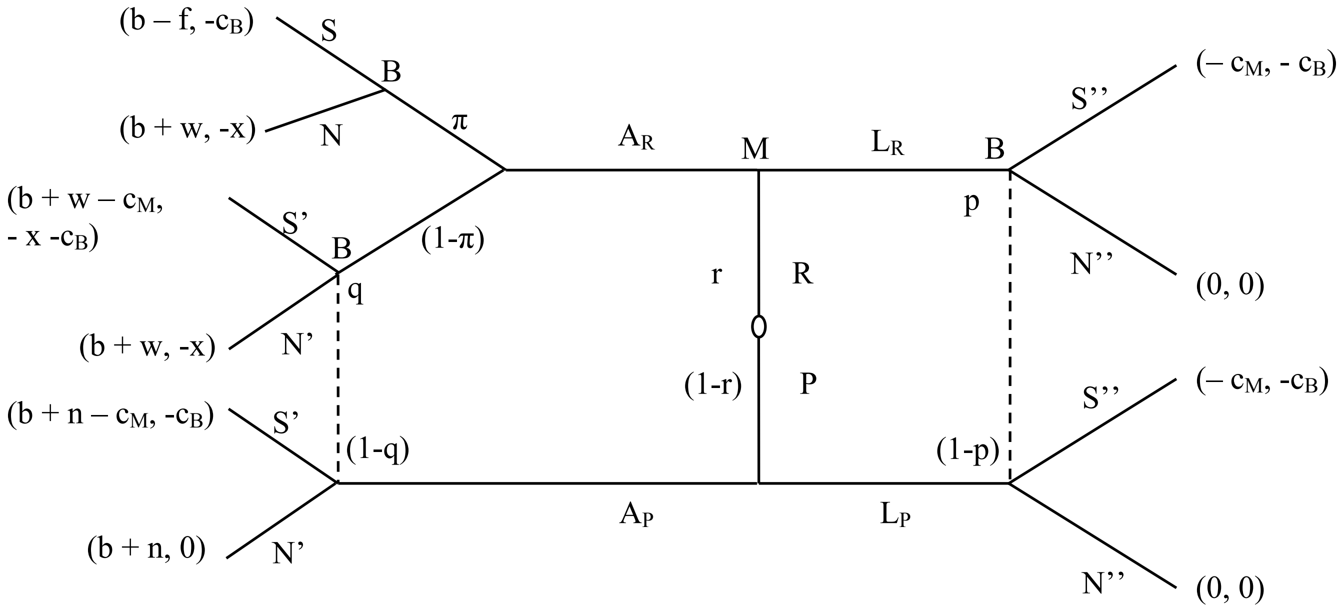
There is also a question about the specific goals of the NPT-AP structure. Should the international community hope to draw rogue proliferators into such an agreement, in order to inspect suspected nuclear weapons facilities? Or should the regime work to scare potential proliferators out of the NPT, thereby denying them the benefits of technology exchange? The challenge, then, is either to encourage rogue states to sign the AP, even at a high probability of successful detection, or to force rogue states to reveal themselves and sanction them accordingly.

AP Model B described in the previous section does have a limited ability to separate proliferators from non-proliferators, but at a very low level of successful detection of a state's rogue status. If the non-proliferation regime wishes to be taken seriously, its inspections must have a chance at succeeding in detecting rogue states, and in conducting effective inspections on states that may decide to develop nuclear weapons in the future. Additionally, the regime must not reward states that do not sign the AP with continued NPT status, but must instead force states to choose between signing the AP and expulsion from the NPT. Although the international community assumed that the construction of the NPT would grant the IAEA the ability to monitor states hoping to develop nuclear weapons and thereby prevent proliferation, such an assumption proved fundamentally flawed: the benefits associated with NPT membership actually *aided* those states hoping to develop nuclear weapons. If a state decides that it wants to develop nuclear weapons, why facilitate the process by providing such a state with technology and nuclear materials under the structure of the NPT? The following model will suggest an NPT-AP structure that will be more effective in curtailing nuclear proliferation.

Proposition F: Forcing states to choose whether to sign the AP or leave the NPT, and offering positive incentives to join the AP, can cause proliferators to separate from non-proliferators, or cause both types of state to sign the AP even with a significant probability of detecting a rogue state, thereby granting the IAEA a greater ability to prevent nuclear proliferation.

The following model presents a new version of the NPT-AP regime: one that will be better able to prevent nuclear proliferation. In the following game, the small state must choose whether to sign the AP (strategy S) or to leave the NPT structure (strategy L). Once again, signing the AP grants the big state a limited ability to detect a state's rogue status, and the probability of detecting a state as rogue is represented by (π). The left side of this game is the same as that of AP Model B, with a positive incentive (b) associated with signing the AP. The right side of the model is identical to the *left* side of the earliest NPT model, and represents the payoffs the players receive when the small state does not sign the AP and thus leaves the NPT. Note that when the small state chooses to leave the NPT, he no longer receives the benefits of

(w) and (n) associated with the transfer of peaceful nuclear energy technologies under the structure of the NPT. All payoff variables remain the same as those in previous models, and the game is structured as follows:



(Figure F.1: NPT-AP Normative Model)

NPT-AP Normative Model: Solutions

The model is solved under the assumption that $(w > n > c_M)$. For all possible strategies pursued by player M, player B's best response is to sanction those rogue players that it is able to detect and not to impose sanctions at any other juncture (S, N', N''). For a separating equilibrium to exist in which the rogue player M leaves the NPT and in which the peaceful player M signs the AP, then it must be true that the rogue player M receives a payoff from leaving the NPT that is greater than or equal to that which it would receive by signing the AP, and that the peaceful player receives a payoff from signing the AP that is greater than or equal to the payoff that it would receive by not singing the AP and thereby leaving the NPT, given player

B's strategy. Because player B is playing (N", N''), then the peaceful player M's best response is to sign the AP and receive the payoff of (b + n). The rogue player M must again consider the possibility that signing the AP will grant the big state the ability to detect its rogue status; but the stakes are much higher now because if he chooses not to sign the AP, then player M loses the benefits of nuclear technology transfer and the enhanced ability to develop a nuclear weapon (w). If the rogue player's status were not detected, given that player B only imposes sanctions on a detected rogue player, then the rogue player M would prefer to sign the AP and receive the payoff of (b + w) rather than (0). However, by signing the AP, the rogue player M does risk being detected and receiving a payoff of (b – f), in which case he would prefer not to sign the AP at all and simply receive the payoff of (0). The rogue player M therefore compares the expected value of his payoff from signing the AP with the payoff he would receive by not signing and thereby exiting the NPT. For the desired separating equilibrium to exist, the rogue player M will not sign the AP when $[(\pi)(b - f) + (1 - \pi)(b + w) \leq 0]$ and the rogue player M will sign the AP when the reverse is true: $[(\pi)(b - f) + (1 - \pi)(b + w) \geq 0]$. Solving for (π), with the knowledge that the peaceful player M will always sign the AP, given player B's strategy, the model yields the following equilibria:

$$\{(L_R, A_p), (S, N', N''), p = 1, q = 0, \pi \geq [(b+w)/(f+w)]\} \quad \text{(f.1)}$$

$$\{(A_R, A_p), (S, N', N''), 0 \leq p \leq 1, q = r, \pi \leq [(b+w)/(f+w)]\} \quad \text{(f.2)}$$

In the first equilibrium (f.1), the value of (π) is sufficiently high that the rogue player decides not to sign the AP and therefore leaves the NPT, yielding the desired separating equilibrium. By not signing the AP, the rogue player M gives up the benefits of the technology transfers stipulated by the NPT (w). In the second equilibrium, the value of (π) is sufficiently low that both types of player M will sign the AP. Note that because player B is pursuing a strategy of (S, N', N''), the relative values of (w), (n) and (c_M) do not have a bearing on the equilibria, and the same set of equilibria is obtained when the model is solved under the assumption that ($w > c_M > n$) or that ($c_M > w > n$).

The crucial aspect of these equilibria is the value of (π) under which the rogue player will sign the AP or decide to leave the AP. Note that the numerator of the fraction determining the value of (π) is now (b+w), whereas in the standard AP Model B in the previous section of this paper, the value of the numerator was only (b) (equilibria e.1 and e.2). Since the value of (b) is assumed to be quite small in comparison to the other variables, AP Model B indicated that the rogue player M would refuse to sign the AP (and retain the benefits of the NPT) at a very low threshold value of (π). If the inspections had even a small chance of succeeding, the rogue player would not sign the AP; since the international community presumably wants its inspections to be effective, this means that such an equilibrium is not necessarily an appropriate goal. Only if the success of inspections was perceived to be very low (equilibrium e.2) would the inspections actually be carried out on rogue states. And in AP Model B, a state that chooses not to sign the AP may remain a member of the NPT and continue to receive technical assistance from other countries.

Under this normative framework, however, the threshold value of (π) is much higher, because the numerator now includes the term (w). The value of the technical advantage conferred by the NPT (w) did not factor into the equilibria in the standard AP Models because states received those benefits regardless of whether they signed the AP. Under this normative model, however, the state loses this transfer of technology if it chooses not to sign the AP and leaves the NPT. There is now much more at stake in deciding whether to sign the AP and retain the benefits of the NPT, and thus the rogue state will sign the AP at a much higher probability of being detected as rogue—that is, at a higher threshold value of (π). The rogue states that choose to sign the AP submit themselves to inspections that have a chance of detecting their rogue status and thereby halting progress on their nuclear weapons plans.

Finally, in this model the separating equilibrium becomes very meaningful when states face a deadline by which they must decide whether to sign the AP or leave the NPT. Since there are only two equilibria in this game, one in which both types of player M sign the NPT and one in which only the peaceful player M signs the AP, then a state that chooses not to sign the AP could be assumed to be rogue. Under the incentive structure of this model, the peaceful player chooses to sign the AP: there is a positive incentive to do so, particularly since a failure to do so eliminates the gains of peaceful technology transfers. States that choose not to sign the AP and to leave the NPT could thus be assumed to be rogue states seeking to develop nuclear weapons, and could be inspected and sanctioned accordingly. On the other hand, if a player does choose to sign the AP, that does not necessarily mean that the state is peaceful; but, it does grant the IAEA the ability to inspect the state's facilities with greater efficiency and likelihood of success. This structure forces rogue states to choose whether to sign the AP and risk revealing belligerent intentions, or to leave the NPT and thereby signal to the world those belligerent intentions.

Rewriting the NPT: Summary

The creation of the Additional Protocol was an attempt to fix some of the problems of the NPT. By granting inspectors greater access to facilities in countries that signed the AP, its architects hoped to prevent another failure of detection like that which occurred in Iraq during the 1980s. Because signing the AP is not required of all members of the NPT, and because there is no deadline by which countries must sign, the current framework has only a limited ability to check nuclear weapons proliferation.

The Normative AP Model presented herein corrects some of these problems. By forcing states to choose whether to sign the AP or leave the NPT, and by enforcing a deadline by which they must make this decision, the "rewritten" version of the NPT-AP structure may be better able to prevent proliferation. Under this structure, rogue states are willing to sign the AP even at a higher probability of being detected as rogue; when they do sign, they grant the IAEA the ability to inspect a much wider range of their nuclear facilities. According to this model, a state that chooses not to sign the AP must be a rogue state. States that choose not to sign the AP by the appointed deadline could thus be sanctioned, subjected to forced inspections, etc. to halt or prevent any weapons proliferation. Once states have signed the AP, the IAEA should continue to work to improve its inspections capabilities in order to detect continuing (or beginning) proliferation. And for the "fine" (f) associated with being detected as a rogue state and sanctioned to be a credible threat, the UN should institute an automatic response structure, whereby states that are detected to be in violation of the NPT or that refuse inspection under the terms of the AP face immediate consequences for their actions.

4. Conclusions

Nuclear proliferation by small, “rogue” states is a very real and very immediate problem, and one with which the international community will continue to wrestle in the coming years. No model, no matter how sophisticated, can completely capture the nuances of reality; nor can a game theorist set up the world so that it will always play by the rules of the game. Nevertheless, the models in this paper demonstrate many of the shortcomings of the current nonproliferation regime and suggest some concrete changes to the current structure of the NPT-AP regime that could help stem the tide of nuclear proliferation.

The Treaty on the Non-Proliferation of Nuclear Weapons (NPT) went into effect in 1968, with the goal of preventing the spread of nuclear weapons. The treaty provides signatories with the benefits of nuclear energy technologies in exchange for renouncing the right to develop nuclear weaponry. This transfer of technology has, however, enabled states such as Iraq to develop a nuclear weapons program. Being a signatory to the treaty has actually been viewed as a way to deceive the international community about a state’s nefarious intentions (Kay 1995). As the NPT Models demonstrate, the implementation of general sanctions under the structure of the NPT cannot separate proliferators from non-proliferators; a program of targeted sanctions may be able to separate proliferators from non-proliferators, if the targeted sanctions render the cost of developing a nuclear weapon sufficiently high.

Attempts to rectify problems with the NPT have focused on improving the ability of inspections to detect illicit activity. The framers of the Additional Protocol hoped to increase the ability of IAEA inspectors to access sensitive sites within countries that sign. AP Model B shows that offering positive incentives to sign the NPT may cause states to separate themselves; but the rogue state is likely to sign the AP only at a very low probability of being detected as rogue—otherwise, he chooses not to sign the AP and instead remains within the regular framework of the NPT. This framework, however, fails to address the problem of states exploiting the NPT for the purposes of developing nuclear weapons.

The AP Normative Model provides a new framework for addressing nuclear proliferation. Rather than permitting states who do not submit to the additional inspections of the AP to retain the benefits of nuclear technology transfer, this normative model forces states to choose between signing the AP or withdrawing from the NPT. There are only two equilibria in this model: in the separating equilibrium, the rogue state chooses to leave the NPT, because the probability of being detected as a rogue player under the AP is too high; in the pooling equilibrium, both types of small state choose to sign the AP and thereby submit to increased inspections. States that choose not to sign the AP lose the benefits of technological transfer that have been exploited to develop nuclear weapons, and those rogue states that do choose to sign the AP must submit to increased inspections of a wider range of facilities and activities. This structure thus provides a better way of forcing rogue states to either identify themselves as rogue, by not signing the AP, or submit to increased inspections that have a greater chance of monitoring their status than those conducted under the current framework of the NPT.

This paper is not the first to point out the flaws of the NPT, nor the first to suggest an alternative mechanism for addressing proliferation. Although some have asserted that the current treaty system should be abandoned altogether, I have argued for a refinement of the current NPT-AP system that will be more effective in curtailing nuclear proliferation. This is not to suggest that the normative model that I present is the best way of preventing nuclear proliferation—but it is certainly a more effective way of structuring the current NPT to prevent the illicit acquisition of nuclear weapons.

