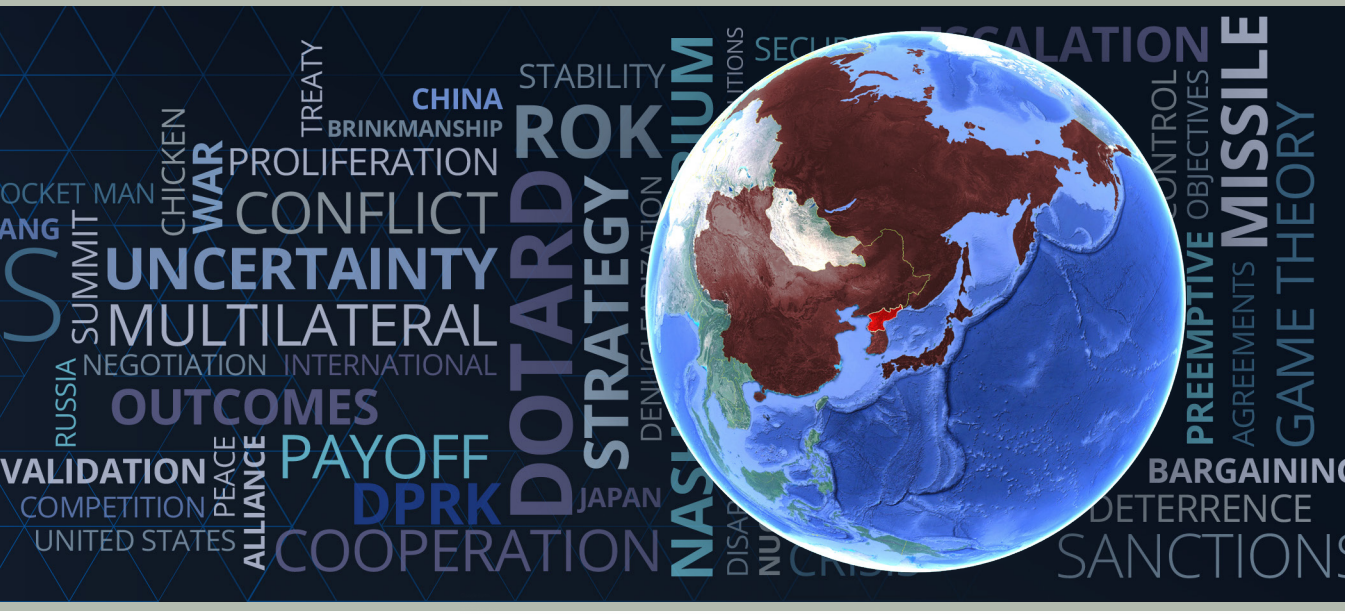
GAME THEORY AND NUCLEAR STABILITY IN NORTHEAST ASIA

Workshop Proceedings



Lauren Ice | James Scouras | Kelly Rooker | Robert Leonhard | David McGarvey

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

GAME THEORY AND NUCLEAR STABILITY IN NORTHEAST ASIA

Lauren Ice

James Scouras

Kelly Rooker

Robert Leonhard

David McGarvey

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

The figure on page 37 is reprinted by permission of SAGE Publications, Ltd. from Kathleen M. Carley, Geoffrey P. Morgan, and Michael J. Lanham, "Deterring the Development and Use of Nuclear Weapons: A Multi-Level Modeling Approach," *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 14, no. 1 (2017): 95–105, © 2017 SAGE Publications, Ltd.

The figure on page 38 is reprinted by permission of IEEE from Geoffrey P. Morgan, Michael J. Lanham, William Frankenstein, and Kathleen M. Carley, "Sociocultural Models of Nuclear Deterrence," *IEEE Transactions on Computational Social Systems* 4, no. 3 (2017): 121–134, © 2017 IEEE.

NSAD-R-19-026

CONTENTS

1

INTRODUCTION

Over the past half-decade, the Johns Hopkins University Applied Physics Laboratory (JHU/APL) has developed a broad portfolio of internally funded deterrence research. One important dimension of this research is multilateral nuclear stability. The focus on this topic is motivated by the belief that multilateral nuclear stability is far less understood than its bilateral counterpart, the dominant Cold War nuclear concern, and will not be the primary emerging nuclear challenge over the coming decades.

