Aalto University
School of Business

The Application of Game Theory in Nuclear Deterrence

Bachelor's Thesis
Daniel Harper Kakkonen
Aalto University School of Business
Department of Economics
Fall 2021

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of bachelor's thesis

Author Daniel Harper Kakkonen

Title of thesis The Application of Game Theory in Nuclear Deterrence

Degree Bachelor of Science in Economics and Business Administration

Degree programme Economics

Thesis advisor(s) Pauli Murto & Marko Terviö

Year of approval 2021 **Number of pages** 29 **Language** English

Abstract

The reasons for the absence of nuclear warfare between two nations despite the widespread ownership of nuclear weapons have been studied for decades. An often-used explanation for this is nuclear deterrence, whereby owning nuclear weapons deters other nations with hostile interests from attacking in any way. This literature review studies whether game theory can be applied to explain the absence of nuclear war and if nuclear deterrence is a rational explanation for the nuclear peace that has prevailed so far since 1945. I explain the basic elements of game theory and dive deeper into the concepts that are often used in game-theoretic applications of nuclear deterrence. I also present the two-player extensive-form games developed by Zagare (1992) and Kraig (1999) for the purpose of modelling deterrence. I discuss the applicability of models such as these, and whether in general game-theoretic models can be used to accurately explain nuclear deterrence or not. I find that under certain assumptions, game-theoretic models can explain nuclear deterrence. However, criticism of the application of game theory to nuclear deterrence generally revolves around the assumptions that must be made for this theory to apply here.

Keywords extensive-form games, rationality, threats, credibility, capability

Table of Contents

1 Introduction

It is commonly thought that nuclear peace has been maintained since 1945 thanks to nuclear deterrence. Nuclear deterrence is often used as justification for the ownership of weapons of mass destruction, namely, nuclear weapons such as atomic bombs. This justification has been used by academics and politicians alike (see e.g. Slocombe, 2000; Waltz, 1981) and it is often thought that for a nation under threat to have a credible national defense, it should have access to nuclear weapons (Powell, 2003). Advocates for nuclear proliferation, i.e., the spread of nuclear weapons to nations not recognized as "Nuclear Weapon States" by the United Nations Treaty on the Non-Proliferation of Nuclear Weapons, claim that the existence of nuclear weapons in a nation's arsenal deters potential adversaries from attacking even on a smaller scale using conventional methods of warfare (Waltz K., 1990). In other words, owning nuclear weapons is said to be an effective way of maintaining peace and preventing war.

This literature review attempts to answer the following research question: Can game theory be used to explain nuclear deterrence and the absence of nuclear war? The interactions between nuclear states with possible nuclear tensions are strategic by nature, which is why nuclear deterrence contains the essential characteristics of game theory. These strategic interactions have been analyzed from a game-theoretical perspective for decades, with help from theories such as those introduced in Thomas Schelling's important book on game theory, *The Strategy of Conflict* (1960). The game-theoretical concepts explained in his work such as threats, deterrence, credibility, and rationality can provide the skeleton for analyzing nuclear deterrence. In this thesis I go through the game-theoretical characteristics related to the posed research question and investigate whether models of game theory can help explain why nuclear weapons have not been used in attack since the second world war.

Ever since the atomic bomb was first used in military conflict in August 1945, when the United States detonated nuclear weapons over the Japanese cites of Hiroshima and Nagasaki within three days of each other, nuclear weapons have been a sensitive but important topic of political discussion. Though weapons of mass destruction have only been used twice in combat during the history of mankind (Hakim, 2002), the

number of nations with access to weapons of mass destruction has multiplied since then (Kraig, 1999). Nuclear weapons are of economic significance as well, due to the vast amounts of investment used to uphold nuclear defense. According to a report by ICAN (2021), $72,6 billion was spent on nuclear weapons in 2020 worldwide, an inflation adjusted increase of $1,4 billion from 2019. Of this sum, around 52% ($37,4 billion) was spent by the United States, accounting for approximately 3,4% of its GDP. The other four largest spenders on nuclear weaponry were China ($10,1 billion), Russia ($8 billion), the United Kingdom ($6,2 billion), and France ($5,7 billion). In addition to these nations, India, Israel[1], Pakistan and North Korea all invested in nuclear weapons in 2020, accounting for approximately 7,2% of global spending. It is important to ask whether nuclear weapons are worth the amount currently spent on them, or whether the money would be better spent elsewhere. Two-sided nuclear warfare has never occurred, and nations have remained content with using conventional methods of warfare without resorting to nuclear weapons even though the technological capacity for a nuclear war between two nuclear powers exists, and a significant figure is spent on developing these weapons annually. Considering that, other than for testing and demonstrative purposes, no nuclear weapons have been used since 1945, why is it that they have still been invested in and developed? This thesis investigates whether game theory has a role in justifying this.

The game-theoretical analysis that can and has been performed on nuclear deterrence is also what makes the topic relevant to economics. Although much of existing literature on nuclear deterrence falls into the field of political science, ethics or international relations, the strategic interactions that are inherent in nuclear deterrence are strongly game-theoretical, and thus economically relevant. Arguably, game theory could provide the tools needed to guide politicians and leaders of an increasingly tense international political landscape towards better decisions. The consequences of a nuclear war are economic as well as humanitarian, another reason why studying this topic through an economist's lens is important.

[1] Israel maintains a policy of deliberate ambiguity regarding their ownership of nuclear weapons (see e.g. Cochran, 1996), meaning that it has not been confirmed by the state itself. It is, however, widely thought that the state has developed them.

The thesis continues as follows. In Section 2, the key game-theoretical concepts related to nuclear deterrence, such as threats and credibility, brought to light by Schelling (1960), are explained. Section 3 introduces models of nuclear deterrence that were developed by Zagare (1992) and Kraig (1999). Section 4 uses these models and existing literature on the topic to discuss the applicability of game theory in nuclear deterrence and the challenges of nuclear deterrence research. Section 5 concludes.

2 Key Game-Theoretical Concepts of Nuclear Deterrence

In this section, I will first introduce the basics of game theory using teachings of, for example, Varian (2020), Osborne (2004). Then, I will discuss the elements of game theory that are key in analyzing nuclear deterrence using the work of Schelling (1960) among other literature. An introduction to the basic elements of the game theory helps set the stage for more in-depth discussion related to nuclear deterrence in later sections of this thesis. Although the concepts explained in the latter part do not necessarily use examples directly connected to the general topic of this thesis, the concepts will all be used later in connection to the application of game theory in nuclear deterrence. For example, they will be of use when discussing extensive-form models of deterrence.