The enforcement of such a structure would require the UN to implement an automatic system of punishments for those states discovered to be in violation of the terms of the NPT—otherwise, the threat of the payoff of (-f) associated with being discovered and sanctioned as a rogue state would not be credible. And the international community must be willing to withhold the benefits of nuclear technologies from those who will not agree to the full range of inspections mandated by the Additional Protocol. The IAEA should also focus on improving its inspections capabilities, to detect proliferation already in process or that may arise in the future. The international community must also work to rein in countries or individuals that provide nuclear assistance to countries that choose not to belong to the NPT, or that are suspected of developing nuclear weapons, since such assistance aided the nuclear aspirations of South Africa and Libya. The international nonproliferation framework cannot be a viable system for preventing proliferation if some actors are allowed to violate the rules of the game.

Any game theoretic modeling of a real world situation is bound to omit some nuances for the sake of simplicity. Nonetheless, the simplicity of the models in this paper helps identify the important features of the nonproliferation regime's incentive structure. Future modeling of the nonproliferation regime could employ other tools of game theory to examine different aspects of the nonproliferation regime—for example, how the actions of the international community affect the calculations of states deciding whether to pursue a nuclear weapons program. Additionally, further modeling could examine the international community's employment of tools other than sanctions: how the impact of positive incentives such as trade agreements or improved relations with the international community would affect states' decisions to acquire nuclear weapons. Indeed, the latter strategy seems to have encouraged Libya to renounce its weapons of mass destruction (Bahgat 2004). Or, game theorists could model the equilibria that occur when a small state does not have a utility function like those in this paper but rather has an "irrational" view of the payoffs associated with various actions.

The problem of nuclear proliferation is one that will continue to plague the international community, particularly as the world struggles to find alternatives to fossil fuels and nuclear energy presents a viable option. Ultimately, the debate over nuclear weapons is one that concerns the security of the peoples of the world. Efforts to address nuclear proliferation must remember that such strategies are not simply struggles of power or prestige, but rather efforts to protect the thousands or millions of individual lives that would be lost in any offensive (or accidental) use of a nuclear weapon. The tools of game theory can provide insights into how to better manage proliferation in the twenty-first century; but efforts to halt the spread of nuclear weapons must combine the tools of many disciplines, including science, game theory, psychology and diplomacy, and must stem from a firm commitment to making the world a safer place.

References

Bahgat, Gawdat. 2004. Oil, Terrorism and Weapons of Mass Destruction: The Libyan Diplomatic Coup. *The Journal of Social, Political and Economic Studies* 29:4 (December): 373-94.

Betts, Richard K. 1977. Paranoids, Pygmies, Pariahs & Nonproliferation. *Foreign Policy* 26 (Spring): 157-183.

Bidwai, Praful and Achin Vanaik. 2000. *New Nukes: India, Pakistan and Global Nuclear Disarmament.* NY: Olive Branch Press.

Bowen, Wyn Q. and Joanna Kidd. 2004. The Iranian Nuclear Challenge. *International Affairs* 80:2: 257-276.

Bracken, Paul. 2003. The Structure of the Second Nuclear Age. *Orvis* 47 (Summer): 399-413.

Brams, Steven J. 1985. *Superpower Games: Applying Game Theory to Superpower Conflict.* New Haven; London: Yale University Press.

Brams, Steven J. and D. Marc Kilgour. 1988. *Game Theory and National Security*. NY: Basil Blackwell.

Broad, William J. and David E. Sanger. As Crisis Brews, Iran Hits Bumps in Atomic Path. *New York Times* 5 March 2006.

Buzan, Barry. *An Introduction to Strategic Studies: Military Technology and International Relations*. NY: St. Martin's Press, 1987.

Carpenter, Ted Galen. 1994. Closing the Nuclear Umbrella (Life After Proliferation). *Foreign Affairs* 73:2 (March/April): 8-13.

Cohen, Avner. 1998. *Israel and the Bomb.* New York: Columbia University Press.

Cohen, Avner and Graham, Thomas Jr. 2004. An NPT for Non-Members. *Bulletin of the Atomic Scientists* 60:03 (May/June): 40-44.

Cole, Paul M. 1997. Atomic Bombast: Nuclear Weapon Decision-Making in Sweden, 1946-72. *The Washington Quarterly* 20:2: 233-51.

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

Doyle, James E. and Col. Peter Engstrom. 1998. The Utility of Nuclear Weapons: Tradeoffs and Opportunity Costs. In *Pulling Back from the Nuclear Brink: Reducing and Countering Nuclear Threats,* ed. Barry R. Schneider and William L. Dowdy, 39-59. London: Frank Cass.

Dutta, Prajit K. 1999. *Strategies and Games*. Cambridge, MA: MIT University Press.

Faiola, Anthony and Dafna Linzer. 2005. Japan Reports Missile Test by North Korea. *Washington Post.com*, May 2, A11. http://www.washingtonpost.com/wp-dyn/content/article/2005/05/01/AR2005050100204.html.

Fig, David. 1999. Sanctions and the Nuclear Industry. In *How Sanctions Work: Lessons from South Africa,* ed. Neta C. Crawford and Audie Klotz, 75-102. NY: St. Martin's Press.

Freedman, Lawrence. 1986. The First Two Generations of Nuclear Strategists. In *Makers of Modern Strategy from Machiavelli to the Nuclear Age,* ed. Peter Paret, 735-778. Princeton: Princeton University Press.

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

Garwin, Richard and Georges Charpak. 2001. *Megawatts and Megatons: A Turning Point in the Nuclear Age?* NY: Alfred A. Knopf.

Geller, Daniel S. 1990. Nuclear Weapons, Deterrence and Crisis Escalation. *The Journal of Conflict Resolution* 34 (2): 291-310.

Greenlaw, Steven A. 2006. *Doing Economics: A Guide to Understanding and Carrying Out Economic Research.* Boston: NY: Houghton Mifflin Company.

Hirsch, Theodore. 2004. The IAEA Additional Protocol: What It Is and Why It Matters. *The Nonproliferation Review* (Fall-Winter): 140-166.

Huth, Paul. 1990. The Extended Deterrent Value of Nuclear Weapons. *The Journal of Conflict Resolution* 34 (2): 270-290.

IAEA Board of Governors. *Implementation of the NPT Safeguards Agreement in the Islamic Republic of Iran: Draft Resolution Submitted by France, Germany and the United Kingdom.* IAEA: 4 February 2006.

International Institute for Strategic Studies. 2004. *North Korea's Weapons Programmes: a Net Assessment.* Houndsmills: Palgrave Macmillan.

Jasper, James M. 1990. *Nuclear Politics: Energy and the State in the United States, Sweden, and France.* Princeton: Princeton University Press.

Kay, David A. 1995. Denial and Deception Practices of WMD Proliferators: Iraq and Beyond. *The Washington Quarterly* 18:1 (Winter): 85-105.

Kibaroğlu, Mustafa. 2002. An Assessment of Iran's Nuclear Program. *The Review of International Affairs.* 1:3 (Spring): 33-48.

Krause, Keith. 1999. Rationality and Deterrence in Theory and Practice. In *Contemporary Strategy and Security,* ed. Craig A. Snyder. NY: Routledge.

Kreps, David M. 1990. *A Course in Microeconomic Theory.* Princeton: Princeton University Press.

Lewis, Patricia. 2004. The New Urgency of Effective Arms Control Cooperation. *Foresight.* 6(4): 246-50.

Manning, Robert A. 1997-98. The Nuclear Age: The Next Chapter. *Foreign Policy* 109 (Winter): 70-84.

N. Korea Agrees to Give Up Nuclear Program. *CNN.com,* Sep 19 2005. http://www.cnn.com/2005/WORLD/asiapcf/09/19/korea.north.talks/index.html.

N. Korea May Have 5 or 6 Nukes. *CNN.com,* May 9 2005. http://www.cnn.com/2005/WORLD/asiapcf/05/09/north.korea/index.html.

Nye, Joseph S. Jr. 1985. NPT: The Logic of Inequality. *Foreign Policy* 59 (Summer): 123-131.

Nuclear Strategy Study Group. 1994. The Future of Arms Control. In *Toward a Nuclear Peace: The Future of Nuclear Weapons*, ed. Michael J. Mazarr and Alexander T. Lennon. NY: St. Martin's Press.

Ollapally, Deepa and Raja Ramanna. 1995. U.S.-India Tensions: Misperceptions on Nuclear Proliferation. *Foreign Affairs* 74:1 (Jan/Feb): 13-18.

Ozga, Deborah. 2000. Back to Basics on the NPT Review Process. *Security Dialogue* 31:1 (March): 41-54.

Payne, Keith. 2003. The Fallacies of Cold War Deterrence. *Comparative Strategy* 22 (5): 411-428.

Pfundstein, Dianne. 2005. Disarming Rogue States: A Game Theoretic Perspective. Paper submitted for Professor Ashok Rai's ECON 385 class at Williams College.

Pierre, Andrew J. with Claudia W. Moyne. 1976. *Nuclear Proliferation: A Strategy for Control.* NY: Foreign Policy Association.

Quester, George H. 1977. Reducing the Incentives to Proliferation. *Annals of the American Academy of Political and Social Science* 430 (March): 70-81.

Schell, Jonathan. 2000. The Folly of Arms Control. *Foreign Affairs* 79 (5): 22.

Schelling, Thomas C. and Morton H. Halperin. 1961. *Strategy and Arms Control*. NY: Twentieth Century Fund.

Sciolino, Elaine. 2006. Nuclear Panel Votes to Report Tehran to U.N. *New York Times* 5 Feb 2006.

Simons, Geoff. 2003. *Libya and the West: From Independence to Lockerbie.* Oxford: Centre for Libyan Studies.

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

Snyder, Craig A. Contemporary Security and Strategy. In *Contemporary Security and Strategy,* ed. Craig A. Snyder. NY: Routledge.

Snyder, Jack. 2003. Imperial Temptations. *The National Interest* (Spring 2003): 29-40.

Spector, Leonard S. 1987. Nuclear Proliferation: Who's Next? *Bulletin of the Atomic Scientists* 43:4 (May): 17-20.

Thomson, William. 2001. *A Guide for the Young Economist*. Cambridge, MA: MIT Press.

United Nations. 2005. Background Information: 2005 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons (NPT). http://www.un.org/events/npt2005/background.html.

United Nations. 2000. The Treaty on the Non-Proliferation of Nuclear Weapons. http://www.un.org/events/npt2005/npttreaty.html.

Walker, Ronald. 2000. What is to be Done About Nuclear Weapons? A Rejoinder. *Security Dialogue* 31:2 (June): 179-84.

Walters, Ronald W. 1987. *South Africa and the Bomb*. Lexington, MA: Lexington Books.

Waltz, Kenneth N. 1981. The Spread of Nuclear Weapons: More May Be Better. From *Adelphi Papers*, 17. Oxford: Oxford University Press.

Wohlstetter, Albert. 1976-77. Spreading the Bomb without Quite Breaking the Rules. *Foreign Policy* 25 (Winter): 88-96+145-179.

Appendix A:

Factors Driving Nuclear Proliferation and the Efficacy of Nuclear Weapons in Providing a State with Greater Security (Nuclear Deterrence Theory)

For many states, pursuing nuclear weapons is “central to the state-making project” (Bracken 2003, 405). Particularly for states that have solidified within the last half century, the pursuit of nuclear weapons technology may represent a state’s independence and an ability to be a major player on the international stage. For such a state, nuclear weapons serve as “symbols of power, legitimacy, and status” (Bracken 2003, 405), signaling to the international community that a state deserves to be taken seriously in the world arena. Thus a small state may perceive the acquisition of nuclear weapons as the path to greater power and prestige in the international system (Pierre and Moyne 1976, 10). Efforts to disarm a nation for which a nuclear weapons program serves as a symbol of its prestige and power will have to take into consideration the fact that these disarmament efforts will likely be seen as assaults on the nation’s status and independence. The international system may need to take measures to compensate a state for the status that it would believe itself to be giving up when surrendering its nuclear weapons program.

In addition to a desire to gain prestige internationally by acquiring nuclear weapons, domestic politics may drive some states to pursue nuclear weapons. For a country experiencing serious domestic social or economic problems, a nuclear weapons program may be a means of distracting attention from such difficulties and focusing national pride and prestige on the government. A program that is technologically demanding and that grants a country attention in the international system may be a powerful way of promoting nationalism (Pierre and Moyne 1976, 11). For a state largely motivated to proliferate due to internal political, social and economic unrest, the ability of an outside agent to convince the state to give up its weapons program may be limited, unless the outside actor is willing to help to alleviate the nation’s internal grievances. This supposes, of course, that the regime in power in the proliferating state does not have an interest in maintaining such difficulties as a means of ensuring its own position.

Another motivation and perhaps the most intuitive reason for which a state chooses to arm itself with nuclear weapons is national security (Pierre and Moyne 1976, 9); however, for a variety of reasons, including both the moral imperative against the use of nuclear weapons and the ability of many countries to retaliate after a nuclear strike, nuclear weapons are increasingly viewed as "inappropriate instruments for achieving tangible foreign and military policy objectives" (Doyle and Engstrom 1998, 41). The only way in which nuclear weapons may increase a state's security and power is by deterring the use of nuclear weapons by other states (Doyle and Engstrom 1998, 41). The ability to deter nuclear weapons use by other states is particularly crucial for a state facing a threat from a regional neighbor. A state that perceives that a neighbor possesses or may be acquiring nuclear weapons may feel a strong pressure to acquire nuclear weapons of its own in order to protect itself from such a regional rival (Pierre and Moyne 1976). The fear of rivals is acute for those states that do not enjoy a guarantee of security from the world superpower, and thus states that feel marginalized by the international system may be even more inclined to acquire nuclear weapons when there is no guarantee of protection from attackers (Pierre and Moyne 1976). Disarming a state that seeks nuclear weapons to protect itself from a perceived threat, regional or otherwise, is likely to be difficult if such a state's security concerns are not met.

The extent to which the possession of nuclear weapons would deter their use by a great power possessing an extensive nuclear arsenal is unclear; whether nuclear weapons may serve as "equalizers" (Pierre and Moyne 1976, 10), allowing states to prevent an attack by a great power, is heavily debated by scholars. As mentioned in the body of this paper, traditional theories of deterrence assert that two states possessing large and relatively equal nuclear arsenals do not launch their weapons against each other because each state knows that doing so would only invite its own destruction. Although Mutually Assured Destruction (MAD) does not apply to cases in which two opponents are unevenly matched in weapons capabilities, deterrence theory still provides a useful lens for the consideration of current proliferators and the likelihood that newly-acquired weapons would ever be deployed.

Since the stability of deterrence depended on the possession of enough nuclear weapons that a state could survive a strike designed to eliminate its weapons, such an equilibrium would only be secure when both states possessed large amounts of nuclear weapons. This seems to be a perverse conclusion: deterrence theory suggests that, as the United States and Soviet Union acquired *more* weapons, the likelihood that they would be used in any conflict was actually lessened. By this calculation, "as nuclear weapons became more numerous and more powerful, they also became less usable; but as nuclear weapons became less usable, one needed more of them to deter others who possessed them" (Gaddis 1997, 101). Thus proliferation by the United States and Soviet Union may have actually helped make the world more secure during the Cold War—indeed, Gaddis asserts that the existence of nuclear weapons forced conflicts such as the Korean War to remain fairly localized, rather than erupting into global warfare (Gaddis 1997, 104). Deterrence logic would therefore suggest that the absence of great-power war during the latter half of the twentieth century was in part due to the limitless destruction that both the Soviet Union and United States could have unleashed on one another with their nuclear weapons.