While bilateral nuclear stability remains relevant in today's international security environment, increased analytic attention to multilateral nuclear stability is needed. Since the end of the Cold War, India, Pakistan, and North Korea have become nuclear states, and more may follow. These nuclear states are involved in both bilateral competitions, such as between India and Pakistan and between North Korea and the United States, and multilateral competitions, such as among India, Pakistan, and China and among North Korea, South Korea, China, Russia, Japan, and the United States. Thus, purely bilateral stability analyses of these competitions are inherently incomplete.

Finally, even though the world cannot yet be considered from a purely multilateral perspective, multilateral nuclear stability analysis is innately hard and will take years, if not decades, to mature. Thus, it is prudent to begin such analyses now, rather than wait until the world is truly multilateral.

Prior JHU/APL research in this area culminated in a 2014–2015 State Department–sponsored project that assessed the contributions and limitations of game theory as applied to Cold War bilateral nuclear stability analysis; identified the principal additional challenges in extending this work to multilateral stability analysis; and drew some tentative conclusions about the utility of game theory to shed light on multilateral nuclear stability. This work is documented in an annotated briefing, "Multilateral Nuclear Stability: Potential Contributions from Game Theory,"[1] available by request from JHU/APL. A report documenting this work is being prepared, with an anticipated publication date of 2019. This report will be available on JHU/APL's public website. The conclusions of this work are briefly summarized here.

The primary conclusion from this prior body of research is that game theory is a potentially useful analytic framework for raising and thinking through multilateral nuclear stability. Interesting insights can be derived from simple multilateral games. It is important to note that it is not always necessary to "solve" the game; formulating it can provide significant value.

However, if game theory is to make significant contributions to understanding multilateral nuclear stability, analysts need to learn appropriate lessons from the large body of Cold War bilateral game theory analysis. First, game theorists need to work closely with policy analysts, and second, uncertainties must be recognized and addressed. Additionally, game theory needs further development to address important aspects of multilateral stability, including in the areas of sequential games, coalition games, overlapping games, and validation of games.

The first of these lessons is the primary motivation for this project. A workshop was designed to bring together nuclear policy and strategy analysts, international relations experts, and game theorists as a first step in encouraging collaborations among the disciplines. Nuclear strategists and regional experts could gain greater insights into the utility and limitations of game theory, while game theorists could gain a greater understanding of how to make their work relevant to policy formulation.

The workshop focused on the North Korean nuclear crisis. This crisis presents a clear challenge to multilateral stability, directly involving at least six countries: North and South Korea, the United States, China, Japan, and Russia. More important, the North Korean nuclear crisis is high on the agenda of international concerns regarding nuclear stability, so there is urgency in addressing this crisis and evaluating alternative policies to resolve it.

[1] T. Smyth, J. Scouras, and D. McGarvey, "Multilateral Nuclear Stability: Potential Contributions from Game Theory," unpublished briefing, April 15, 2015, PowerPoint file.

Structure

The objectives of the workshop were to (1) assess the utility of game theory in providing insight into multilateral nuclear stability, focusing on the North Korean nuclear crisis; (2) understand the types of problems game theory can be most useful in helping to solve; and (3) address communication and motivational disconnects between the policy and game theory communities. Additionally, the workshop explored how these communities can work together to produce more policy-relevant games and collaborate on future work.

Participants

Approximately one-third of the participants were game theorists with a published record of applying game theoretic methods to international relations and nuclear issues. This group was augmented by several additional academics with diverse backgrounds who had also published articles on applying mathematical methods to understanding nuclear stability or the North Korean nuclear crisis.

Most of the other workshop participants were international relations experts, nuclear policy and strategy analysts, and representatives of the sponsoring government agencies. Included in this group were experts on the countries most closely involved with the North Korean nuclear crisis (the Democratic People's Republic of Korea, the Republic of Korea, the United States, and Russia). Unfortunately, the invited regional experts from China and Japan were not able to attend.

Overall, these participants brought a wide variety of backgrounds and research interests to the workshop. The broad spectrum of perspectives brought to light some of the divides and disconnects among the communities, but also yielded insight into potential avenues of collaboration on future work.