Game theory is used to analyze strategic interactions between economic agents and is relevant to a variety of fields in addition to economics and business (Varian, 2020). In business, common applications of game theory are, for example, firms competing for customers or bidders in auctions (Osborne, 2004). In game theory, the models used to interpret these strategic interactions are known as games, and the decision-makers participating in these games are called players, each of whom are looking to maximize their own expected payoff whilst considering the optimal actions, or strategies, of other players. Each player has preferences regarding the possible outcomes of a game, which are decided by the different combinations of actions taken by players. These preferences are visible in the payoffs that each player obtains from different outcomes, meaning that a preferable outcome has a higher payoff than a less preferable one. Games model the effects of interaction between multiple decision-makers by allowing the preferences of players and thus payoffs at different outcomes to be affected by the actions of other players (Osborne, 2004). The outcome of a game, where no player

would wish to change his strategy given the strategies of the other players is called Nash equilibrium, which will be described in Section 2.2 in more detail.

Typically, games are modelled in two different ways. They are either in normal form and illustrated using a payoff matrix, or in extensive form and illustrated as a game tree. Although both forms are used in this thesis, extensive-form games are more prominent in the analysis of nuclear deterrence. This is because they allow for a more intuitive graphical representation of games where choices are sequential, whereas normal-form games are of more use in examples where choices by players are made simultaneously. The outcomes of extensive-form games are typically solved using backwards induction, by starting at the final decisions of the game. Both normal-form and extensive-form games allow for more than two players, but in this thesis, focus is maintained strictly on two-player games, as real-life tensions between nuclear states are generally between two nations at a time.

2.1 Rationality

Equilibria in models of game theory are dependent on the rationality of the players involved, but what exactly is rationality and how can it be defined? Interpretations of rationality in game theory differ. In relation to rationality in general game theory Osborne (2004) describes the theory of rational choice as players choosing the best action out of all available actions according to their preferences. That is, when facing a choice between n actions, a player will be assumed to choose the action that has the highest expected payoff, which is determined by his preferences. This means that no stance is taken on whether the preferences themselves are rational, as long as the players are acting according to those preferences.

Some academics believe that rationality implies all-knowingness (Zagare, 1990). In game theory, this means that the players are completely knowing of not only their own possible actions and payoffs, but also those of the other players. For example, Verba (1961) describes a rational actor as someone who can make a cool-headed decision based on calculations about every possible action and outcome. In a case of two nations undergoing a game of nuclear deterrence, this is perhaps not realistic to assume, as no nation can know the precise preferences and payoffs of another. This type of rationality

is described by Zagare (1990) as procedural, and it is a common assumption in game-theoretic models, regardless of its real-life applicability.

A more realistic assumption is that players are not all-knowing. A theory of rationality that conforms to this view is instrumental rationality, which is defined by Luce and Raiffa (1957) as simply choosing the action which yields the preferred outcome when confronted with two options. All-knowingness is not implied, so this would be a better fit to the topic of nuclear deterrence. Many models of nuclear deterrence, including those discussed in this thesis make the simplifying assumption of perfect information, which leads to a situation of procedural rationality. Despite it not necessarily being the most realistic of the two aforementioned definitions, it does help to draw conclusions and focus on some of the more relevant aspects of these models.

2.2 Nash Equilibrium

One of the most important concepts of game theory is Nash equilibrium. Nash equilibrium is the outcome of a game where each player is using a strategy that maximizes his own expected payoff, given the strategies of other players (Nash, 1950). That is, no player will wish to change their own strategy given the actions chosen by the other players. This means that in a two-player game, a pair of strategies leads to a Nash equilibrium if the choices of both players are optimal given the other player's choices (Varian, 2020). As stated earlier, the objective of this literature review is to discuss whether nuclear deterrence and the absence of nuclear war can be explained using game theory. Thus, in a game-theoretical model of nuclear deterrence, for mutual deterrence and peace to hold true, equilibrium should be an outcome in which neither nation decides to attack the other. This would be simple if the nations were willing to cooperate, as rational nations could simply agree not to escalate with nuclear weapons or go to war of any kind. However, the topic of interest in nuclear deterrence is nations that are not necessarily immediately willing to cooperate. After all, there would be no need to deter if the other player is in no way threatening. Non-cooperative game theory, i.e., game theory in which players cannot or will not cooperate is fitting when it comes to nuclear deterrence (Zagare, 1990). With imperfect communication, reaching stable mutual deterrence is less straightforward than in cases where players are willing and able to communicate with each other.

The prisoner's dilemma, first introduced by Albert W. Tucker in 1950 (Poundstone, 1993) is a classic example of a non-cooperative game in which the Nash equilibrium is Pareto inefficient, meaning that there exists another outcome in which both players' payoffs can be increased without decreasing that of the other player. The dilemma is illustrated in Figure 1 using an example by Plous (1993) of an arms race between the Soviet Union and the USA. This example is a one-shot game, i.e., a game which is played once and never repeated.

		USSR: Disarm	USSR: Arm
USA	Disarm	3, 3	1, 4
USA	Arm	4, 1	2, 2

Figure 1. The Prisoner's dilemma of an arms race between the USSR and USA (Plous 1993). Equilibrium is at (Arm, Arm) even though (Disarm, Disarm) is the Pareto efficient outcome.

The Nash equilibrium is at (Arm, Arm), but there is clearly a Pareto improvement to be made by moving to (Disarm, Disarm) where payoffs are higher for both. With no communication, the nations must individually decide what action to choose, leading to an outcome where both nations have nuclear arms in a world where they are slightly worse off than in the Pareto efficient outcome. If the players were able to communicate, they could try and change the outcome. Say USSR told USA that it would disarm. Should the nations agree to disarm, the outcome would be coordinated to the Pareto efficient outcome of (Disarm, Disarm). However, in a non-repeated prisoner's dilemma, or even a finitely repeated one, economic theory states that cooperation, or disarmament in this case, is never a Nash equilibrium. There is a consensus that the Nash equilibrium of a finitely repeated game, i.e., a game repeated more than once but less than infinitely many times, is to play the Nash equilibrium of the one-shot game every single time (Benoit and Krishna 1985) even if communication were possible.

2.3 Threats and Credibility

The USSR and USA struggled to agree on nuclear disarmament, and the Cold War witnessed an arms race between the two so-called superpowers. One explanation for this could be that cooperation in a finite game is irrational for both. If USA knew that the USSR was going to disarm, then USA would arm itself to gain the better payoff at (Arm, Disarm). This would be the case in Figure 1 even if communication was possible. The USSR could threaten to retaliate if USA 'cheats' by deviating from the Pareto efficient choice. In a one-shot game that is never repeated and ends after choices have been made, this threat would have to be based on something outside of the game. A player cannot threaten to hurt the other's payoff in the next game if the game is not going to be repeated. In such a case the Soviet Union could threaten to invade one of the USA's allies after the game, for example. Providing that this has no effect on the payoffs of the one-shot game, this threat would be meaningless to the USA because it has no bearing on the game. Cooperation becomes more likely the more a game is repeated. The higher the probability is that the game will continue, the more likely cooperation becomes (Dal Bó & R. Fréchette, 2018). This means that a threat is most likely to lead to cooperation in an infinitely repeated game. Threats can also be effective in sequential games, games where choices are not made simultaneously, but sequentially, one after the other. Although these games are not necessarily repeated, sequential decisions mean that threats can still have an effect on what the other player chooses at an earlier stage of the game.