Does the idea of enhanced world security associated with increasing weapons stocks extend to proliferation by small, weak states? Kenneth Waltz suggests that the spread of nuclear weapons may actually make the world system more stable and less prone to conflict escalation (Waltz 1981). Waltz claims that states will be less likely to be unpredictable or take unnecessary risks because of the possible consequences if the use of force is not controlled. He asserts that this observation extends to small states, claiming that, "nuclear weapons induce caution, especially in weak states" (Waltz 1981, 457). Furthermore, a great power could not be entirely sure of eliminating a small state's nuclear arsenal in a preemptive strike and would thereby risk being hit with a nuclear weapon not eliminated in that strike; thus large states will be induced to act more cautiously towards a nuclear-armed state, precisely because of the inability to be absolutely certain that one could eliminate even a small nuclear arsenal (Waltz 1981, 457-8). For Waltz, then, nuclear weapons encourage rational actors to behave more cautiously in the international arena, thereby suggesting that the spread of nuclear weapons would have a stabilizing effect on the international system.

In analyses of the outcomes of standoffs between states in which MAD is not in place, Paul Huth finds that, "possession of a nuclear retaliatory capability has contributed to extended deterrence success" (Huth 1990, 286). His work draws on his prior probit analysis of 56 cases of military deterrence that occurred between 1885 and 1983; in this paper, he narrows his focus to the 15 cases that involved at least one nuclear power and added interactive variables to the probit analysis. By adding these variables, he hoped to examine the interactions among nuclear retaliatory capability, the balance of conventional forces and tit-for-tat military escalation in determining whether deterrence succeeded. From his analysis, he concludes that the "possession of a nuclear retaliatory capability did enhance the prospects of extended deterrence success," but noted that this nuclear capability did not eliminate the importance of conventional military forces in determining the success of deterrence (282). This conclusion reinforces Waltz's theory that the chance that a state's first strike would fail to eliminate a weapon held by an opponent would induce such a state to be cautious in a nuclear conflict. It follows that the ability to stand up to an opponent in the international system would be a positive incentive encouraging proliferation.

Some theorists believe, however, that deterrence breaks down, even in the case of evenly-matched adversaries. Keith Payne asserts that challengers facing a superior rival are not always rational in their actions, even though their actions may be *reasonable* given their own utility functions. Because opposing states may have different norms or standards of behavior, because leaders may possess radical personal beliefs, or because leaders may view inaction as more costly than running a risk in a game of nuclear standoff, a country such as the United States cannot assume that its opponent will have a rational utility function in the face of a nuclear threat—or even in the face of a conventional military assault. This problem is particularly exacerbated because the United States is highly averse to both collateral damage and risk taking in its calculations about nuclear deployment (Payne 2003).

An analysis by Daniel Geller arrives at a conclusion about the deterrent ability of nuclear weapons that differs from that reached by Huth. Geller examines the pattern of escalation of conflicts between states, in cases both when states are evenly matched in nuclear capability and when they are not. He focuses on the fact that, "the decision maker's dilemma is to construct a strategy to secure political interests through coercive actions that raise the possibility of war without pushing the risk to an extreme level" (294). He employs a data set of 393 militarized conflicts between 1946 and 1976, 111 of which involved at least one nuclear weapon state (classified according to whether a state had detonated a nuclear weapon and thus including India by 1974). The conflicts were classified according to the highest level of force employed, ranging from no action or threat to war. The author constructed and analyzed a contingency table that examined the threat classification of a conflict, according to the nuclear weapons status of the instigator and the target state. Geller finds that, "the possession of nuclear weapons appears to have no deterrent effect in disputes with nonnuclear states" (302). By conducting a Markov analysis in which the probability of a response from a target was calculated based on the hostility level of the initiator, Geller finds that nonnuclear targets and initiators behave more aggressively in conflicts with nuclear states than do nuclear states. Geller's analysis suggests that a weak proliferator or a state that is in the process of acquiring a nuclear weapon is not deterred from aggressive behavior by an opponent's possession of nuclear weapons.

On the other hand, Huth's analysis suggests that nuclear weapons do have a strong deterrent effect. How may these results be interpreted? I posit that a smaller state that is undeterred by its opponent's overwhelming nuclear capability may feel it has less to lose in the international system, or may simply be more willing to take risks, than its stronger opponent. If an American monopoly on nuclear weapons at the end of WWII encouraged Stalin to take risks (Gaddis 1997, 92), then perhaps the existence of states with superior nuclear capabilities pressures other states to take risks as well—even to go against the international community to build nuclear weapons. In Appendix B, I provide a series of simple models illustrating deterrence during both the Cold War and the twenty-first century. These models show that the launch of a nuclear weapon is still unlikely in the absence of MAD, but still possible if one of the players has an "irrational" utility function that causes it to misjudge the costs of its actions.

Although it may provide some useful insights into the behavior of states possessing nuclear weapons, deterrence theory does possess a fundamental flaw: it cannot guide or predict a state's actions if deterrence fails (Freeman 1986, 758). The game theorist would recognize that, under the conditions of MAD, the threat to retaliate against a nuclear strike is not credible because it will only inflict further disutility on one's self—unless, of course, the victim would receive a high utility from knowing that it has retaliated against its foe. Since nuclear weapons have not been used against an opponent since 1945, we cannot observe whether a state that has been attacked would choose to retaliate with its own nuclear weapons. On the other hand, it seems intuitive to expect that a state possessing overwhelming nuclear superiority may have little to fear in retaliating against a nuclear attacker with only a small store of weapons, most of which it would probably employ in its initial attack. I model the conditions of deterrence in both the Cold War and the twenty-first century in Appendix B.

Appendix B:

Nuclear Deterrence Models

In the following pages, I develop a series of simple models to illustrate deterrence theory and to demonstrate why deterrence may break down, and therefore weapons be launched, in a conflict between two nuclear rivals that are unmatched in weapons capability. According to deterrence theory, the ability of both the United States and Soviet Union to launch an overwhelming, retaliatory attack after sustaining a strike from its opponent prevented both nations from launching their nuclear weapons. The question I seek to answer is: under what conditions may nuclear weapons be launched in a conflict between two states with vastly different nuclear weapons capabilities? I then consider the conditions under which a state would launch a preemptive, nuclear strike in order to disarm an inferior nuclear power.

Although the bulk of this paper addresses the nonproliferation regime and the role of beliefs and signaling in nuclear proliferation, it is also important to examine whether the possession of nuclear weapons by small "rogue" actors is really a threat—that is, whether nuclear weapons would ever be used by such states, or used by opponents of such states. Even though nuclear weapons have not been launched offensively since their debut in 1945, the international system is currently very worried about the acquisition of nuclear weapons by small states or by terrorists—presumably because of a fear that such weapons might actually be launched. The following models are therefore designed to illustrate the conditions under which a nuclear weapons launch is possible. I accomplish this by constructing a series of models that illustrate conditions under which deterrence may break down and when nuclear weapons may be launched against an opponent.

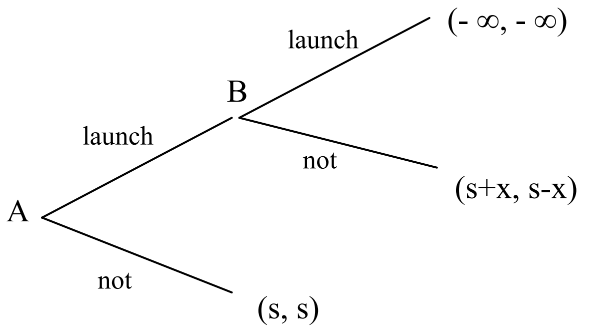
The earliest game theoretic models of deterrence were simultaneous-move games, in which evenly matched nuclear adversaries decided whether or not to launch their weapons against each other. The simplest of these games were prisoner's dilemma models, in which the equilibrium reached by the two opponents was a conflict or arms race that could have been avoided had the other side been able to trust its opponent (as reviewed by Krause 1999). Later models attempted to model the "brinkmanship" that occurs in an actual crisis, when neither rival wishes to be the first to back down in a conflict over nuclear weapons use. Such games were structured as games of "Chicken," in which it was to each player's advantage to convince the other side that he was irrational enough to fail to back down from a crisis situation, even at great personal cost to himself, in order to compel his opponent to back down instead (Krause 1999, 131).

In his series of games modeling superpower conflict, Steven Brams takes the traditional models of "Chicken" one step further by making them into sequential games to better reflect the decision-making process that actually occurs in nuclear weapons standoffs (Brams 1985, 19). He builds on the earlier, non-sequential game of Chicken in which the two players' payoffs are ranked according to their preference of outcomes. His game is solved by backwards induction, based on each player's belief that its opponent would choose to cooperate or not. To apply his deterrence model to a real-life crisis, Brams models the Cuban Missile Crisis of 1962 as a sequential game solved by backwards induction (Brams 1985, 60). Although Brams' models are based on the Cold War era, I use his concept of sequential games of deterrence as the starting point for the following models but without the strict ranking of payoffs employed by Brams.

Cold War Launch Model

The following model illustrates the principles underlying traditional, Cold War era deterrence theory. The Cold War was characterized by two nuclear rivals of relatively equal strength and from which traditional thinking about nuclear theory was derived. This game has two players, A and B, representing two countries that possess the same number of nuclear weapons. In this game, both players possess sufficient nuclear weapons capabilities to completely destroy their opponent's population, infrastructure, industrial capacity, etc. In the first iteration, player A must decide whether to launch its weapons or not; if it does not launch its weapons, the game is over, but if he does launch, then player B must decide whether or not to launch its own weapons in retaliation.

The game is structured as follows, with payoffs (A, B):



(Figure W.1: Cold War Launch Model)

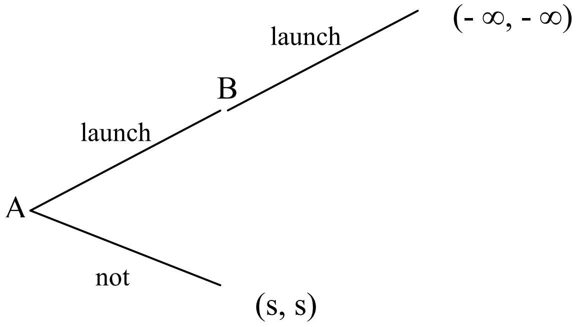
A payoff of (- ∞) for both players when both A and B decide to launch reflects the fact that a retaliation by player B would prompt player A to retaliate, and so forth, and thus the conflict would degenerate into total nuclear war, resulting in the complete destruction of both players. Consequently, the payoff of (s) that each side earns when player A decides not to launch reflects the utility each state receives from the status quo or "safety". Finally, (x) refers to the utility associated with the delivery of (x) weapons: player A adds (x) units of utility to its status quo level with the delivery of (x) weapons to state B, while player B loses (x) units of utility for the delivery of (x) weapons to state B.

By backwards induction, player B should choose not to launch its weapons in response to an attack by player A. Therefore player A, believing that player B would not retaliate, could increase its own utility by launching weapons against its opponent. Thus in equilibrium, player

A launches its weapons and B does not; A receives a payoff of (s + x) and B receives (s – x). According to deterrence theory, the ability to inflict an overwhelming retaliatory strike against one's opponent should deter the opponent from launching weapons. Yet in this case, player A is not deterred by player B's ability to inflict a punishing response because player B would lose so much utility by launching a retaliatory attack. This equilibrium illustrates one of three main challenges for deterrence theorists: the challenge of how to make one's threat of retaliation *credible* (Krause 1999, 123), so that it actually deters the opponent from launching its weapons.

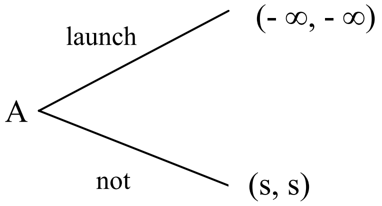
Cold War Deterrence Model

I now model the situation under which the threat of retaliation was successful in deterring the use of nuclear weapons during the Cold War. Attempts to make such threats credible included placing weapons on hair triggers or computerized, automatic response systems, such that a state had no ability to choose whether to launch a retaliatory attack. Such actions eliminated the ability to choose *not* to launch in the event of an attack. The Cold War game is now structured as follows:



(Figure W.2: Cold War Deterrence Model.0)

Player A may choose from the same two options, but now player A knows that, if he chooses to launch its weapons, it will receive a payoff of (- ∞). Player B has essentially removed the ability to choose whether or not to launch its weapons, and thus the game collapses into:



(Figure W.3: Cold War Deterrence Model)

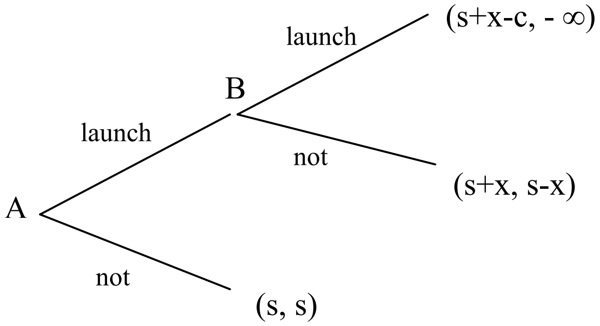
It is now clear that, faced with a payoff of (-∞) and total destruction when he launches his weapons, player A now chooses not to launch his weapons, and deterrence is achieved.

Symmetry in Cold War Deterrence

When both players possess equal weapons capabilities and have identical preferences, player B faces exactly the same decision tree as player A. Identical Cold War Games could be drawn with B making the initial decision of whether or not to launch its weapons and A responding to its launch; the difference in such a case would be that B would acquire x units of utility for launching its weapons when A does not launch its own, while A would lose the x units of utility from being hit by B's weapons without retaliating. Essentially, both players simultaneously face the decision of whether to launch their weapons in a first strike. Thus, as in the Cold War Launch Game in which player A is not deterred from launching his weapons, in an identical situation, player B would not be deterred from launching his weapons. But when A places its weapons on automatic trigger systems, such that it will definitely launch a retaliatory attack in the event that it is attacked by player B, then player B will be deterred from launching its weapons. Thus eliminating the ability to choose whether or not to launch a retaliatory strike actually deters the launch of an initial strike, by both players A and B, when both players possess equal nuclear weapons capabilities.