Agenda

The workshop was conducted on August 15–16, 2018, at JHU/APL in Laurel, Maryland. It included (1) an introduction to game theory and the North Korean nuclear crisis, as well as a review of the relevant literature; (2) presentations on national objectives and game theoretic methods applied (or potentially applicable) to the North Korean nuclear crisis; and (3) a discussion of the utility of game theory. The full agenda is on the following page.

August 15, 2018

Game Theory Tutorial—Jun Zhuang

The North Korean Nuclear Crisis—Robert Leonhard

The Objectives and Perspectives of the United States—Elaine Bunn

North Korea and Nuclear Weapons—Stakeholders' Objectives—Robert Leonhard

The Objectives and Perspectives of South Korea—Lisa Collins

Russia's Vital Interests in South Korea—Stephen Blank

Game Theory and North Korea: A Literature Review—Kelly Rooker

Nuclear Crisis Bargaining and Escalation Revisited—Robert Powell

Extended Deterrence of North Korea: A Three-Party Game—Stephen Quackenbush

Report on a Conversation with Tom Schelling about North Korea—Yaneer Bar-Yam

Stabilizing Cooperative Outcomes in Nuclear Conflicts: Theory and Cases—Steven Brams

August 16, 2018

Bargaining and North Korea—William Spaniel

A Game Theory Analysis of the Stability–Instability Paradox—Barry O'Neill

Strategic Consequences of Psychological Factors and Emotional Misrepresentation in Negotiation—Alexandra Mislin

Nuclear Weapons and Nuclear Risk on the Korean Peninsula: Two Game Theoretic Takes—James Fearon

Denuclearization or Not? A Multiple-Player Sequential Game Model—Jun Zhuang

Common Conjectures and International Norms and Law—James Morrow

Strategic Causes of Proliferation: Northeast Asia in Comparative Perspective—Alexandre Debs

The Utility of Game Theory for Understanding Multilateral Nuclear Stability—Discussion

2

BACKGROUND

After a tutorial on game theory provided by Jun Zhuang, the background session of the workshop proceeded to address three main topics.

First, the history of US–Democratic People's Republic of Korea (DPRK) relations from the early twentieth century to modern day was presented by Robert Leonhard. This overview included motivations for North Korea to develop nuclear weapons and how the history between the two countries has unfolded into the present nuclear crisis. Additionally, this presentation highlighted recent developments in the crisis, including the war of words between US president Donald Trump and DPRK leader Kim Jong Un and the June 2018 summit. This presentation is summarized in the paper titled "The North Korean Nuclear Crisis."

Second, to provide a better understanding of the perspectives of several stakeholder nations, presentations were given on the objectives of the United States (Elaine Bunn), the DPRK (Robert Leonhard), the Republic of Korea (ROK) (Lisa Collins), and Russia (Stephen Blank). The "National Objectives and Perspectives" paper summarizes the key considerations when developing an understanding of national objectives and includes a discussion of decision-making models by Robert Leonhard.

Last, the background session included a review of the current state of the literature on game theoretic methods applied to the North Korean nuclear crisis by Kelly Rooker. This presentation, described in "Previous Game Theory Work on North Korea," summarized the merits, limitations, and faults of the available published work on this topic.

The North Korean Nuclear Crisis

Goryeo Nexus[1]—Origins of Conflict

The origins of a historical episode can be challenging to find. Historians tend to march leftward on the timeline in search of the elusive seed corn, because each event and person is inextricably linked to an antecedent. But in trying to pinpoint the most decisive launch point for the twenty-first-century nuclear crisis in Northeast Asia, perhaps the best place to start is with the 1868 Meiji Restoration in Japan. A powerful faction of Japanese militarists took control of the emperor and directed the country toward modernization and military expansion. At about the same time, China was retreating from the modern world into its traditional Confucian past. The resulting imbalance led to the rapid expansion of Japanese power throughout Northeast Asia. In the Sino-Japanese War of 1894–1895, the Japanese army and navy completely dominated the outdated forces of the Qing Dynasty and ended China's domination of the Korean Peninsula. The Japanese went on to defeat the Russians in the Russo-Japanese War of 1904–1905, removing the remaining great power that might aid Korea against them, at which point they forced the Korean emperor to accept a Japanese protectorate. Five years later, Tokyo annexed Korea completely.[2] The regime that would eventually come to rule North Korea thus inherited a clear lesson from history: military weakness leads to foreign occupation.