For a threat to succeed, it must be credible enough for the threatened player to believe that it will indeed be carried out. The more likely it is that a player will go through with a threat, the more likely it is that the threat will work and not have to be carried out. In game theory, the purpose of a threat is not to serve as revenge if it fails, it is there to deter the other player from choosing an action that is undesirable to the player making the threat (Schelling, 1960). Another effect of threats is that they may inflict damage to payoffs of both the threatener, and the threatened player. In such a case, the costs that will occur if the threat needs to be fulfilled must also be visible to the other player for it to be credible. However, if the player making the threat will also suffer from the threat being carried out, then the player being threatened is unlikely to believe the threat because it would be irrational to carry it

out. The threatener thus lacks credibility. For instance, in nuclear deterrence, one nation threatens to retaliate with nuclear weapons if it is attacked. If the player who makes the initial attack also possesses nuclear weapons, then the outcome is likely to be nuclear war. If both nations know that this is will be the outcome, then the threatener has a credibility issue because it would be irrational to be willing to risk complete destruction just to fend off a small-scale conventional attack, unless the threatener has perverse preferences.

Another concept closely related to credibility is reputation. A player may be able to increase the credibility of his threat by putting his reputation with the other player on the line (Myerson, 2009). For example, a business may want to maintain a reputation for lowering its prices whenever a competitor enters the market so that competitors are deterred from doing so. In the case of nuclear deterrence, a nation can make its deterrent strategy significantly more credible by having a reputation for acting according to this strategy. A downside of this is that it then must maintain this reputation, as any deviation from this strategy would significantly reduce its credibility in the future. Thinking of real-life cases, the United States are so far the only nation that may have a reputation for escalating with nuclear weapons, as they are the only nation to have used them to attack.

The final requirement of a credible threat is that the threatener must be able to go through with the threat. In other words, the threat must be capable (Boulding, 1963). In nuclear deterrence, a capable nuclear threat implies an ability to carry out nuclear strikes among other things. The equilibria of games of nuclear deterrence depends significantly on the symmetry of the capabilities of participant nations. This means that the outcomes vary from cases where both have capable nuclear threats, to those where one or neither of the players have them.

2.4 Commitment

A closely linked concept of game theory related to threats and credibility is commitment. Commitment can be a way of ensuring the credibility of, for example, a threat such as nuclear retaliation. Committing to going through with a threat can take some of the decision-making power of away from the threatener or make it less costly

for him to retaliate as he has made some of the costs of retaliation sunk. This is what makes the threat more credible to the player being threatened. Commitment must be both observable to the other player and irreversible for it to be credible (Varian, 2020). Danilovic (2001) describes the two main methods of commitment as sinking costs and tying hands. The sunk-cost method involves undergoing costly actions before the game begins, in order to manipulate the payoffs of the game. For example, in nuclear deterrence and warfare in general, setting up a missile defense system or mobilizing troops can reduce the costs associated with warfare in the game.

Fearon (1997) finds that tying hands is, on average, a more effective method of commitment than sinking costs. This may be because of the irreversibility involved in tying hands, which, as stated above, is key to effective commitment. Hand-tying reduces the options of the threatener before any decisions have been made. By tying his hands, the threatener can commit to a certain action and increase the costs that would be associated with detracting from the commitment. Fearon (1994) talks about using domestic audience costs as a method of tying hands to a commitment, and this could be applicable to nuclear deterrence as well. He argues that audience costs may even be the most important method for a nation in crisis negotiations. States can tie their hands using audience costs by, for example, producing a public statement of intent by their state leader (Fearon 1997). By doing this, a state puts its reputation on the line and risks a huge weakening of its leadership should it not carry out the threat that it has publicly stated it would carry out. For example, in cases such as the Cuban Missile Crisis of 1963, the leaders of the Soviet Union and United States could have been considered weak if they had backed down. By tying hands, the nation also provides a signal to its adversary, who now sees the threat as more credible because the threatener has committed to it publicly.

These methods of commitment are highly relevant to single events, but in reality, events are interdependent. This means that actions in one event or game may have an effect on reputation in another. Due to the interdependence of events, it is important to maintain a reputation for seeing through commitments. USA seems to have acknowledged this, as during the cold war it made sure that it had a reputation of fulfilling its commitments, sometimes to the extent that it performed actions even though they may not have been wise context-specifically (Danilovic, 2001). The United

States are so far the only country to have ever attacked another with an atomic bomb, meaning that it is the only country that could base its nuclear threat credibility on a reputation through historical actions, as mentioned in the previous section.

2.5 A Noncredible Threat

An example of a noncredible threat is shown by an extensive-form game of Gibbons (1997) in Figure 2[2] with perfect and complete information. The game has been adapted to a simple model of deterrence between India and Pakistan. In this example, both nations have capable nuclear threats, and are in a politically tense scenario, possibly on the brink of war. Both players are looking to maximize their own payoffs. India is the first to make its choice of attacking or cooperating. If India cooperates, Pakistan will attack, and India will lose the war. If India attacks first, Pakistan will face the same choice of attacking or cooperating. If it attacks, the war will escalate, and nuclear war will begin, leading to the worst outcome for both (-1, -1). If Pakistan refrains from attacking, India will win the war and payoffs will be (2, 0).

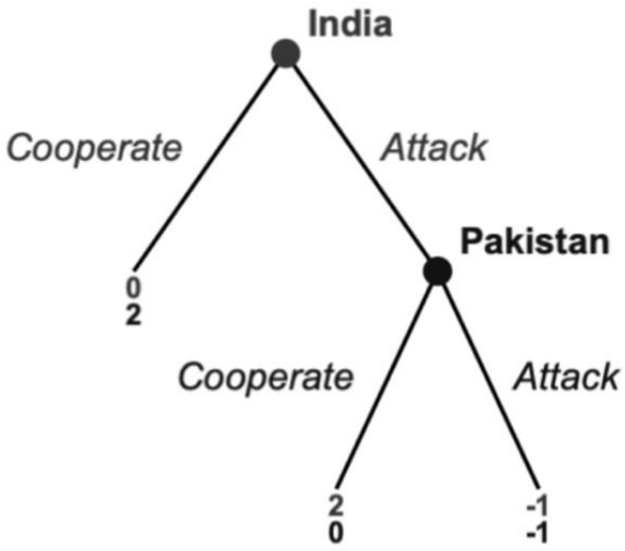

Figure 2. An extensive-form game with two players (Gibbons, 1997). The Nash equilibrium under the given payoffs is (Attack, Cooperate).

Using backward induction, the method for solving extensive-form games, we can see that the Nash Equilibrium is for Pakistan to choose not to attack, and therefore, India

[2] The games used in thesis, namely those of Gibbons (1997), Zagare (1992) and Kraig (1999), have been slightly modified for this thesis. They are not visually identical to those seen in their original papers.

will choose to attack, leading to an outcome of (Attack, Cooperate) with payoffs of (2, 0). Pakistan could threaten to choose to attack if India chooses to do so, to coerce India into not attacking in the beginning. However, given the payoffs, the threat is noncredible when information is perfect and complete, and both players are aware of each other's payoffs. Unless the payoffs change in favour of the threat, Pakistan would have to be irrational for the threat to be of any substance to India.