Tiger in the Grass Model

I now leave the Cold War era and model a nuclear conflict between two states that are not equally matched in capability. In the following game, player A possesses a large stockpile of nuclear weapons, capable of completely destroying its opponent. Player B possesses only a few nuclear weapons, similar to North Korea, which may have six to twelve nuclear warheads. Player A will launch a limited number of weapons in any initial strike, but would retaliate to any strike by player B by completely destroying player B; I assume that player B uses all six of its weapons in any strike. Finally, I assume that neither country's weapons are on an "automatic launch" setting, as in the Cold War Deterrence Games:



(Figure W.4: Tiger in the Grass Game)

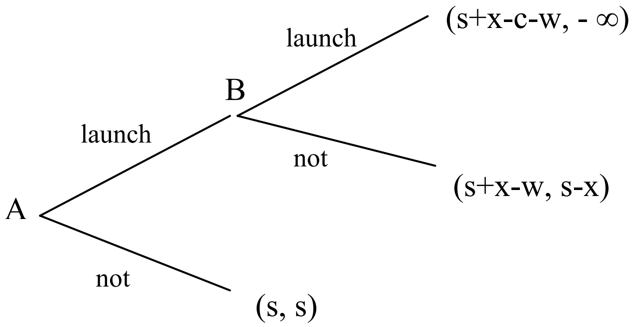
As in the Cold War Games, players A and B both receive a payoff of (s) for the maintenance of the status quo. In this case, however, only player B risks receiving a payoff of (- ∞), because only player A possesses the ability to deliver an apocalyptic nuclear attack against its opponent. Player A again receives (x) units of utility from launching (x) weapons, while losing (c) units of utility for a launch of c weapons by player B and (x > c). Similarly, player B loses (x) units of utility when player A launches (x) weapons and gains (c) units of utility for launching its own weapons against player A. Technically, player B would receive a payoff of (- ∞ + c) when it retaliates in the face of player A’s attack, since it would gain (c) units of utility before being destroyed by player A’s retaliatory strike; for ease of notation, I have simplified this payoff to ($-\infty$).

Solving this game by backwards induction, it is clear that player B would choose not to launch its weapons after absorbing an attack by player A, since launching a retaliatory attack would spur player A to hit back with the full strength of its nuclear arsenal and destroy player B completely. Knowing this, player A receives a higher utility from launching than from not launching, and will choose to launch its weapons. Even if player B was able to place its few weapons on an automatic retaliation trigger, as long as ($x > c$), then player A would still choose to launch an attack against player B.

The crucial difference from the Cold War Games is the fact that player A no longer faces complete destruction when player B retaliates to an attack by player A. When an overwhelmingly powerful player A considers whether to launch its nuclear weapons at an adversary with vastly inferior nuclear capabilities, deterrence fails and player A maximizes his utility by launching his nuclear weapons against player B.

Tiger in the Grass Model: Extension (the Squeamish Tiger in the Grass)

Is it entirely realistic to expect that a strong state such as the United States would launch a nuclear attack against a small state such as North Korea? What if player A incurred a high loss of utility not only from being hit by nuclear weapons, but from launching its own weapons? The following model captures the situation in which player A is reluctant to launch its own weapons:



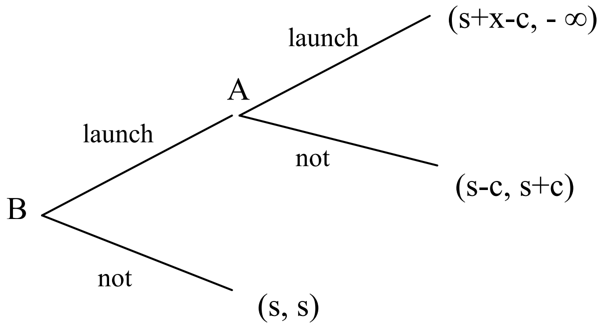
(Figure W.5: Squeamish Tiger in the Grass Game)

In this game, the new term (w) captures the loss in utility that A incurs from launching its own weapons. Solving for equilibrium, player B would still refrain from launching its own weapons if attacked by player A. Player A's decision to launch now depends on the relationship between variables (w) and (x). If ($w \leq x$), then player A will choose to launch, as in the original Tiger Game. However, if ($w > x$), then player A will choose not to launch its weapons. Thus if player A experiences a net loss in utility from launching its weapons, even against an opponent that will not retaliate, then player A will not launch its nuclear weapons.

Is this a reasonable payoff structure for a large state with an overwhelming nuclear weapons arsenal? If a state believes that it will suffer a large loss of prestige or power in the international system following a launch of its weapons—particularly against a state that will not strike back—then a state may receive a net loss in utility for launching its nuclear weapons. The fact that nuclear weapons have not been launched against an opponent since 1945 has likely stigmatized the use of nuclear weapons. A state may also suffer a loss in prestige or credibility among its own citizens for launching nuclear weapons. I suggest that a loss of utility stemming from a drop in prestige among its own citizens is much more likely to be a characteristic of democratic regimes than dictatorial regimes, since democratically elected governments must be much more aware of the sentiment of public opinion than dictatorial governments, if they hope to remain in power.

Scorpion in the Sand Model

What about the other side of this story? How does a nuclear-inferior player B decide whether to launch an attack against its superior foe? In the next model, player B faces the initial decision of whether or not to launch its small handful of weapons against its opponent. The parameters remain the same as those defined in the original Tiger in the Grass Model:

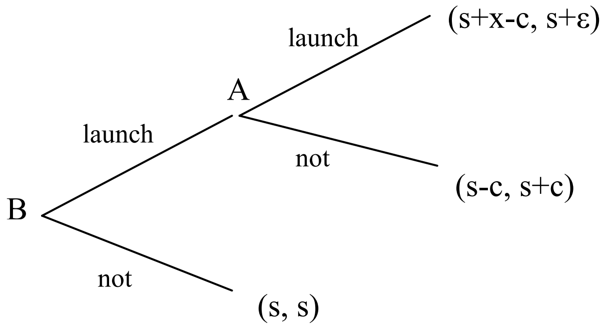


(Figure W.6: Scorpion in the Sand Game)

Even though B moves first, if A retaliates, it would do so with its full strength and thereby wipe out player B. Thus, player B still incurs a payoff of (- ∞) when both players launch. Indeed, by backwards induction, player A would choose to launch its weapons if hit by an attack by player B. Because of this, player B would choose not to launch its weapons in a first strike. While the player A in the original Tiger Game was not deterred from launching its nuclear weapons by player B's nuclear weapons capability, the weaker player B is deterred from launching its weapons because of player A's ability to inflict overwhelming destruction on player B.

Reckless Scorpion in the Sand Model

What if player B were not "rational"?[10] In the following model, I will examine the expected equilibrium when player B still moves first, but is now "irrational" because he does not value his own destruction at (-∞). Note that player B's behavior in the following game is perfectly rational given the payoff structure described below; he is "irrational" or reckless in the sense that he does not value the destruction of its own country at (-∞). I now suppose that player B actually gains a small amount of utility from inflicting damage on player A:



(Figure W.7: Reckless Scorpion in the Sand Game)

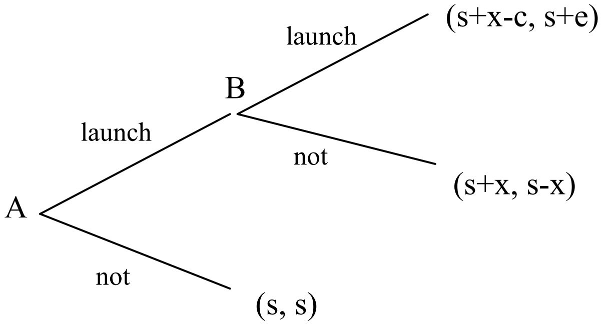
In this game, the new payoff (ε), epsilon, represents a small net gain in utility from launching weapons against player A, even in the face of player A's retaliation and subsequent destruction of player B. Solving this model by backwards induction, player A again chooses to launch its weapons in response to an attack by player B. Now, instead of being deterred by player A's retaliatory capability, player B would choose to launch its weapons, because it values the launch against player A more highly than the status quo.

[10] The following models incorporate theories about modeling irrationality from Kreps, David M., *A Course in Microeconomic Theory* (Princeton: Princeton University Press, 1990). According to Kreps, a player may be "irrational" if his behavior or utility function seems to violate the player's "best interests" (Kreps 1990, 480). The payoffs in the game are structured to reflect this irrationality, such that the player still acts "rationally" given the modified payoff structure—i.e. the player plays the game rationally, based on an "irrational" utility function.

Is it realistic to model a state that does not receive a payoff of (-∞) when completely destroyed by a nuclear attack? While such a utility function may seem unrealistic at first glance, I suggest that there are plausible explanations for such a utility function: when the decisions about player B's actions are made by a group or even a single individual that places a high premium on inflicting damage on player A, and may have the ability to exit the country before an attack inflicts any damage on themselves; when the regime in charge of making decisions for player B faces an imminent coup from within its own country and feels it has little to lose from attacking a rival, since it is likely to be violently overthrown in the near future; or when the leaders of state B adhere to an extreme ideology that privileges attacking an enemy over their own deaths. This model suggests that, if the actions of player B are being determined by an individual or a group of individuals that would bear little personal cost from a retaliatory attack by player A, then player B will be more willing to launch its weapons against player A, even though such an act may virtually guarantee player B's own destruction. This is a troubling conclusion to make about the ability of a small group of actors whose decisions may determine the fates of millions of people.

Reckless Scorpion in the Sand Model: Extension (The Reckless Response Model)

In the preceding model, I evaluated the sustainability of deterrence when a reckless player B made the first decision about whether or not to launch its weapons. In the following model, I analyze whether a reckless player B would retaliate against an attack by player A, since I demonstrated by backwards induction in the Tiger in the Grass Game that player B would choose not to retaliate against player A. The following model retains the same payoffs as those outlined in the Reckless Scorpion in the Sand Game, and simply changes the order of the players' actions such that player A once again makes the initial decision of whether or not to launch its weapons:



(Figure W.8: Reckless Response Model)

In this case, the reckless player B again receives a marginal increase in its utility when both players choose to launch their weapons. Because of this, by backwards induction, player B will choose to launch its weapons. Knowing this, and assuming that (x > c), then player A will also choose to launch its weapons. With the same payoff structure as the preceding model, a reckless player B will choose to launch a retaliatory strike against player A.

Are there situations in which player B may receive less disutility from being destroyed by player A than from simply being attacked? If the individual(s) responsible for deciding player B's moves and for setting its payoffs receive higher disutility from failing to respond to an attack by player A, for reasons of national pride, because of a need to be viewed as tough by a rival, or because such leaders believe they would bear little cost personally in the event of a strike, then such a utility function may be plausible. I suggest that, as in the Reckless Scorpion in the Sand Game, a reckless player B is more likely to be a politically closed society. On the other hand, a democratic player may be much more risk-averse to a nuclear weapons strike. In the above model, if the individuals responsible for making player A's decisions were elected democratically, their payoff structure may be such that (c) greatly exceeds (x). For such a situation, by backwards induction player A does not launch and player B does launch. Thus, by appearing "reckless" and making it clear that it would retaliate against a strike by player A, player B would actually deter a risk-averse player A from launching its weapons.

Summary: Deterrence in Sequential Games of Unequal Nuclear Opponents

The models that I have constructed to illustrate nuclear standoffs between two unevenly matched powers reveal many different conclusions about the ability to deter a launch of nuclear weapons in standoffs between asymmetric powers. In these sequential models, the strong state represented by player A is not deterred from launching its weapons against an opponent whose weapons are not on an automatic response system and whose utility-maximizing move is to absorb an attack by player A without retaliating (Tiger in the Grass Game). In the case of a player A that incurs a net loss in utility from launching its own weapons (Squeamish Tiger Game), player A will not launch a first attack against player B. It may still be possible for a Squeamish Tiger to launch its weapons in a first strike if the loss in prestige and influence it suffers from launching its own nuclear weapons is less than the gain in utility it receives from punishing its opponent.

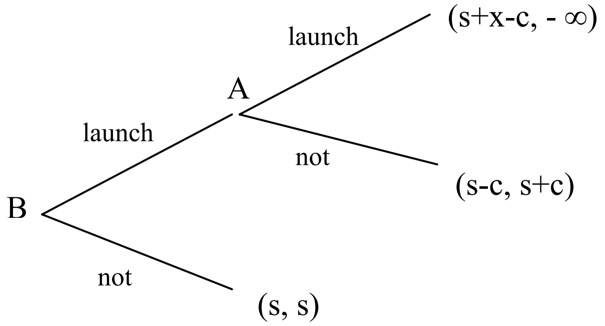
Turning to the opposite situation, in which the weak player B considers whether to launch its own nuclear weapons against player A in a first strike, player B is deterred from launching its weapons by the fact that player A will retaliate with its full strength and completely destroy player B's state (Scorpion in the Sand Game). However, when I evaluate a player B that is "irrational" or reckless—one that does not incur a disutility of $(- \infty)$ when player A retaliates to player B's launch—I demonstrate that player B would launch its weapons, even when facing elimination by a nuclear-superior rival (Reckless Scorpion in the Sand Game). Reversing the order of the players' decision-making, I demonstrate that the reckless player B would also launch a retaliatory attack against player A (Reckless Response Game). This "reckless retaliator" has the same payoff structure as the "reckless first striker," and will launch its nuclear weapons after an attack by player A, despite the fact that doing so would guarantee its own destruction. If player A is highly averse to being hit by nuclear weapons, such that (c > x), then the reckless player B, by convincing player A that it would retaliate, may actually deter player A from launching a first strike.

I have therefore demonstrated that there are conditions under which nuclear weapons may in fact be utilized by two opponents of differing nuclear weapons capabilities. In general, deterrence is most likely to fail when a player places a low value on the disutility that it receives from an attack or when it places a high value on attacking its opponent; deterrence is more likely to succeed and prevent weapons launch when a player receives a very high disutility from even a "small" nuclear weapons attack or receives little utility from launching its own weapons.

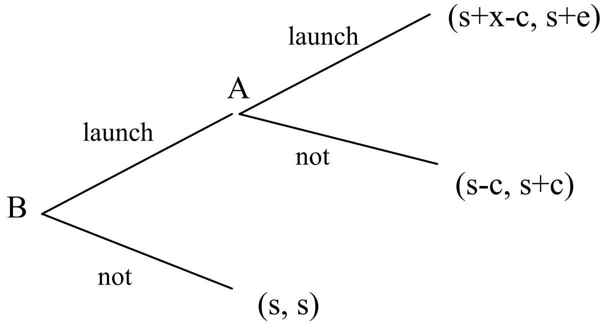
The Threat of Recklessness: The Preemptive Tiger Model

In the following model, I illustrate the situation in which player A considers whether to launch a preemptive strike against player B to wipe out player B's small nuclear arsenal. Player A considers this decision because player A is unsure of whether or not player B is "rational"—i.e. player A does not know whether player B is a "Prudent" or "Reckless" Scorpion. Player A knows that, if player B is prudent, then player B will choose not to launch a first strike against player A. If, on the other hand, player B is reckless, with the utilities described in the Reckless Scorpion in the Sand Game and Reckless Response Game, then player B would launch a first strike against player A, and would also retaliate to any strike by player A by launching its weapons. If player B is reckless and would launch a first strike against player A, then it may be to player A’s advantage to eliminate player B’s arsenal *before* player B has a chance to attack. Because player A has such a large number of nuclear weapons, it has the ability to launch a strike against player B and eliminate its weapons before player B has a chance to launch a nuclear strike. Player B does not possess enough weapons to eliminate player A’s arsenal, and therefore only player A considers whether to launch a preemptive strike. Of course, despite its nuclear superiority, player A cannot be completely guaranteed of eliminating player B’s nuclear arsenal, and therefore does incur some risk in attacking player B.