Japanese rule was harsh and oriented on benefiting the mother country. There were strong racist and chauvinistic overtones to the occupation as Japanese officials attempted to eradicate Korean culture and replaced Korean currency with their own. A Korean government in exile continued to advocate for liberation, and on the peninsula, there were sporadic demonstrations that the Japanese police and military put down with great bloodshed. Sustained popular reaction to Japanese (and before that, Chinese) domination developed into a national ideology that championed Korean ethnicity and eventual autonomy.[3] During World War II, Korean insurgent leaders—including future ruler of North Korea Kim Il Sung—joined with China's People's Liberation Army in fighting against Japanese occupiers in Manchuria. As the war against Japan drew to a close, the Koreans and their various international sponsors (primarily the United States, the USSR, and Chinese communists) contemplated what form of government would follow once the Japanese were gone. At the 1943 Cairo Conference, Allied leaders declared that Korea

[1] Goryeo is the Anglicized version of an ancient name for Korea, derived from the name of one of the Three Kingdoms, Goguryeo.

[2] Jansen, *The Making of Modern Japan*, 333–414.

[3] Seth, *North Korea*, 6–26.

would be free and independent. Two years later, at the Yalta Conference (February 1945), they agreed that the Allied forces would set up a trusteeship on the peninsula, with the eventual goal of establishing a capable Korean government.[4]

The Americans had long agitated for the Soviet Union to enter the war against Japan, and in August 1945 they got their wish. The Soviets invaded Manchuria on August 9 and quickly overran the Japanese defenses. On August 15, Emperor Hirohito communicated his order to surrender to the Allies, but because of delays in communicating a cease-fire to Japanese troops and the Soviets' desire to seize Sakhalin Island (and, if possible, Hokkaido), the campaign continued.[5] The Soviet Army invaded the Japanese-occupied Korean Peninsula and drove south. The Allied powers scrambled to transition from world war to some sort of postwar international order, and Korea was not among the chief priorities. Nevertheless, the Americans grew anxious knowing that Soviet armies were marching southward on the peninsula, and they scrambled to find a temporary demarcation line between the Soviet and American occupying forces. A Pentagon staff officer, Lieutenant Colonel Dean Rusk (who would later become one of the United States' longest-serving secretaries of state) described how he and another staff officer found a convenient dividing line.

> During a meeting on August 14, 1945, the same day as the Japanese surrender, [Bonesteel] and I retired to an adjacent room late at night and studied intently a map of the Korean peninsula. Working in haste and under great pressure, we had a formidable task: to pick a zone for the American occupation. Neither Tic nor I was a Korea expert, but it seemed to us that Seoul, the capital, should be in the American sector. We also knew that the U.S. Army opposed an extensive area of occupation. Using a *National Geographic* map, we looked just north of Seoul for a convenient dividing line but could not find a natural geographical line. We saw instead the thirty-eighth parallel and decided to recommend that . . . [Our commanders] accepted it without too much haggling, and surprisingly, so did the Soviets.[6]

The Fruits of Bifurcation—The Cold War and Korea

Thus, the Korean Peninsula was split into a northern section administered by the Soviets and a southern section administered by the United States. The temporary nature of the division, however, was overcome by the developing Cold War between the Soviet Union and the United States, along with their respective blocs. Diplomats met repeatedly to try

[4] Seth, *A Concise History*, 325–343.

[5] Dower, *Embracing Defeat*, 34; and Wilson, "The Bomb Didn't Beat Japan."

[6] Rusk, *As I Saw It*.

to work out an agreement on a national government, but they deadlocked. The Soviets wanted to set up a communist client state, while the United States wanted a capitalist democracy. Frustrated by Soviet intransigence, US president Harry Truman submitted the matter to the newly formed United Nations (UN), and the UN called for elections in 1947.[7] The Soviets refused to allow UN officials into the north to monitor the elections. A communist-led uprising on the island of Jeju began in April 1948 and was bloodily suppressed by the South Korean army by May 1949.[8] In December 1948, the UN General Assembly recognized the government of the Republic of Korea (South Korea) as the sole legitimate regime on the peninsula. Both the northern communist government in Pyongyang and the Republic of Korea (ROK) government claimed sovereignty over the entire peninsula, making a violent clash likely.