The threat could be made credible by incorporating one of the two methods introduced earlier. These can be used to change the payoff structure of the extensive game model in Figure 1. If Pakistan ties its hands, say, by means of domestic audience costs, it may reduce its payoff at (Attack, Cooperate) so that it becomes smaller than its payoff (Attack, Attack) and is no longer the rational choice. By tying its hands, it may also remove this choice completely from the game. If Pakistan were to use sunk costs to make its threat more credible, it would have to do so before the game begins. This could be used to decrease the payoff at (Attack, Attack). Another method would be for Pakistan to somehow make itself seem irrational to India. By doing so, it would make Pakistan's decision seem uncertain to India, leaving it more to chance than before. Leaving a threat to chance can be an effective way for a player to coerce the other player into a more favourable decision for himself (Schelling, 1960). If nuclear war was not anymore totally in Pakistan's hands but could occur at some probability due to Pakistan's irrationality, India may choose not to take this chance.

3 Models of Nuclear Deterrence

In this section, models of deterrence developed by Zagare (1992) and Kraig (1999) are introduced. I introduce a model of deterrence where nations do not have nuclear weapons and a model of deterrence where the possibility of nuclear escalation exists. By doing so, it is possible to compare the conditions required for mutual deterrence when nuclear weapons exist to the conditions required when they do not. These models help give a deeper insight into what circumstances may be needed to apply game-theoretical models successfully and realistically to nuclear deterrence. The models introduced in this section show that under certain assumptions, mutual nuclear deterrence can indeed exist. However, opposition to nuclear deterrence and game-theoretical models of deterrence generally stems from doubts about the

assumptions the models make, rather than the models themselves. The discussion related to the applicability of these models will be discussed in Section 4.

3.1 A Basic Model of Deterrence

First, we look at a basic deterrence game in extensive form (Figure 3) with sequential choices and perfect and complete information. There are two players, the USA and Russia (RU), and in this game, Russia makes the first choice, although the choice of who acts first is arbitrary. Players make their choices at the decision nodes labelled with the abbreviation of the nation whose turn it is. This game does not yet contain the option of nuclear escalation for either player. The game begins with Russia choosing to cooperate (C) or defect (D). Cooperation means maintaining peace, whilst defecting means initiating conflict with conventional warfare methods. If Russia chooses to cooperate, USA is faced with the same choice that Russia had in the first node. If USA chooses to cooperate, peace is maintained at the outcome of CC. By choosing to defect, Russia is faced with another choice of whether to cooperate, which would mean losing in this scenario (CD), or defect, leading to a conventional two-sided war (DD). Should Russia choose to defect at the very start, USA is faced with the choice of cooperating or defecting, leading to outcomes of DC or DD.

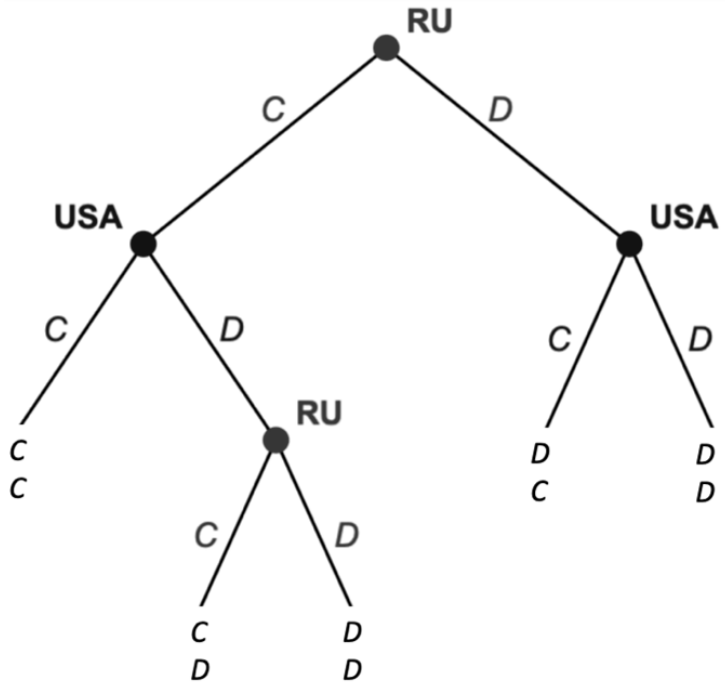

Figure 3. A basic deterrence game (Zagare, 1992; Kraig, 1999). RU makes the first choice and can either cooperate (C), or defect (D). Defecting means attacking with methods of conventional warfare.

Figure 3 does not yet provide the information required for an equilibrium to exist, as there are no payoffs. Nothing has been stated about the preferred outcomes of either player, so we cannot yet solve for equilibrium. Zagare (1992) and Kraig (1999) postulate that there are three conditions on player preferences that must be met for an equilibrium to be found, and for that equilibrium to be mutual deterrence. First, if one player defects, the non-defecting player's payoff must be smaller than at peace. Thus, for Russia, CC > CD and for USA, CC > DC. This suggests that players would rather be at peace than be attacked and not respond. Second, for the theory of deterrence to be relevant in this model, one or both players must have an incentive to defect. If this was not true, there would be no need to deter as there would be no threat from the other player. This means that Russia would want to defect if USA doesn't, and vice versa. So, for Russia, DC > CC and for USA, CD > CC. This is perhaps less intuitive than the first condition, but it is necessary for the model to provide the outcome of peace. It is also worth mentioning that even if only one player wishes to upset peace, the equilibrium will remain the same. Third, the payoffs in conflict (DD) must be greater than in the scenario where one player defects and the other cooperates, so for Russia, DD > CD and USA, DD > DC. This means that players would rather go to war than surrender to an attack from the other player, and thus signals that the threat of mutual conflict is credible and capable, as players are able and willing to inflict damage on the other in conflict. Finally, the model assumes that for both players peace is better than war, so CC > DD for both. As a result of these conditions, the following payoff conditions are obtained.

$$Russia: DC > CC > DD > CD$$

$$USA: CD > CC > DD > DC$$

Having deduced these preferences, it is possible to create a game with payoffs to find out the equilibrium (Zagare, 1992) and further demonstrate that the outcome, under the conditions stated above, is mutual deterrence (Figure 4). In this model, the top numbers of the outcomes indicate Russia's payoff, and the bottom numbers indicate USA's payoffs. Both players are looking to maximize their own payoffs.