Player A does not know which of the following games player B is playing when considering whether to launch an attack against player A:



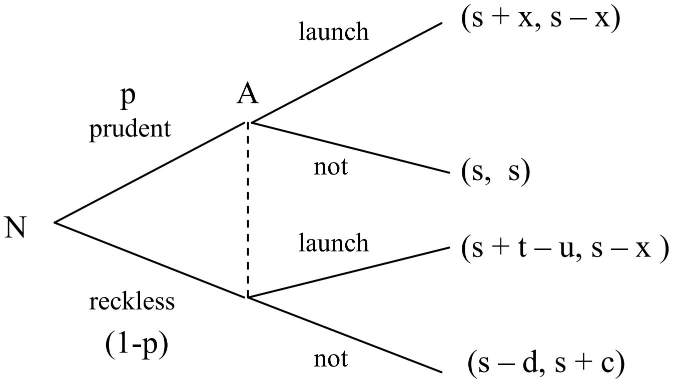
(Figure W.9: “Prudent” Scorpion)



(Figure W.10: “Reckless” Scorpion)

Player A must now decide whether it wants to launch a preemptive strike against player B, without knowing whether or not player B is prudent. If player A could know with certainty that player B was reckless *and* would launch an attack, then player A would want to launch a preemptive attack to wipe out player B's weapons (with as much certainty of success as possible).

Suppose that (p) is the probability that player B is prudent and has the payoffs described in the original Scorpion Game, while (1-p) is the probability that player B is reckless and has the payoffs described in the Reckless Scorpion Game. In the following game, nature moves first to determine whether player B is prudent or reckless:



(Figure W.11: Preemptive Tiger Game)

This game combines the payoffs from the "Prudent" and "Reckless" Scorpion games into a single game, with player A deciding whether to anticipate the actions of player B and launch an attack to eliminate player B's nuclear weapons before it has a chance to use them. Because the prudent player B would not launch a first strike against player A, and would not retaliate against player A even if some of its weapons escaped player A's initial strike, in the top half, the game ends after player A decides whether or not to launch against the prudent player B. When player A launches against the reckless player B, I assume that player A eliminates all of player B's weapons and that player B would not be able to retaliate with nuclear weapons; however, when player A does not launch against the reckless player B, then player B launches its weapons. For player A, the worst outcome occurs when it chooses not to launch against the reckless player B and player B decides to launch against its enemy; player A in this case loses utility (d), where ($d > x$). Even if player B is reckless, if player A launches a first strike he will do so with a limited number of weapons and not with the purpose of completely destroying player B (and rendering player B unable to inflict damage on its opponent); thus player B receives a payoff of (s-x) when hit with a preemptive strike by player A.

Player A gains extra utility from presumably eliminating the threat from the irrational player B's weapons, which is captured by variable (t). Player A receives a higher utility from eliminating the weapons of the irrational player B than from striking the rational player B because the rational player B would not have used its weapons, whereas the irrational player B poses a very real threat to player A. The new variable (u) represents the disutility that player A incurs by launching an attack against player B. This can be thought of as an "uncertainty" variable: because player A is launching an attack to try to eliminate an irrational player B's weapons before it has the chance to use them, player A assumes that the irrational player B will not be able to retaliate. But player A cannot be sure that it has eliminated every nuclear weapon in player B's possession, nor that player B would not launch an attack with conventional forces in response to the strike by player A. This "uncertainty" disutility is captured in the variable (u).

Because player A does not know what type of player B he is facing, he will launch a preemptive attack when the expected payoff from launching meets or exceeds the expected payoff from not launching, based on the probabilities of encountering the prudent or reckless player B:

$$p\ (s + x) + (1\text{-}p)(s + t - u) \geq p\ (s) + (1\text{-}p)(s - d)$$

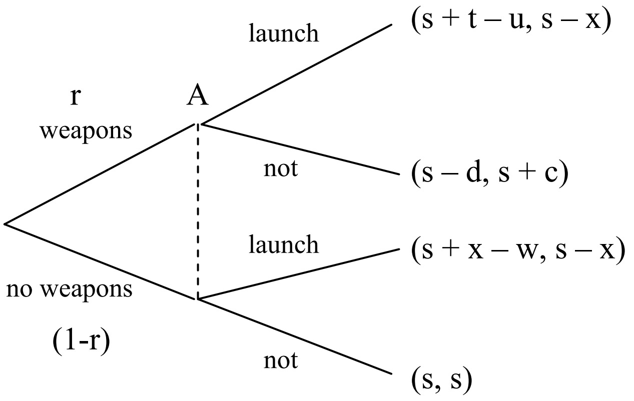
Solving for p, A will launch the preemptive attack when:

$$p \leq \frac{t + d - u}{t + d - u - x}$$

This equilibrium suggests that, as player A's uncertainty disutility (u) increases, he will launch his weapons when there is a decreasing probability that player B is prudent, and therefore an increasing probability that he is reckless. To be willing to incur this uncertainty disutility, player A must have a high certainty that player B is in fact reckless and would use its weapons against player A if it did not eliminate those weapons.

Preemptive Blind Tiger Model

In this final model, I consider a situation in which player A does not know whether player B actually possesses nuclear weapons. In the following game, player A assumes that player B is reckless, and that player B would use nuclear weapons (if it had any) in either a first attack or retaliatory strike if player A does not eliminate such weapons. In the following game, (r) represents the probability that player B possesses nuclear weapons, while (1-r) is the probability that it does not have nuclear weapons. Player A makes the decision about whether to launch a preemptive strike against player A to remove player B's weapons without knowing whether player B actually possesses such weapons:



(Figure W.12: Preemptive Blind Tiger Game)

Player A's payoffs of (t), (u) and (d) are the same as those from the preceding Preemptive Game—Player A once again receives a high amount of utility from eliminating weapons that would be used against it, and loses a lot of utility from failing to eliminate the reckless player B's nuclear weapons. Now, when player A launches its weapons against a reckless player B that does not possess any weapons, he receives the same loss in utility as the Squeamish Tiger because of the public outcry that would result from launching nuclear weapons against a state that was found not to possess any nuclear weapons of its own (presumably, after the strike by player A, it would be discovered that player B had not actually possessed any weapons).

In equilibrium, player A will choose to launch its weapons when its expected payoff from doing so meets or exceeds its expected payoff from not launching, based on the probability that player B possesses nuclear weapons:

$$r\,(s + t - u) + (1\text{-}r)(s + x - w) \geq r\,(s - d) + (1\text{-}r)(s)$$

Solving for r, player A will launch a preemptive strike in this case when:

$$r \geq \frac{w - x}{w + t - u - x - d}$$

This equilibrium suggests that, for an increasing disutility from launching against an unarmed opponent (w), player A will launch only at an increasingly high probability that player B actually possesses nuclear weapons. The same is true of an increasing uncertainty disutility (u) associated with launching an attack against an armed opponent and an increasing disutility from failing to eliminate player B's weapons (d): as both increase, so must player A's belief that player B possesses nuclear weapons in order for player A to launch the preemptive strike.

Is it entirely realistic to expect that a country could be uncertain about an opponent's nuclear weapons capability? Such uncertainty could stem from the fact that player B boasts about having weapons without having presented or demonstrated any hard proof of possessing the weapons; or because player A suspects that player B may have secretly developed nuclear weapons without making such a weapons program public. Regardless of the source of the uncertainty, the equilibrium of this game suggests that a player A that is conscious of world opinion would be hesitant to launch its own nuclear weapons to eliminate those of a reckless opponent, unless there was a high probability that player B did in fact possess nuclear weapons.

Summary

In the preceding models, I demonstrated the conditions under which nuclear weapons may be used in a conflict between two unevenly matched nuclear powers and the conditions under which a large state may wish to launch a preemptive strike to eliminate the nuclear weapons of a smaller state. In the original models of Cold War deterrence, the possession of a nuclear arsenal placed on an automatic launch system, that had the capability of sustaining an initial strike by an opponent and inflicting utter destruction on that opponent, deterred the era's two large nuclear powers from launching their weapons against each other. Moving into the current era, I modeled conflicts between two states with uneven weapons capabilities: player A possessed an immense nuclear arsenal capable of completely destroying its opponent, while player B possessed only a small handful of weapons. In the original Tiger in the Grass Game, I demonstrated that player A would not be deterred from launching its weapons by player B's nuclear weapons capability; however, a "squeamish" player A that incurs a net loss in utility from launching its weapons would be prevented from launching a first strike against player B. Looking at the situation from the opposite point of view, I demonstrated that the "prudent" player B would not launch a first strike against player A, while an irrational or "reckless" player B would launch both a first and retaliatory strike against player A.

From these games of complete information, I moved into an evaluation of games of incomplete information. Because player A possesses overwhelming superiority in weapons capabilities, he has the ability to eliminate player B's nuclear weapons before he has the chance to use them. In the Preemptive Tiger Game, I demonstrated that there are conditions under which player A would be willing to launch a preemptive strike when it is unsure whether player B is prudent or reckless. I then moved to the situation in which player A believes that player B is reckless, but is uncertain of whether player B actually possesses nuclear weapons. In this game, player A must decide whether to launch a nuclear strike against its opponent without knowing whether this reckless opponent actually possesses the offending weapons. In equilibrium in this game, as player A's loss in prestige and credibility from launching its nuclear weapons against an unarmed opponent (w) increases, player A is willing to launch its weapons only at an increasingly high probability that its opponent actually possesses the weapons. Similarly, as the loss of utility it incurs from the uncertainty of launching against a nuclear-armed opponent (u) increases, so must the certainty of player B's possession of weapons in order for player A to launch its own nuclear weapons.

These models present some important conclusions about the conditions under which nuclear weapons may be launched in conflicts between two unevenly matched powers; by elucidating the payoff structures necessary for a launch to take place, they also provide some hints about why nuclear weapons have not been launched offensively since 1945. I assert that the model with which I concluded—the Blind Preemptive Tiger Model—most closely resembles the current situation in which the international community as a whole is evaluating its options for eliminating North Korea's nuclear weapons. The game suggests that a state (or coalition of states) averse to launching nuclear weapons and to being hit by a retaliatory attack will launch a preemptive attack against a proliferating state only when there is a high probability that the state actually does possess nuclear weapons. The games also suggest that only an irrational new proliferator with a handful of nuclear weapons would be willing to launch these weapons against a superior opponent.

Of course, launching a nuclear strike is not the only means by which a state may be confronted or disarmed. The international community has devised a treaty system to prevent nuclear proliferation that certainly does not contain any provision for the use of nuclear weapons against an offending state. It is still important to consider these models, however, as they offer insight into the actual threat posed by nuclear weapons in the hands of rogue actors. Not every rogue state is irrational, and indeed a rogue state pursuing nuclear weapons may be a very rational actor given its own security concerns—particularly if a state such as the US is indeed "squeamish" and would likely be prevented from launching a nuclear attack against a small nuclear-armed opponent. This paper explicitly addresses the world nonproliferation regime and its success in preventing the accumulation of nuclear weapons by small states, but the world may not be fully safe from nuclear proliferation until all states—including the US and other current nuclear powers—agree to forever renounce nuclear weapons.

Appendix C:

Case Studies of Select Countries That Have Violated the International Nonproliferation Regime

South Africa

Until the mid-1960's, the nuclear programs of both the United States and the United Kingdom were dependent on South Africa's uranium deposits. In fact, these two countries provided the funding for South Africa's extraction plants. Under Eisenhower's "Atoms for Peace" program of the 1950's, the United States supplied South Africa with a nuclear reactor; by 1958 South Africa had determined to start its own nuclear research program. Partly because of the US dependence on South African uranium, administrations were reluctant to criticize the South African apartheid regime. South Africa announced plans to begin its own uranium enrichment program in 1970, although the state did not declare that it had any military intentions. At this time, the state decided not to sign the NPT (Fig 1999).

Even though South Africa did not sign the NPT, the sale of highly enriched uranium to South Africa was not halted until 1976; in 1977 the UN initiated a mandatory arms embargo against the South African regime. Nonetheless, France had signed a contract with South Africa for a new reactor in 1976 and continued to train South African nuclear scientists. Evidence obtained in the 1990s also indicates that West Germany helped to train South African nuclear specialists. Europe did not end its collaboration with South Africa until 1985, at which time the United States also halted its purchase of uranium and sale of nuclear materials and technologies to South Africa (which had occurred prior to this time despite the fact that South Africa had not signed the NPT and thus was not eligible to receive nuclear technology from nuclear weapons states). By this time, South Africa had already decided to develop nuclear weapons (Fig 1999).

In 1993, South Africa admitted that it had developed nuclear weapons. In fact, its indigenous uranium enrichment plant supplied enough highly enriched uranium for the production of six nuclear weapons between 1978 and 1990. In 1990 the regime, perhaps fearing the installation of a new government, or desiring to gain entrance into the international nuclear circle, destroyed all of its weapons. The country signed the NPT in 1991. Evidence from South African disclosures after 1993 indicates that the infliction of sanctions against South Africa and withholding of highly enriched uranium encouraged South Africa to develop its own enrichment plants—likely with the help of West Germany, France, and also Israel. The more targeted sanctions instituted in the 1980s were largely too late to halt South Africa's nuclear weapons production, which was already underway by the time the Reagan administration blocked nuclear technology sales in 1986. In fact, the sanctions may have galvanized South African scientists to try to beat the restrictions (Fig 1999).

In addition to a desire to "beat" the restrictions of the sanctions, South Africa also had a number of other incentives for developing nuclear weapons. Decolonization in southern Africa threatened South Africa's hegemony and left the country surrounded with newly independent, leftist states. Of course the South African apartheid regime was highly threatened by the imbalance of white minority rule of the country; the state likely sought the ability to signal both to critical outsiders and to its own citizenry that it was powerful and deserved to be taken seriously (Walters 1987).

South Africa thus managed to develop a nuclear weapon without signing the NPT, but this process was largely facilitated by the fact that the country received much of the technology transfer that should have occurred only under the auspices of the NPT. Additionally, South Africa had its own source of uranium and may have been pushed to refine this uranium itself because of the sanctions imposed against it. IAEA inspections of South Africa's nuclear program following its signing of the NPT in 1991 suggests that the sanctions may have limited South Africa to developing only crude nuclear devices (Fig 1999); regardless, the country still exploited the nuclear energy assistance that it received and the foreign support for its uranium industry for the creation of a nuclear weapons program.

Libya

Libya signed the NPT in 1969, before the rise of leader Mummar al-Qadhafi, and ratified the treaty in 1975 (Bahgat 2004, 388). The IAEA has asserted that, starting in the 1980s and continuing through 2003, Libya secretly developed a program for the conversion and enrichment of uranium (Bahgat 2004, 389), presumably for the purposes of producing the materials necessary for a nuclear weapon. Constant squabbling over its borders likely contributed to Libya's decision to develop chemical and nuclear weapons (Simons 2003, 129). The IAEA also concluded that Libya's program was much less sophisticated than those of Iran and North Korea, and that the country relied on outside help for the development of much of its nuclear technology. This lack of indigenous expertise, combined with the strict UN sanctions imposed on the country in the early 90s mainly in response to Libya's sponsor of several terrorist acts (Bahgat 2004, 375), prevented the country from making much progress in developing a nuclear weapon (Bahgat 2004, 389). The economic sanctions imposed on Libya are estimated to have cost the country $26.5 billion (Bahgat 2004, 384). American sanctions imposed against the country prohibited further American investment in Libya's oil fields; banned the sale of US manufactured oil equipment to the country; and halted the importation of Libyan crude oil (Simons 2003, 132-3).