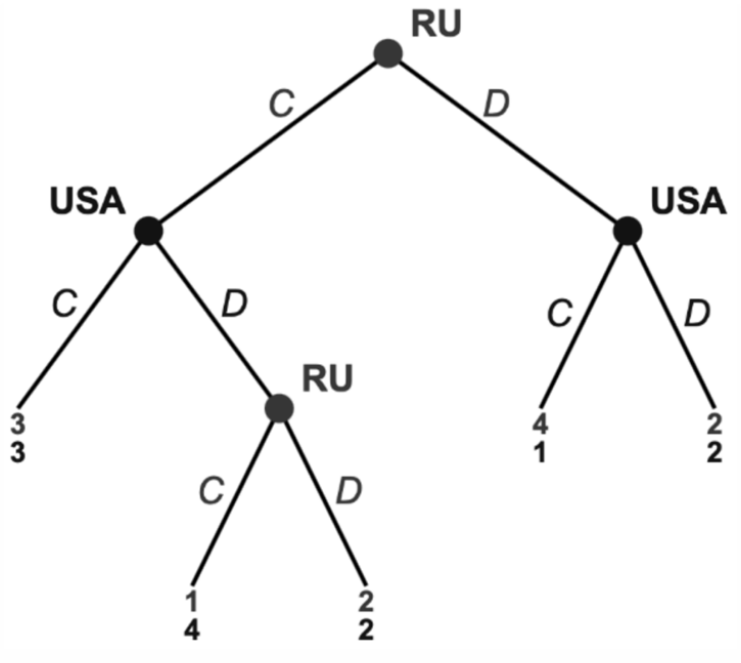

Figure 4. The basic deterrence game with payoffs based on the preferences explained by Zagare (1992) and Kraig (1999). Nash equilibrium is at (C, C). The top numbers of the outcomes indicate Russia's payoff, and the numbers below indicate USA's payoffs.

The equilibrium of the game, given the payoffs and solved using backward induction, is peace (CC). If Russia chooses D at the first node, USA will also choose D and the outcome will be (2, 2), because it is preferred by USA (2 > 1). When Russia chooses to cooperate, USA will choose to cooperate as well, because defecting would lead to the outcome of (2, 2) due to Russia's preferred choice. The theory of stable mutual deterrence holds in this scenario under the assumptions that both threats are credible and capable (Zagare, 1987). Capability is assumed to be fixed in conventional deterrence, as most nations have capable conventional threats. However, credibility is not, and equilibrium changes when both players' threats are not assumed to be credible. When only one player has a credible threat, the Nash equilibrium is for the player with a credible threat to gain the advantage. For example, the payoffs in Figure 4 change in a way that leads to an equilibrium of DC if only Russia's threat is credible and CD if only USA's is. Zagare (1992) describes this scenario as a 'Called Bluff' game, because the 'bluff' of the player with a noncredible threat is called by the player with a credible one. If neither threat is credible, whoever has the first choice in the extensive-form game will choose to defect, meaning that the first mover gains an advantage. Thus, if Russia moves first as in Figure 4, the outcome is DC, and if USA moves first the outcome is CD. In this game of symmetrical non-credibility, with features akin to

a game of chicken (Schelling, 1960), the payoffs of conflict become lower than losing the war (for Russia, DD < CD and for USA, DD < DC). The consequence of this is that once the player choosing first chooses to defect, the other player is better off surrendering. In this model of basic deterrence, the outcome is peace only when both players have credible and capable threats.

3.2 A Two-Stage Nuclear Model of Deterrence

Now, the option of nuclear escalation is incorporated into the game, hereby turning it into the two-stage deterrence game developed by Zagare (1992) and Kraig (1999), which builds on the game of basic deterrence introduced in the Section 3.1. This nuclear deterrence game in extensive form starts off as the basic deterrence game but includes second-stage nuclear options for both players, with additional branches and payoffs for outcomes where one or both nations escalate with nuclear weapons (ED, DE and EE). In this model, any form of attack involving nuclear weaponry is considered a nuclear escalation, ignoring the fact that in reality, nuclear escalation is not necessarily either performed or not performed, and is multi-staged (see e.g., Kahn 1962, 1965). As in the one-stage model of deterrence, there are conditions regarding player preferences that must be met for nuclear deterrence to be the equilibrium outcome of this model. We start off by assuming deterrence in this game to be symmetric, meaning that the nuclear threats of both players are equally capable and credible. After this, the effects of variation in both conditions are discussed.

Russia is once again assumed to act first and can again either defect or cooperate. The game remains the same as the basic deterrence game until either Russia or USA decide to defect. Then, the other nation has the possibility of escalating using nuclear weapons (E). Once nuclear escalation has been chosen by either player, the other player must take the decision to either retaliate using nuclear weaponry (E) or continue with conventional methods of warfare by defecting (D), meaning that it does not further escalate the conflict. In addition to the outcomes introduced in the previous model (CC, DC, CD and DD), we have three new possible outcomes. If USA chooses to respond by defecting, Russia can escalate with nuclear weapons and choose E. If USA decides not to retaliate, the outcome is ED. If USA chooses to retaliate and choose E,

the outcome is EE, nuclear war. Russia faces the same decisions whenever USA defects for the first time, leading to the outcome of DE if Russia does not escalate, and EE, if it does.

Nuclear escalation (E) is never a first option in this model. One player must have first chosen to defect conventionally first for this to happen. This is a strong assumption, but we will see that when other more realistic assumptions are in place the nonexistence of this option does not alter equilibrium outcomes. It does however bias the game in favour of those arguing for nuclear proliferation (Kraig, 1999).

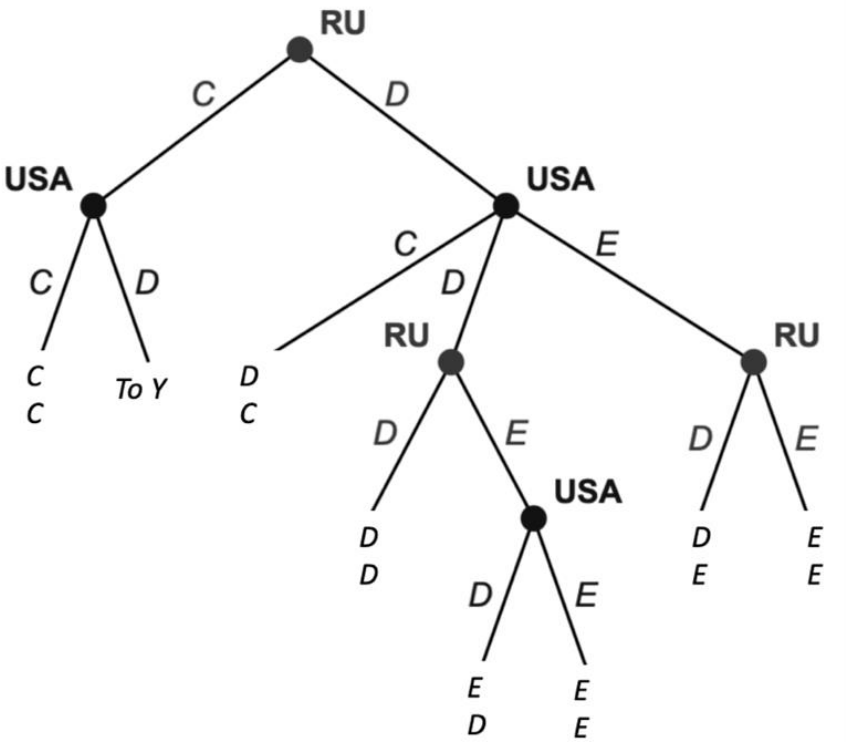

Figure 1. The two-stage deterrence game with nuclear escalation. In the first stage, players can either cooperate (C), or defect (D). In the second stage, nuclear escalation (E) is an option. At 'To Y', the game moves to Russia's choice between D and E to the right.

The basic structure of the two-stage escalation game is now clear, but nothing can be said about equilibria without payoffs, and payoffs are difficult to deduce without assumptions or knowledge of the preferences of players. As in Figure 3, several assumptions are to be made about these preferences (Zagare 1992; Kraig 1999). Based on these preferences it is possible to model an extensive-form game with payoffs, and thus make deductions about the equilibria of a game of nuclear deterrence and compare them to the basic deterrence game. There are three key assumptions. First, a

nation gains more utility from unilateral escalation than from conventional war, or cooperating when the other defects, leading to the following utilities.

$$U_{RU}(ED) > U_{RU}(DD, CD)$$
$$U_{USA}(DE) > U_{USA}(DD, DC)$$

This preference makes nuclear weapons realistic options and means that nations would much rather 'beat' their adversary than settle for a conventional war, let alone concede defeat in the case of the adversary defecting first. This assumption has its drawbacks, as it is not a given that a nation would indeed rather win a war by nuclear escalation than play out a conventional war. Second, the new outcomes introduced in the two-stage model of deterrence lead to payoffs that are always smaller than the payoffs of first-stage outcomes.

$$U_{RU,USA}(EE) < U_{RU,USA}(CC, DD, DC, CD)$$

This makes mutual destruction less desirable than any outcome of the basic deterrence game with conventional warfare and ensures that the significant costs of nuclear war are represented (Kraig, 1999). Lastly, the payoff of conceding defeat or not escalating when the other player escalates is smaller than doing the same at the first stage during conventional warfare.

$$U_{RU}(DE) < U_{RU}(CD)$$
$$U_{USA}(ED) < U_{USA}(DC)$$

This means that players would much rather concede defeat at an earlier stage when less damage has been done. Without the second and third assumptions, nuclear escalation and war would be more viable outcomes within the game. The preferences presented in this last point ensure that nuclear war, and unilateral nuclear attacks are not the outcomes with the highest payoffs in the game.

These preferences lead to the logic of the payoff structure of Zagare's (1992) two-stage escalation game (Figure 5). Based on the method of backwards induction introduced in Section 2, the result is the equilibrium outcome of stable mutual deterrence where both players choose to maintain peace. In Figure 5, the players of the game are Russia (RU) and USA. In addition to the possible actions of cooperation (C) and conventional warfare (D), nuclear escalation (E) is now included. The payoffs at the outcomes of the game tree reflect the preferences introduced above.

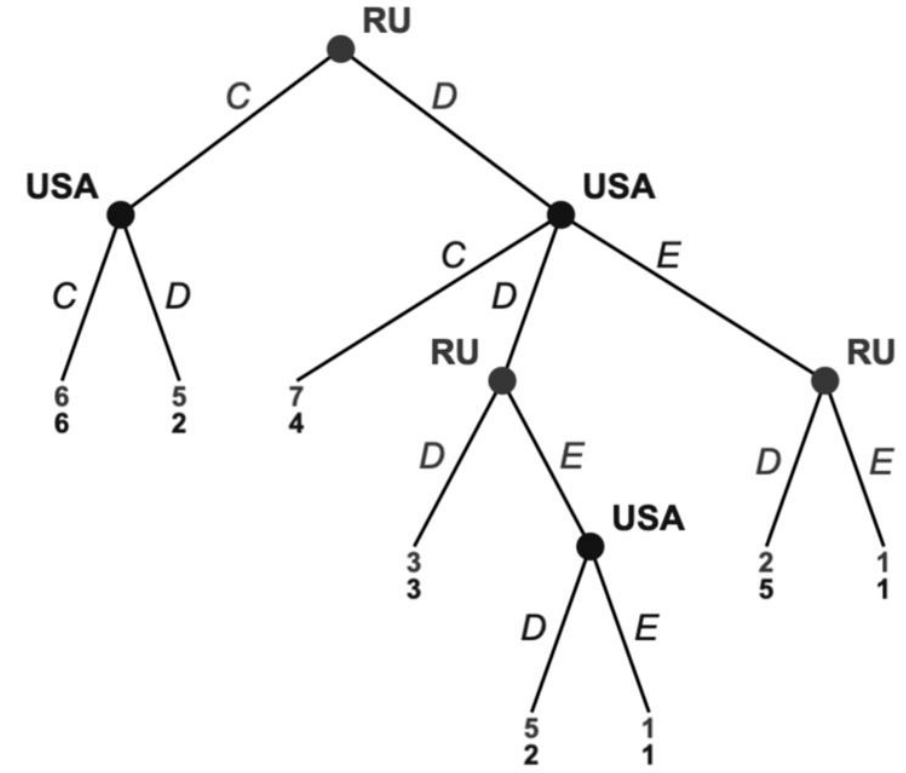

Figure 2. The two-stage deterrence game with payoffs. Nash equilibrium is at (C, C) where payoffs are (6, 6), meaning that peace is maintained.

The outcome of the two-stage model presented in Figure 5 is mutual deterrence and thus peace, the same as in the basic deterrence model. In his paper, Zagare (1992) interprets how sensitive peace is to changes in the credibility of threats in both the first stage (conventional warfare) and the second stage (nuclear escalation). The finding is that peace is maintained in two different scenarios, the first being a situation with credible threats in both first and second stages, and the second being a situation where neither player has a credible second-stage threat. If these conditions are not in place, mutual deterrence does not hold. However, another significant finding is that under perfect and complete information, nuclear escalation is never performed by either

player. Although peace is not always the outcome, variations in credibility and capability at most cause unilateral or bilateral conventional conflicts.

Kraig (1999) uses the same models and makes the same deductions about the outcomes. However, he adds another layer of interest by letting the capabilities of players' nuclear arsenals vary as well. There are, of course, real-life cases where the distribution of nuclear weaponry is not symmetric, which is why this is of interest. Both nations in a conflict do not necessarily possess nuclear weapons, and an example of such a conflict is Ukraine and Russia. Kraig produces asymmetric models of deterrence to illustrate such a situation, and a key finding of these models is that the nation in possession of nuclear weapons defects, and the nation who isn't, cooperates. In other words, according to this model nuclear deterrence does not hold when only one player has a credible and capable threat.

3.3 Comparison of Extensive-Form Models of Deterrence

Kraig (1999) finds that out of the 14 equilibria that are possible when varying the threats and capabilities of the players in a two-stage deterrence game, four lead to an outcome of peace. Similarly, Zagare (1992) finds that when capability is held constant, varying credibility leads to four equilibria of peace out of ten possible outcomes. Both find that when neither player has credible nuclear threats, the outcome is always peace. This leads to the conclusion that nuclear weapons are not required to maintain peace if conventional warfare methods are symmetrically credible and capable. If neither player had nuclear weapons, peace could be maintained without them. When the possibility of nuclear escalation is introduced for either player, both players must then have symmetrically capable and credible threats at both a conventional level and nuclear level for peace to be the outcome. If this is not the case for either player, the outcome is not mutual deterrence as the player with nuclear weapons gains the upper hand. This conclusion backs the argument that nuclear weapons can act as stabilizers of peace in hostile situations, as long as both nations also have credible conventional methods of warfare.