In March of 2003, Libya initiated talks with Britain and the United States to indicate that the country would voluntarily abandon its pursuit of weapons of mass destruction, including nuclear weapons (Bahgat 2004 387). The move prompted the lifting of sanctions against Libya (Bahgat 2004 374), and the country was warmly embraced by the international community.

Indeed, President Bush announced that nations that pursued similar paths would be invited to "rejoin the international community" (Bahgat 2004 373). Libya signed the NPT Additional Protocol in March of 2004 and has cooperated with the IAEA's inspections (Bahgat 2004, 389). Oil companies that were forced out of Libya by the sanctions regime have started to return to the country to tap its extensive resources (Bahgat 2004, 377). Although Libya's situation has been painted with a rosy outlook lately, what is important for my analysis is the fact that Libya pursued its secret weapons program while it was a member of the NPT, and received assistance for its endeavors from outside agents.

Iraq

The failure to detect Iraq's clandestine nuclear weapons program was largely responsible for the introduction of the Additional Protocol (Hirsch 2004, 142). Iraq signed the NPT at its inception; it was believed to be developing nuclear weapons during the 1970s, until Israel's bombing campaign destroyed the Osiraq reactor in 1981. During the 1980s, the international community believed that the destruction of the reactor effectively destroyed Iraq's ability to develop a nuclear weapon (Spector 1987, 19); Iraq was well aware of this perception and Iraqi officials pointed to the destruction of the reactor as proof that they were not developing a nuclear weapon (Kay 1995). Indeed, during the 1980s Iraq was actively and visibly involved with the IAEA (Kay 1995).

When the Gulf War ended and IAEA inspectors were finally granted wide access to nuclear facilities in Iraq, they discovered that Iraq had been actively acquiring the technology and materials necessary to produce highly enriched uranium that could be used in a nuclear weapon (Kay 1995). In fact, the decision to pursue the uranium enrichment program was made following the 1981 attack on the Osiraq reactor. According to Jaffar dhia Jaffar, the head scientist overseeing Iraq's nuclear weapons program, Iraqi officials explicitly decided to pursue the uranium enrichment program while remaining a member of the NPT. According to Jaffar, this decision to stay in the NPT was made based on a desire, "not to attract any undue attention to Iraq's developing nuclear program that would complicate procurement and development efforts" (Kay 1995). The scientists wished the world to believe that Israel had, "'destroyed [its] nuclear capacity'" (Kay 1995). The Iraqis' deceptions were so complete that the Iraqi IAEA governor, while acting as *chairman* of the Iraq Atomic Energy Commission, was overseeing the development of a nuclear weapons program during the early 1980s. Perhaps the greatest failure of the IAEA was the fact that the Iraqi nuclear weapons program at Tuwaitha was based at the same compound where IAEA inspectors were evaluating Iraq's supposedly peaceful nuclear energy program (Kay 1995). In other words, Iraq remained fully in compliance with the safeguards imposed by the NPT and was still able to develop a sophisticated nuclear weapons program (Lewis 2004, 247).

Iraq conducted many different activities to deceive the international community about its nuclear weapons capabilities. Personnel involved with the project circumvented export controls by altering project codes, lying on purchase orders, and making multiple purchases of sensitive materials. For greater security, Iraq dispersed its nuclear activities across many different sites, and executed scientific personnel who even hinted at disloyalty to the Iraqi regime. Additionally, Iraq offered positive incentives, including food and entertainment, to the hand-picked IAEA inspectors that it allowed to visit limited sites within Iraq. Fortunately the IAEA access to Iraq following the First Gulf War discovered and effectively shut down Iraq's nuclear weapons programs—Iraq was only eighteen to twenty-four months away from possessing a crude nuclear weapon at the time of the invasion (Kay 1995). Most critical for the purposes of my models is the fact that Iraq developed its nuclear weapons program while still a member of the NPT—in fact, it used its status as an NPT member to convince the world that it was not pursuing a nuclear weapons program, and utilized the technology transfers to which it was legally entitled to aid in its weapons program.

Iran

Like Iraq, Iran signed the NPT at its introduction (Spector 1987, 19). The country also signed the Additional Protocol in December of 2003 (Bowen and Kidd 2004, 257). IAEA inspections in 2003 caused Iran to make its first public admission of a program to develop uranium enrichment capabilities (Bowen and Kidd 2004). Iran has obtained most of its assistance in developing its nuclear capabilities from China, Russia and North Korea, two of which are declared weapons states under the NPT (Kibaroğlu 2002, 33).

Iran's nuclear program dates to the days of Shah Reza Pahlavi. The Shah purchased a nuclear reactor from the US in 1967 and acquired contracts for nuclear fuel and additional reactors with the US, France, Germany, England and India during the 1970s. During the 1970s Iran also pursued a clandestine nuclear weapons research program and attempted to purchase highly-enriched uranium; at the end of the Shah's reign, the country lacked the resources and personnel for an advanced weapons program (Kibaroğlu 2002, 34-5). In 1984, Iran built a research center that it failed to declare to the IAEA until 1992. In 1987 the country signed an agreement with Pakistan and sent its scientists to receive training there; during the 1980s the country also received substantial assistance from China in developing its "peaceful" nuclear infrastructure (Kibaroğlu 2002, 35-7). Iran has also signed agreements with Russia for the sale of Russian nuclear reactors, all of which is allowed under the NPT framework. Most troubling, Iran has also developed a ballistic missile program, with the help of Libya, North Korea and

Russia, and tested missiles capable of delivering nuclear warheads in 1998 and 2000 (Kibaroğlu 2002, 39).

Although Iran has substantial fossil fuel reserves, the country insists that the development of its nuclear capacity is necessary to meet its energy needs (Kibaroğlu 2002, 38). It is largely believed that a strong impetus for the development of Iran's nuclear program was its war with Iraq in the 1980s (Bowen and Kidd 2004, 263), and an increasing desire to be a leader in the Middle East (Kibaroğlu 2002, 43). Most troubling for the international community is the fact that Iran has consistently deceived the world about its nuclear capabilities and intentions. On February 4, 2006 the board of the IAEA voted to report Iran to the UN Security Council (Sciolino 2006). The Resolution stated that the IAEA, "after nearly three years of intensive verification activity…is not yet in a position to clarify some important issues relating to Iran's nuclear programme or to conclude that there are no undeclared nuclear materials or activities in Iran" (IAEA Board of Governors 2006). The report demands greater transparency in Iran's nuclear activities and a halt of all reprocessing and enrichment activities but stresses that the Board still prefers a diplomatic solution to the problem (IAEA Board of Governors 2006). Upon the passage of the resolution, Iran announced that it would terminate its relationship with the IAEA and restart full production of highly enriched uranium (Sciolino 2006). While some alarmists insist that Iran may be able to enrich enough uranium to build a warhead within months, US Intelligence estimates that Iran is likely five to ten years away from its first device. The more conservative judgments of when Iran would be able to build its first weapon assume that the country would acquire materials such that it could quickly produce a small arsenal when its first weapon is within reach (Broad and Sanger 2006). Most important for my models is the fact that Iran, like Iraq, has likely pursued a nuclear weapons program while still a member of the NPT—in fact, Iran has likely exploited the technology transfers permitted under the NPT to promote its weapons program by receiving technology and equipment from Russia and China.

North Korea

US intelligence discovered a new research reactor under construction at the Yongbyon Nuclear Research Center in 1980, at which time both the US and Soviet Union pressured North Korea to accept IAEA inspectors. North Korea signed the NPT in December of 1985, becoming party to the treaty as a state that did not possess nuclear weapons, but the country avoided IAEA inspections and began building a larger nuclear reactor by the end of the decade. North Korea finally acceded to IAEA safeguards and inspections, and director Hans Blix toured North Korea's now-declared facility at Yongbyon. Samples taken by the IAEA in 1992 indicated that North Korea had reprocessed much more plutonium than it had declared to the agency—possibly enough for one to two nuclear weapons (IISS 2004, 3-7).

When the IAEA pressed for "special inspections" of North Korea's facilities, the country threatened to withdraw from the NPT. It cited the nuclear threat posed by the United States as its justification for withdrawal. In July of 1993 North Korea agreed to abandon its reprocessing program in exchange for a light-water reactor from the US. Extensive negotiations involving the US, South Korea and the IAEA over the next year resulted in the 1994 Agreed Framework, under which North Korea would receive peaceful nuclear reactors in exchange for a freeze on its reprocessing activities and expanded IAEA access to its facilities. North Korea remained a member of the NPT, which some felt undermined the treaty, "because it allowed North Korea to remain in violation of its safeguards obligations and to retain a small amount of undeclared plutonium" (IISS 2004, 11).

Implementation of the Framework proved difficult, and the US struggled to limit North Korea's missile program. In 1998, US intelligence detected the possible construction of a new nuclear facility in North Korea and in August the country tested a long-range missile that flew over Japan. In negotiations in 2000, North Korea appeared interested in a bargain to give up its missiles in exchange for the launch of civilian satellites. In 2002, President George W. Bush labeled North Korea a member of the 'axis of evil,' and pressured North Korea to give up its weapons and missiles in exchange for the lifting of economic sanctions and better political relations. In negotiations in October, North Korea acknowledged that it was pursuing a secret enrichment program, which a North Korean official declared was "justified by the Bush administration's threats and hostility" (IISS 2004, 17).

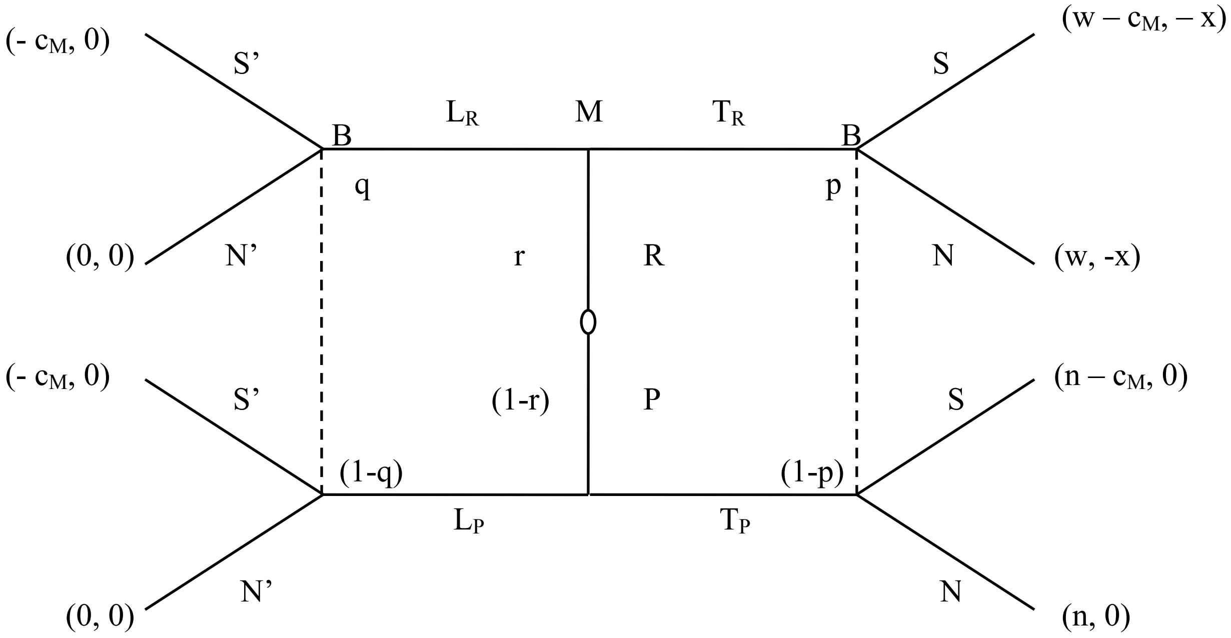
North Korea was thus admitting to violating the Agreed Framework. The Bush administration hoped that the threat of new sanctions would deter North Korea from leaving the NPT and encourage it to take advantage of the better political relations it would receive if it renounced its weapons; however, North Korea unfroze its plutonium production and expelled IAEA inspectors in December of 2002. North Korea announced its withdrawal from the NPT in January 2003. In private meetings in July 2003, North Korean officials asserted that the country had reprocessed 8,000 spent fuel rods, which US intelligence could not independently verify (IISS 2004 17-23). In October 2003 President Bush indicated that the US would issue written security guarantees to the country in exchange for full disarmament (IISS 2004, 26). However, in February 2005, North Korea announced that it possessed nuclear weapons; in May of the same year, the director of the IAEA announced that North Korea likely had enough plutonium for five or six nuclear weapons (CNN.com 9 May 05). Although North Korea announced in September 2005 that it would give up its nuclear weapons program (CNN.com 19 Sep 2005), negotiations have since stalled and the country’s weapons program remains a major concern for the international community.

Appendix D:

NPT Model Variations

NPT Model A Variation: $c_B = 0$

The following variation of NPT Model A examines the equilibria that arise when imposing sanctions on player M is costless for player B—i.e. when ($c_B = 0$). When player B does not incur any disutility from the imposition of sanctions, the game is structured as follows, with the payoff variables the same as those in the models in the body of the paper:



(Figure A.4: Basic NPT "Costless" Sanctions Model)

At each node at which player B moves, he is indifferent between sanctioning and not sanctioning player M. Such a situation may occur when the sanction imposed is almost costless to the large country, as in the case where suspending trade relations with the small country does not hurt the large country very much because the lost trade constitutes a very small part of the large state's overall economy.

NPT Model A.V: Solutions

The model is solved under the assumption that ($w > n > c_M$). Because imposing sanctions is costless for player B, he is indifferent between sanctioning or not sanctioning player M at every node. Thus for any combination of strategies pursued by player M, there are four strategies that constitute a best response for player B: (N', N), (N', S), (S', N) and (S', S). For there to exist a separating equilibrium in which the rogue player M leaves the NPT and the peaceful player does not, it must be true that the rogue player receives a payoff from leaving the treaty that is greater than or equal to the payoff that he receives from remaining in the treaty, and that the peaceful player receives a payoff from remaining in the treaty that is greater than or equal to the payoff that he receives by leaving the treaty, given the strategies pursued by player B. Yet because the benefits of treaty membership exceed the costs of sanctions for both types of player M, for any strategy pursued by player B, the best response of both types of player M is to remain in the NPT (strategy T). The model thus yields the following four equilibria:

$$\{(T_R, T_p), (N', N), p = r, 0 \leq q \leq 1\} \tag{a.2}$$

$$\{(T_R, T_p), (N', S), p = r, 0 \leq q \leq 1\} \tag{a.3}$$

$$\{(T_R, T_p), (S', N), p = r, 0 \leq q \leq 1\} \tag{a.4}$$

$$\{(T_R, T_p), (S', S), p = r, 0 \leq q \leq 1\} \tag{a.5}$$

Because the benefits from remaining a member of the NPT outweigh the costs to player M for the imposition of sanctions, both types of player M always choose to remain within the NPT, for any combination of strategies pursued by player M. The incentive structure of the treaty seems to be such that, regardless of the threat of sanctions by player B, both types of player M are still willing to sign the treaty, and thus the treaty fails to separate proliferators from non-proliferators.

This model is now solved under the assumption that ($w > c_M > n$). Player B is still indifferent between sanctioning and not sanctioning at every node, and his best response to player M's strategy is still any of the four strategies (N', N), (N', S), (S', N) and (S', S). The benefits of remaining in the NPT still exceed the costs of sanctions for the rogue player M, and thus his best response to any strategy pursued by player B is still to remain within the NPT (strategy T). There will therefore be no equilibria in which the rogue player M chooses to leave the NPT. For the peaceful player M, however, the costs of sanctions now exceed the benefits of treaty membership. This constraint is binding only when player B sanctions those players that remain in the NPT and does not sanction those that leave the NPT. Under this condition, the peaceful player M's best response is to leave the NPT. For all other possible strategies pursued by player B, the peaceful player M's best response is to remain within the NPT. Under the assumption that ($w > c_M > n$), the model thus yields the following equilibria:

$\{(T_R, T_p), (N', N), p = r, 0 \leq q \leq 1\}$ (a.2)

$\{(T_R, T_p), (S', N), p = r, 0 \leq q \leq 1\}$ (a.4)

$\{(T_R, T_p), (S', S), p = r, 0 \leq q \leq 1\}$ (a.5)

$\{(T_R, L_p), (N', S), p = 1, q = 0\}$ (a.6)

This game yields three pooling equilibria identical to equilibria (a.2), (a.4) and (a.5) in the first solution set; however, under the current assumption the model yields the separating equilibrium in (a.6). Because the benefit from remaining in the NPT still exceeds the cost of sanctions for the rogue player M ($w > c_M$), the rogue player M always chooses to sign the treaty. Player B's strategy in equilibrium (a.6) thus causes the rogue player M to remain within the treaty while the peaceful player M finds it in his best interest to leave. This constraint is not binding on the peaceful player M for any other set of strategies pursued by player B and thus in the other three equilibria (a.2), (a.4) and (a.5), both types of player M choose to sign the treaty. Interestingly, when player B plays a strategy of (S', S) and player M realizes that he will be sanctioned regardless of what he chooses, then it is still in the peaceful player M's best interest to sign the treaty and receive the benefits of nuclear technology transfer.

Finally, the game is solved under the assumption that ($c_M > w > n$). Player B still has the same four possible equilibrium strategies as above; but now, the cost of sanctions exceeds the benefits of NPT membership for both types of player M. This constraint is only binding when player B is playing (N', S), sanctioning those players that remain in the NPT and not sanctioning those that leave. Under these conditions, the best response of both types of player M is to leave the NPT. All the other equilibria calculations remain the same, and thus solving under the assumption that ($c_M > w > n$) yields the following equilibria:

$\{(T_R, T_p), (N', N), p = r, 0 \leq q \leq 1\}$ (a.2)

$\{(T_R, T_p), (S', N), p = r, 0 \leq q \leq 1\}$ (a.4)

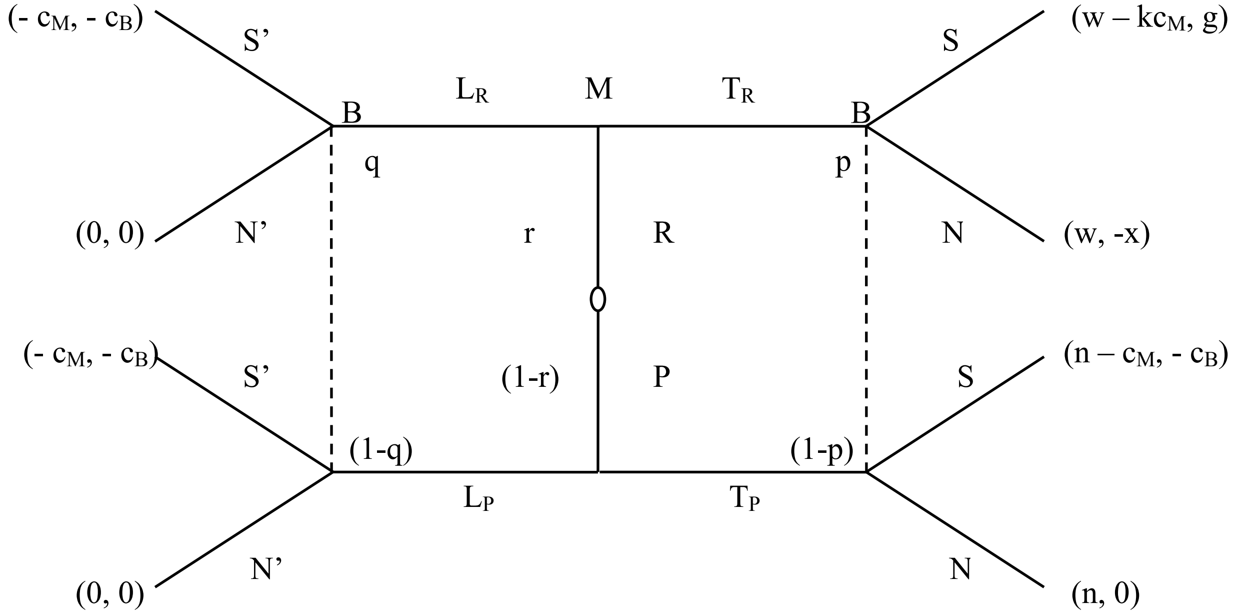
$\{(T_R, T_p), (S', S), p = r, 0 \leq q \leq 1\}$ (a.5)

$\{(L_R, L_p), (N', S), q = r, 0 \leq p \leq 1\}$ (a.7)

All of the equilibria in this game are still pooling equilibria, and (a.2), (a.4) and (a.5) are identical to three of those observed under the first two sets of assumptions. But now there is a pooling equilibrium in which both types of player M choose not to sign the NPT (a.7). The cost of sanctions is so high in this case that it drives both types of player M out of the treaty and the two types of player M again fail to separate in the desired manner, with the rogue player M leaving the NPT and the peaceful player remaining a signatory. In equilibria (a.2) and (a.4), player B is not sanctioning those players that remain within the NPT and thus both players choose to remain signatories; in equilibrium (a.5), player M faces sanctions no matter what it plays and thus chooses to receive the benefits associated with remaining within the NPT.

NPT Model C Variation: "Nervous" Targeted Sanctions

In this variation of NPT Model C, the large state is nervous about imposing sanctions on players that sign the NPT and thus receives a net gain in utility (g) only when he sanctions the rogue player that signs the NPT, and not when he sanctions the peaceful member of the NPT. Player B may have such a utility function if he is nervous about imposing sanctions on a peaceful player that has no intention of pursuing a nuclear weapon; or this player B may believe that it does not gain any utility from sanctioning a peaceful player M, because such a player would not have developed nuclear weapons anyway. Thus, when player B sanctions the rogue player M that remains within the NPT, he receives the gain in utility (g) and believes that he is eliminating the ability of player M to develop nuclear weapons, and ($x > g$). In all the other cases in which player B imposes sanctions, he receives the payoff of ($-c_B$) and ($c_B > 0$). Again, the rogue player M receives a payoff of ($-kc_M$) when he remains within the NPT and player B imposes targeted sanctions. The model is structured as follows:



(Figure C.2: NPT "Nervous" Targeted Sanctions Game)

NPT Model C.V: Solutions

The model is first solved under the assumption that ($w > n > c_M$). When player M leaves the NPT (strategies L_R and L_p), then player B's best response is not to impose sanctions (strategy N'). However, when player M remains in the treaty, player B's best response depends on the likelihood that he is facing a rogue player—i.e. depends on the value of (p). For player B to play a strategy in which it sanctions those players that sign the NPT (S), it must be the case that player B receives an expected payoff from sanctioning player M that is greater than the expected payoff of not sanctioning player M. Thus in equilibrium, player B will pursue strategy (S) when [$(p)(g) + (1 – p)(- c_B) \geq (p)(-x) + (1 – p)(0)$], and will play strategy (N) when [$(p)(g) + (1 – p)(- c_B) \leq (p)(-x) + (1 – p)(0)$]. Player B will always play strategy (N'), and thus player B has two different possible strategies in equilibrium, depending on the value of (p): (N', N) and (N', S).

Because the value of NPT membership exceeds the cost of sanctions for both types of player M, the peaceful player's best response to either of player B's equilibrium strategies is to remain within the NPT (strategy T_p). For a separating equilibrium to exist in which the rogue player M leaves the treaty, it must be the case that the rogue player M's payoff from leaving the NPT meets or exceeds the payoff he receives from remaining within the NPT. When player B is playing the strategy (N', N), then the rogue player M's best response is to remain in the NPT and receive the payoff of (w). When player B is playing (N', S), then the rogue player must consider the value of the targeted sanctions in relation to the payoff he would receive by leaving the NPT. Player B will remain in the NPT (strategy T) when his payoff from doing so meets or exceeds his payoff from leaving the NPT—that is, when $[(w - kc_M) \geq 0]$. The rogue player M would leave the NPT when $[(w - kc_M) \leq 0]$; but that would mean that player M would be playing a strategy of (L_R, T_p), since the peaceful player always remains in the NPT. But this is not an equilibrium strategy, because player B's best response to a strategy of (L_R, T_p) is *not* (N', S). Thus the only possible equilibrium when player B plays (N', S) is for both types of player M to remain in the NPT. Solving for the values of (p) and (k), the model yields only the following two equilibria:

$$\{(T_R, T_p), (N', S), r = p, 0 \leq q \leq 1, p \geq (c_B/[c_B + g + x]), k \leq (w/c_M)\} \quad \text{(c.6)}$$

$$\{(T_R, T_p), (N', N), r = p, 0 \leq q \leq 1, p \leq (c_B/[c_B + g + x])\} \quad \text{(c.7)}$$

Unlike the version of Model C presented in the body of the paper, in this case there is an equilibrium in which player B does play something other than (N', S). In equilibrium (c.7), player B chooses not to sanction either type of player M because the probability that player M is rogue is sufficiently low. When the probability that player M is rogue is sufficiently high, then player B sanctions the players that sign the NPT and does not sanction those that do not (c.6); since the cost of being sanctioned is sufficiently low, both types of player M still opt to remain in the NPT, as long as the additional cost of the targeted sanctions (k) is sufficiently low. If the value of (k) is too high, which is a consideration only when player B plays (N', S), then the rogue player M leaves the NPT; however, this is not an equilibrium because player B's best response to a strategy of (L_R, T_p) is *not* (N', S) and thus there are only the two pooling equilibria.

Now, the model is solved under the assumption that ($w > c_M > n$). Player B makes the same calculation about his equilibrium strategies as above, and will play (N', S) when (p) is sufficiently high and (N', N) otherwise. In the special case that (p) equals one, then it must be the case according to the model that player M is playing a strategy of (T_R, L_p), to which a strategy of (N', S) by player B is the best response. When the value of (p) is equal to zero, then player M must be playing a strategy of (L_R, T_p), and player B's best response is to play (N', N).

The peaceful player M still chooses to remain in the NPT when player B is playing (N', N); but since the cost of sanctions now exceeds the benefits of NPT membership, the peaceful player M leaves the NPT when player B is playing (N', S). The rogue player's best response to a strategy of (N', N) by player B is still to remain in the treaty. When player B is pursuing a strategy of (N', S), the rogue player M makes the same calculation as in the above solution set and chooses to remain in the treaty if the value of (k) is sufficiently low. If the value of (k) is too high, the rogue player M's best response to player B's strategy of (N', S) is to leave the NPT; since the peaceful player M also leaves under such conditions, player M would thus be pursuing a strategy of (L_R, L_p), to which player B's strategy of (N', S) is a best response and thus an equilibrium. Solving under the assumption that ($w > c_M > n$) thus yields the following equilibria:

$$\{(T_R, L_p), (N', S), q = 0, p = 1, k \leq (w/c_M)\} \qquad \text{(c.8)}$$

$$\{(T_R, T_p), (N', N), r = p, 0 \leq q \leq 1, p \leq (c_B/[c_B + g + x])\} \qquad \text{(c.9)}$$

$$\{(L_R, L_p), (N', S), q = r, p \geq (c_B/[c_B + g + x]), k \geq (w/c_M)\} \qquad \text{(c.10)}$$

Equilibrium (c.8) is a separating equilibrium, but not one that the international community would hope to encourage, since the rogue player M stays in the NPT and the peaceful player M leaves, as long as the value of (k) is sufficiently low. Because the cost of the sanctions is too high, the peaceful player M leaves the treaty when player B plays (S). In equilibrium (c.9), both types of player M decide to remain within the treaty: player B never sanctions player M because the probability that player M is rogue is sufficiently low. In the final equilibrium (c.10), the probability that player M is rogue is sufficiently high that player B sanctions those that remain within the NPT; the additional cost of the targeted sanctions is sufficiently high that both types of player M are thus driven out of the NPT.

Finally, the model is solved under the assumption that ($c_M > w > n$). Player B makes the same calculations as in the solution sets presented above: he plays a strategy of either (N', N) or (N', S) in equilibrium, depending on the value of (p). When player B is playing a strategy of (N', N), the best response for both types of player M is to remain in the NPT. For there to be a separating equilibrium in which the rogue player leaves the NPT and the peaceful player does not, it must be true that the payoff to the rogue state from leaving the NPT meets or exceeds the payoff for remaining in the NPT, and that the payoff to the peaceful state for remaining in the NPT meets or exceeds its payoff for leaving the NPT. But under the assumption that ($c_M > w > n$), the cost of sanctions exceeds the benefits of treaty membership for both types of player M, and thus both the rogue and the peaceful player choose to leave the NPT when player B plays (N', S). The model thus yields the following two pooling equilibria under the constraint that ($c_M > w > n$):

$$\{(T_R, T_p), (N', N), r = p, 0 \leq q \leq 1, p \leq (c_B/[c_B + g + x])\} \qquad (c.11)$$

$$\{(L_R, L_p), (N', S), q = r, p \geq (c_B/[c_B + g + x])\} \qquad (c.12)$$

These equilibria are very similar to (c.9) and (c.10). Because the cost of sanctions exceeds the benefits of signing the treaty for both types of player M, the value of (k) has no bearing on the equilibria. When the probability that player M is rogue is sufficiently high, player B sanctions those players that choose to remain in the NPT; but because the cost of sanctions exceeds the benefits of the treaty for both types of player M, both the peaceful and the rogue player M are driven out of the NPT when player B plays a strategy of (N', S). There are thus no separating equilibria when the cost of sanctions exceeds the benefit of NPT membership for both types of player M.