The conclusions of these models indicate that nuclear weapons can stabilize peace in certain situations. In a two-player game, when a nation with hostile intentions has

acquired nuclear weapons, it is necessary for the other nation to acquire them as well or it will face a conventional attack. In this scenario, peace is maintained once both players obtain credible and capable nuclear threats, which is line with nuclear deterrence theory. However, the outcome of the basic deterrence game shows that peace can be maintained also when neither nation acquires nuclear weapons. A key finding is that under perfect and complete information, nuclear war or escalation on either side never occurs according to these models (Kraig, 1999). Only outcomes of peace or conventional attacks are possible. These models back the arguments of nuclear proliferation advocates, as the acquiring of nuclear weapons by an adversary requires the other nation to do so as well. However, the validity of this result is questionable as the assumptions made in the models lie heavily in favour of advocates.

4 Discussion

In this section, I will first discuss whether nuclear deterrence is backed by game theory and the applicability of models such as those introduced in the Section 3. I will also highlight two opposing schools of thought regarding nuclear proliferation, namely those who are for it and those, as mentioned, who oppose it and are subsequently unconvinced about the efficacy of nuclear weapons as a mechanism for maintaining peace. After this, I will discuss the issues with economic analysis of nuclear weapon states and the tensions between them.

4.1 The Applicability of Nuclear Deterrence Models

If possessing nuclear weapons truly is the key to national security and a surefire way of maintaining the status quo of peace, then it would be in the interest of all nations to allow every other sovereign state to acquire nuclear weapons (Kraig, 1999). Yet nations and academics are reluctant for this to happen (see e.g., Sagan 1994; Kaiser 1989; Miller 1993). The models of Zagare (1992) and Kraig (1999) show that when nations have both credible and capable nuclear and conventional threats, the equilibrium is at peace. This could explain why peace has prevailed for so long between nuclear states. However, there is disagreement about whether game-theoretical models are applicable to real-life nuclear tensions, or even should be used to model nuclear deterrence. There are several reasons why academia is not in agreement regarding this.

For example, according to Lebow and Gross Stein (1995), some of the loudest critics of nuclear deterrence are of the opinion that nuclear weapons were a cause of political tensions during the Cold War, not a reason that tensions did not escalate. Lebow (1987) argues that nuclear tensions and armament provoked the kind of tensions that deterrence was meant to prevent, and that peace was maintained not because of deterrence, but despite it.

A hindrance to the applicability of game theory to nuclear tensions is the number of assumptions that must be met for mutual nuclear deterrence to exist (Zagare 1992; Kraig 1999). Information, the symmetry of capability and credibility between the nations, their preferences and their rationality are all key to reaching the outcome of peace in game-theoretical nuclear deterrence models (Schelling, 1960). The circumstances must be just right for nuclear deterrence to exist, but in reality, these circumstances may not always exist. Though he created the multi-stage models introduced in Section 3, even Zagare (1992, p. 452) himself states that the conditions upon which his conclusions lie on are less than comforting. The potentially catastrophic consequences of misjudging these underlying conditions are part of the reason why there is such strong advocacy against nuclear proliferation and skepticism towards nuclear deterrence. For example, some non-proliferation advocates are of the opinion that leaders of nations with access to nuclear weapons do not always possess the emotional stability and cognitive capabilities that are required for rational behaviour and, thus, for nuclear deterrence to work (Sagan, 1994). This means that one of the most important assumptions in game theory – that players are acting completely rationally and accordingly with their payoffs in games of nuclear deterrence – may not be realistic to assume. Based on this, in the future, an irrational leader who decides to go against his preferences, or has preferences that favour nuclear retaliation could upset the status quo and defy what is perhaps the most important underlying assumption of game theory.

As a counterargument to this, it is worth noting that a departure from rationality may not necessarily lead to a breakdown in mutual deterrence. To help a threat succeed and avoid war, a player may wish to appear irrational to the other player (Schelling, 1960). The threat of nuclear retaliation by one nation may seem irrational to the other, as it would most likely lead to mutually assured destruction, as showed in the

extensive-form models of Section 3. A player would have to be insane to want to attack with nuclear weapons. But what if this is indeed the case? By pretending or appearing to be irrational to the other player, the pretending player can make the threat of nuclear retaliation more credible, or even assign an element of randomness to it from the point of view of the other player. The threatened player may be led to think that attacking is not worth the risk, as the irrational adversary could do anything, and its threat might be credible after all. A real life example of this may be the current tensions between North Korea and the United States. The USA seems unwilling to perform any sort of military intervention in North Korea, as they seem irrational enough to retaliate with nuclear missiles even if it costs them their own existence. This could mean that the inaccuracy of the underlying assumption of rationality may not alter the equilibrium of peace after all.

Another problem arises from the assumption that there is no advantage to striking first. Models such as the ones used by Zagare (1992) and Kraig (1999) developed into payoff form do not account for the fact that there may be a significant advantage to striking first. Although this issue is discounted by many models, it is in reality such a huge worry that it might lead to a leader launching a first strike based on the mere suspition that the enemy is contemplating attack (MccGwire, 2006). The advantage of attacking first stems partly from the possibility of the attack being a strategic one targetting a weak point in the adversary. In performing such a first attack, the attacker may be able to reduce the target nation's capability of retaliation, which would lead to the payoffs of, for example the models presented in the previous section, changing depending on which player was first to strike.

4.2 Challenges of Nuclear Deterrence Research

The literature covered in this review dates largely back to the 1990's and earlier. Research on the topic of nuclear deterrence seems to have peaked during and after the Cold War, meaning that there is a slight research gap related to nuclear deterrence in present-day cases such as between North Korea and the United States. Furthermore, most of the literature on the topic discusses nuclear deterrence from the point of view of political sciences, rather than economics. More research connecting game theory and the politics of nuclear tensions should be performed to reach a consensus about

whether game theory is indeed applicable to nuclear deterrence or not. Based on such studies, it would be possible to have more informed political discussions about the spread of nuclear weapons. If game theory does indeed support mutual deterrence, then it may be wise to introduce nuclear weaponry to states under threat. The risks of this are, however, potentially catastrophic.

A key problem with modeling nuclear deterrence is the lack of empirical data available (Mohan, 2019). It is difficult to obtain tangible data from the real life nuclear tensions that could be used to improve the accuracy of model. Even though there are some modern day examples of two nations with nuclear weapons who have tense political relationships, because two-sided nuclear war has never occurred, we do not know the circumstances that would be required for such an event to occur. In addition to this, it is difficult to measure the preferences of policymakers of nuclear states, and thus the payoffs in nuclear deterrence are far from certain, even though they have been constructed in a logical manner.

5 Conclusion

Nuclear warfare has thus far been avoided, even though we cannot be certain about whether it can be explained using game theory alone. The basic concepts for analyzing nuclear deterrence that were introduced in Section 2 help link game theory to nuclear deterrence and provide a basic understanding on how game theory could be used to explain interactions between nations that own nuclear weapons. Models such as the extensive-form games of Zagare (1992) and Kraig (1999) go deeper into the analysis and discuss the possible outcomes of such games and the circumstances and assumptions that must be in place to reach them. They even reach conclusions about the conditions that lead to mutual deterrence. According to these models, deterrence works in a conventional warfare scenario when both players have credible and capable threats. When nuclear escalation is introduced, Nash equilibrium is peace when both players have capable and credible nuclear threats. When either of these conditions is violated, peace is not the outcome, as long as either nation has hostile intentions.

These concepts and models alone are not, however, enough to provide sophisticated analysis on the interactions between nuclear states with political tensions. Numerous

assumptions are required for the outcomes of nuclear deterrence models to be peace, and this is the case in the models of Section 3 as well. They require conditions such as perfect and complete communication, instrumental rationality, and specific preferences to reach the equilibrium of peace, and in real life these cannot always be assumed to hold. In addition to this the application of game theory in nuclear deterrence is weakened by the lack of empirical evidence and the difficulty with quantifying nuclear tensions. For example, preferences and payoffs are extremely difficult to quantify, which is why assuming them to be a certain way reduces the credibility of nuclear deterrence models.

The strategic interactions between nuclear states continue to be studied and researched. The more information is collected, the more conclusions can be drawn. To further study how applicable game theory is in nuclear deterrence, it would be useful to study current conflicts and tensions, such as those between North Korea and USA, or India and Pakistan. Using game theory to investigate the possible implications of nuclear deterrence could be useful in international relations and politics, but policymakers should remain wary of assumptions when predicting the actions of adversary nations. Current literature on the topic tends to be more political than economic, but the studying of nuclear deterrence from an economics perspective may give us a better understanding on how well game theory can be applied to interactions beyond fields of business and economics. Analyzing conflicts using game theory could give us a better understanding of conflicts and maybe even help avoid escalation to nuclear war.

6 References

Literature

Benoit, J.-P., & Krishna, V. (1985). Finitely repeated games. *Econometrica,* 53, 905-922.

Boulding, K. E. (1963). Towards a pure theory of threat systems. *The American Economic Review,* 53(2), 424-434.

Cochran, Edwin S. (1996). Deliberate ambiguity: An analysis of Israel's nuclear strategy. *Journal of Strategic Studies,* 19(3), 321-342.

Dal Bó, P., & R. Fréchette, G. (2018). On the Determinants of Cooperation in Infinitely Repeated Games: A Survey. *Journal of Economic Literature,* 56(1), 60-114.

Danilovic, V. (2001). The sources of threat credibility in extended deterrence. *Journal of Conflict Resolution,* 45(3), 341-369.

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

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

Gibbons, R. (1997). An Introduction to Applicable Game Theory. *The Journal of Economic Perspectives,* 11(1), 127-149.

Hakim, J. (2002). *War, peace, and all that jazz.* New York: Oxford University Press.

Kahn, H. (1962). *Thinking about the unthinkable.* New York: Praeger.

Kahn, H. (1965). *On Escalation.* New York: Praeger.

Kaiser, K. (1989). Non-proliferation and nuclear deterrence. *Survival,* 31(2), 123-136.

Kraig, M. R. (1999). Nuclear Deterrence in the Developing World: A Game-Theoretic Treatment. *Journal of Peace Research,* 36(2), 141-167.

Lebow, R. N. (1987). Conventional vs Nuclear Deterrence: Are the Lessons Transferable. *Journal of Social Issues,* 43(4), 171-191.

Lebow, R. N., & Gross Stein, J. (1995). Deterrence and the Cold War. *Political Science Quarterly,* 110(2), 157-181.

Luce, D. R., & Raiffa, H. (1957). *Games and Decisions: Introduction and Critical Survey.* New York: John Wiley and Sons.

MccGwire, M. (2006). Nuclear Deterrence. *International Affairs,* 82(4), 771-784.

Miller, S. E. (1993). The Case against a Ukrainian Nuclear Deterrent. *Foreign affairs,* 72(3), 67-80.

Mohan, J. (2019). Gaming Nuclear Deterrence. *Harvard International Review,* 40(3), 33-35.

Myerson, R. B. (1999). Nash Equilibrium and the History of Economic Theory. *Journal of Economic Literature*, 1067-1082.

Myerson, R. B. (2009). Learning from Schelling's Strategy of Conflict. *Journal of Economic Literature,* 47(4), 1109-1125.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences,* 36(1), 48-49.

O'Neill, B. (1994). Game Theory Models of Peace and War. In R. Aumann, & S. Hart, *Handbook of Game Theory with Economic Applications* (Vol. 2, pp. 995-1053). New York: Elsevier.

Osborne, M. J. (2004). *An introduction to game theory* (Vol. 3). New York: Oxford University Press.

Plous, S. (1993). The Nuclear Arms Race: Prisoner's Dilemma or Perceptual Dilemma? *Journal of Peace Research,* 30(2), 163-179.

Poundstone, W. (1993). *Prisoner's Dilemma.* New York: Anchor.

Powell, R. (1985). The Theoretical Foundations of Strategic Nuclear Deterrence. *Political Science Quarterly,* 100(1), 75-96.

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

Sagan, S. D. (1994). The Perils of Proliferation: Organization Theory, Deterrence Theory, and the Spread of Nuclear Weapons. *International Security,* 53(2), 66-107.

Schelling, T. C. (1960). *The Strategy of Conflict.* Cambridge: Harvard University Press.

Slocombe, W. B. (2000). The Administration's Approach. *Washington Quarterly,* 23(3), 77-85

Varian, H. R. (2020). *Intermediate Microeconomics.* New York: W. W. Norton & Company.

Verba, S. (1961). Assumptions of Rationality and Non-rationality in Models of the International System. *World Politics,* 14(1), 93-117.

Waltz, K. N. (1981). *The spread of nuclear weapons: More may be better*. London: Taylor & Francis.

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

Zagare, F. C. (1987). *The dynamics of deterrence*. Chicago: University of Chicago Press.

Zagare, F. C. (1990). Rationality and Deterrence. *World Politics,* 42(2), 238-260.

Zagare, F. C. (1992). NATO, Rational Escalation and Flexible Response. *Journal of Peace Research,* 29(4), 435-454.

Other References

International Campaign to Abolish Nuclear Weapons (2021). *Complicit: 2020 Global Nuclear Weapons Spending*. Genève: ICAN. https://www.icanw.org/2020_global_nuclear_weapons_spending_complicit

United Nations. (1970, May 11th). Treaty on the Non-Proliferation of Nuclear Weapons (NPT). https://www.un.org/disarmament/wmd/nuclear/npt/text