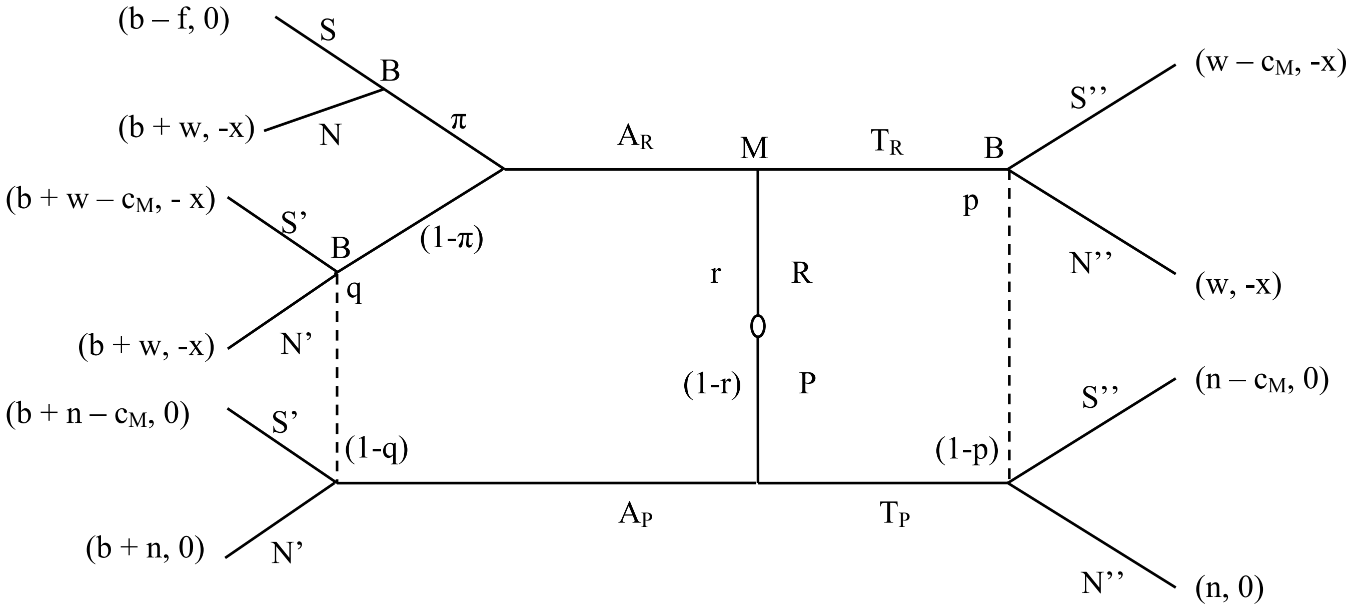
Appendix E:

AP Model Variation

AP Model B Variation: “Costless” Sanctions

The following model assumes that the cost to player B of imposing sanctions on player M is zero, i.e. ($c_B = 0$). This assumption may be appropriate in cases in which the monetary costs of maintaining a sanction regime constitutes a very small part of the big state’s overall economy, or when the large state decides to ignore the cost of sanctions because it believes them to be unimportant compared to the goal of halting nuclear weapons proliferation. Whether it is ever true that there is zero *political* cost associated with sanctioning a small state is uncertain. Nevertheless, the case in which player B faces no cost from imposing sanctions on player M is still a valid avenue of exploration.

When implementing sanctions against player M does not impose a cost on player B, then player B becomes indifferent between sanctioning and not sanctioning player M at every node, except at the node at which the rogue player M is detected. When player B detects that player M is rogue, player B prefers to impose sanctions on player M and thereby avoid incurring the disutility of (-x) that would arise if player M were not sanctioned and thus allowed to develop its nuclear weapons. The AP game is now structured as follows:



(Figure E.2: AP "Costless" Sanctions Model)

AP Model B.V: Solutions

The model is solved under the assumption that ($w > n > c_M$). Except at the node at which the rogue player is detected, in which case player B prefers to impose sanctions (strategy S), player B is indifferent between sanctioning and not sanctioning at every other node. There are thus four possible strategies for player B in equilibrium: (S, S', S''), (S, S', N''), (S, N', S'') and (S, N', N''). The peaceful player M's best response to strategies (S, S', S''), (S, N', S''), and (S, N', N'') is to sign the AP. Only when player B is playing a strategy of (S, S', N'') is the peaceful player M's best response to not sign the AP.

For a separating equilibrium to exist in which the rogue player M does not sign the AP and the peaceful player does, it must be the case that the rogue player's expected payoff from not signing the AP meets or exceeds its expected payoff from signing the AP, and that the peaceful player's best response to player B's strategy is to sign the AP. When player B is playing (S, S', S''), then the rogue player M will not sign the AP when the expected value of not signing meets or exceeds the expected value of signing: $[(w - c_M) \geq (\pi)(b - f) + (1 - \pi)(b + w - c_M)]$. The rogue player M will sign the AP when $[(w - c_M) \leq (\pi)(b - f) + (1 - \pi)(b + w - c_M)]$. When player B plays the strategy (S, N', N''), the rogue player M again compares the expected value of signing with not signing: the rogue player M will not sign the AP when $[(w) \geq (\pi)(b - f) + (1 - \pi)(b + w)]$ and will sign the AP when $[(w) \leq (\pi)(b - f) + (1 - \pi)(b + w)]$. In response to player B employing a strategy of (S, N', S''), the rogue player M's best response is not to sign the AP when $[(w - c_M) \geq (\pi)(b - f) + (1 - \pi)(b + w)]$ and to sign the AP when $[(w - c_M) \leq (\pi)(b - f) + (1 - \pi)(b + w)]$. Finally, when player B is playing (S, S', N''), then the rogue player M's best response is always to not sign the AP, since he will certainly face sanctions if he signs the AP. Under the condition that $(w > n > c_M)$, the model yields the following equilibria:

$$\{(T_R, A_p), (S, S', S''), p = 1, q = 0, \pi \geq [b/(f+w-c_M)]\} \qquad (e.3)$$

$$\{(T_R, A_p), (S, N', N''), p = 1, q = 0, \pi \geq [b/(f+w)]\} \qquad (e.4)$$

$$\{(T_R, A_p), (S, N', S''), p = 1, q = 0, \pi \geq [(b+c_M)/(f+w)]\} \qquad (e.5)$$

$$\{(A_R, A_p), (S, S', S''), 0 \leq p \leq 1, q = r, \pi \leq [b/(f+w-c_M)]\} \qquad (e.6)$$

$$\{(A_R, A_p), (S, N', N''), 0 \leq p \leq 1, q = r, \pi \leq [b/(f+w)]\} \qquad (e.7)$$

$$\{(A_R, A_p), (S, N', S''), 0 \leq p \leq 1, q = r, \pi \leq [(b+c_M)/(f+w)]\} \qquad (e.8)$$

$$\{(T_R, T_p), (S, S', N''), 0 \leq q \leq 1, p = r, \pi \geq 0\} \qquad (e.9)$$

In the first three equilibria (e.3, e.4 and e.5), the rogue player M chooses not to sign the AP and the peaceful player M does sign the AP, much like equilibrium (e.1) in the original AP Model B. In all three of these cases, the probability of detection is sufficiently high that the rogue player chooses not to sign the AP and instead remains within the original NPT framework. The model thus produces the desired separating equilibrium, but the rogue player still retains the benefits of

NPT treaty membership (w) when he chooses not to sign the AP. Comparing the three equilibria, and holding the values of all other variables constant, the probability of detection (π) required to force the rogue player not to sign the AP is highest for equilibrium (e.5): since player B will not sanction the undetected player that signs the AP, there must be higher probability of detection in order for the rogue player not to sign the AP. Of these three separating equilibria, the lowest value of (π) necessary to drive the player away from the AP is that in equilibrium (e.4), in which player B only sanctions the detected rogue player M.

The other four equilibria in this model are pooling equilibria (e.6, e.7, e.8 and e.9). The first three pooling equilibria perfectly mirror the three separating equilibria: now the probability of detection as rogue (π) is sufficiently low that the rogue player M chooses to sign the AP. In the first three equilibria, the value of (π) is above the threshold value, such that the rogue player chooses not to sign the AP because of the risk of detection. Even in equilibrium (e.6), when player B sanctions all players at every node, player M chooses to sign the AP because the probability of detection and thus of incurring the fine of (-f) is sufficiently low. In the final equilibrium (e.9), player B sanctions both the detected and the undetected rogue player M that choose to sign the AP, while not sanctioning those that do not sign the AP. Thus both types of player M do better when they do not sign the AP for all values of ($\pi \geq 0$).

The AP Model B Variation yields the same equilibria when solved under the assumption that ($w > c_M > n$) as when solved under the assumption above. The value of (n) does not figure into any of the conditions placed on (π) that determine the equilibria. Because the value of (b) is small in comparison to (c_M), the change in the relative values of (n) and (c_M) does not change any of the equilibria under this new assumption.

Finally, the model is solved under the assumption that ($c_M > w > n$). Solving the model now demands an additional set of assumptions. When the value of (f) exceeds the absolute value of ($w - c_M$), then solving the model under the assumption that ($c_M > w > n$) yields the exact same set of equilibria as solving under the assumption that ($w > n > c_M$). So, for (f) > $|w - c_M|$, the set of all possible equilibria is exactly the same as that for the first solution set of AP Model B Variation.

For the case in which (f) = $|w - c_M|$, some of the equilibria are no longer feasible. In equilibria (e.3) and (e.6), the equilibrium values of (π) are determined by fractions in which the sum ($f + w - c_M$) appears in the denominator. According to the conditions of this model, ($c_M >$ w); when the value of (f) is equal to $|w - c_M|$, the denominators of the fractions associated with (π) in equilibria (e.3) and (e.6) are zero and thus these equilibria are invalid. Equilibrium (e.9) is also eliminated when (f) = $|w - c_M|$. In extended form, the condition determining the value of (π) in equilibrium (e.9) is: (π) $\geq [(b-c_M)/(f+w-c_M)]$. The probability of detecting a rogue state must be greater than or equal to zero ($\pi \geq 0$). Because ($c_M > b$) by definition, the numerator of this fraction must always be negative and thus the equilibrium holds for all values of (π) ≥ 0 in (e.9). But when the denominator of this fraction ($f+w-c_M$) equals zero, as is the case when (f) = $|w - c_M|$, this equilibrium is invalid. The set of all possible equilibria when $f = |w - c_M|$ and when ($c_M > w > n$) is thus:

$\{(T_R, A_p), (S, N', N''), p = 1, q = 0, \pi \geq [b/(f+w)]\}$ (e.4)

$\{(T_R, A_p), (S, N', S''), p = 1, q = 0, \pi \geq [(b+c_M)/(f+w)]\}$ (e.5)

$\{(A_R, A_p), (S, N', N''), 0 \leq p \leq 1, q = r, \pi \leq [b/(f+w)]\}$ (e.7)

$\{(A_R, A_p), (S, N', S''), 0 \leq p \leq 1, q = r, \pi \leq [(b+c_M)/(f+w)]\}$ (e.8)

There are no longer any equilibria in which the big state always sanctions player M at every node. Neither is there an equilibrium in which both types of player M choose not to sign the AP. Because the fine from being detected as rogue and sanctioned (f) exactly equals the total utility of being sanctioned ($w – c_M$), there is no longer an equilibrium in which both types of player M refuse to sign the AP. This leaves four equilibria: two in which the players separate, the peaceful by signing and the rogue by refusing to sign; and two in which both types of player M sign the AP. In this situation, sanctions impose a sufficient cost on player M that there are no longer any equilibria in which player B plays (S'), i.e. in which player B sanctions those players that sign the AP.

Finally, when the value of (f) is exceeded by $|w - c_M|$, equilibrium (e.5) is still eliminated. Under the condition that $(f) < |w - c_M|$, the denominator of the fraction that determines the equilibrium value of (π) in (e.5) is negative; since a probability cannot be negative, (e.5) is not a possible equilibrium. Although ($f+w-c_M$) appears in the denominator of the fraction determining (π) in (e.3), the conditions of the equilibrium state that $(\pi) \geq [b/(f+w-c_M)]$; since (π) must be greater than or equal to zero, then this condition holds for all values of (π), since the value of the fraction itself is negative. Additionally, equilibrium (e.9) again holds, because the negative value of the denominator in the fraction determining the value of (π) now cancels the negative numerator and renders the value of the fraction positive. Now, however, the equilibrium does not hold true for all values of (π) but only when $(\pi) \geq [(b-c_M)/(f+w-c_M)]$. The set of all equilibria for ($c_M > w > n$) and $(f) < |w - c_M|$ is thus:

$\{(T_R, A_p), (S, S', S''), p = 1, q = 0, \pi \geq 0\}$ (e.10)

$\{(T_R, A_p), (S, N', N''), p = 1, q = 0, \pi \geq [b/(f+w)]\}$ (e.4)

$\{(T_R, A_p), (S, N', S''), p = 1, q = 0, \pi \geq [(b+c_M)/(f+w)]\}$ (e.5)

$\{(A_R, A_p), (S, N', N''), 0 \leq p \leq 1, q = r, \pi \leq [b/(f+w)]\}$ (3.7)

$\{(A_R, A_p), (S, N', S''), 0 \leq p \leq 1, q = r, \pi \leq [(b+c_M)/(f+w)]\}$ (e.8)

$\{(T_R, T_p), (S, S', N''), 0 \leq q \leq 1, p = r, \pi \geq [(b-c_M)/(f+w-c_M)]\}$ (e.11)

These equilibria are again very similar to those observed in the case in which ($w > n > c_M$). In this specific iteration of the model, the rogue player M is again induced not to sign the AP when the probability of detection is sufficiently high (e.10, e.4 and 3.5). When the probability of detection is sufficiently low, the rogue player M and the peaceful player M both decide to sign the AP (e.7, e.8). Finally, when the big player sanctions all players that sign the AP, and when the probability of detection is sufficiently high, both types of player M will choose not to sign the AP (e.11).

The most important feature of the above equilibria obtained under the condition that ($c_M > w > n$) is the fact that they are extremely similar in outcome to those observed in the earlier solution sets of this model, and of the original AP Model B in the body of this paper. Depending on the relative values of the payoff variables and the probability of detecting a state's rogue status, the rogue player can be induced to separate from the peaceful player by not signing the AP (equilibria e.3, e.4, e.5, e.10). This occurs when the value of (π) is sufficiently high that the rogue player M's best response to player B's strategy is to leave the NPT; yet, as in the AP Model B in the body of the paper (equilibrium e.1), this occurs at a very low probability of detecting a rogue player's status (π). When, however, the value of (π) is sufficiently low, then both types of player M choose to sign the AP (equilibria e.6, e.7, e.8). This is similar to the pooling equilibrium (e.2) in the original AP Model B. Finally, there is a unique pooling equilibrium in which both types of player M choose not to sign the AP, because by doing so they may avoid sanctions imposed by player B (equilibria e.9 and e.11). Thus even when the cost to player B of imposing sanctions under the AP structure is zero, the model yields results very similar to those that occur when player B incurs a cost for imposing sanctions on player M. And as in the original AP Model B, the players separate themselves at a very low probability of detecting a rogue player's status (π), a flaw remedied in the NPT-AP Normative Model presented in section 3.