In the South, Syngman Rhee headed up the Provisional Government, set up with American assistance after the war. Kim Il Sung, former guerrilla fighter in Manchuria, emerged as the leader of the Soviet-sponsored communist government in the north. After the abortive Jeju Uprising,[9] he realized that he had no hope of winning the peninsula through political action. He applied to Moscow and Beijing for permission and aid to unite the country through force of arms. Soviet premier Joseph Stalin was persuaded to permit the military adventure because of the triumph of Mao Tse Tung's communist government in China in 1949 and the fact that most US troops had already departed Korea. Further, the Soviets had conducted the first of their underground nuclear tests, breaking the American monopoly on the atomic bomb, thus reducing the chance that a Korean adventure would spark a global war. Mao Tse Tung, though likewise concerned with what the Americans might do, needed the aid that the Soviets promised for his support, so he assented. On June 25, 1950, forces of the Democratic People's Republic of Korea (DPRK) surged across the 38th parallel and quickly overran the ill-trained, poorly equipped South Korean army and the small American contingent. The North Koreans captured Seoul as the desperate defenders retreated into a perimeter around the port city of Pusan.[10]

In September, however, General Douglas MacArthur launched an amphibious invasion at Inchon, and American forces cut off the North Korean invaders. The audacious counter-offensive turned the tide, and North Korean forces retreated. Flushed with victory, the

[7] UN General Assembly, Resolution 112, The Problem of the Independence of Korea.

[8] *New World Encyclopedia Online*, s.v. "Jeju Uprising."

[9] From April 1948 through May 1949, communist guerrillas fought for control of Jeju Island off the southern coast of Korea. The Provisional Government sent troops there and engaged in a brutal suppression aimed at eradicating the communists. Estimates of resulting deaths vary from 14,000 to 30,000.

[10] Fehrenbach, *This Kind of War*, 3–118.

American-led UN forces pushed northward, intending to eradicate the DPRK and reunite the peninsula under ROK leadership. But as early as August, the Chinese contemplated intervention to prevent the collapse of their client communist regime. Mao and the Chinese Communist Party Politburo were receiving desperate pleas from Kim Il Sung, as well as strong advice from Stalin, to send the People's Liberation Army into the peninsula. The specter of victorious American troops possibly continuing beyond the Yalu River in an attempt to overthrow the Chinese communists finally decided the issue.[11] Mao and his generals canceled a planned invasion of Taiwan and instead moved the army to Korea. On October 25, 1950, the renamed People's Volunteer Army crossed the Yalu with some two hundred thousand troops. Disciplined and skilled in camouflage, the Chinese soldiers moved and struck at night to offset American airpower. They pushed the UN forces southward once again, capturing Seoul for the second time. After a long retreat, the US-led coalition forces recovered and counterattacked, recapturing Seoul. In April 1951, President Truman relieved General MacArthur, who had pressed for the authority to use nuclear weapons and wanted to widen the war to defeat China. The fighting stagnated close to the 38th parallel, and the war dragged on until July 1953 when the two sides agreed to an armistice. The war did not officially end, but the fighting ceased for the most part. Both sides settled in for a long-term conflict of words, ideology, and strategic confrontation.[12]

In the South, Syngman Rhee became president of an American-backed republic, but his twelve-year rule degenerated into a harsh anti-communist dictatorship. As often occurred during the Cold War, the United States found itself allied to a less-than-desirable partner that it had to work with in the cause of containing Soviet (and now Chinese) communism. After Rhee was forced out of office in 1960, a succession of strongmen regimes, punctuated by brief attempts at real democracy, finally gave way to popular demands for genuine representative government and the rule of law in 1987. Throughout its history, despite following a bumpy and bloody path to liberalism, the ROK boasted a growing economy and eventually a credible and legitimate democratic government.[13]

The House of Kim

In the North, Kim Il Sung continued to consolidate his power, eliminating and executing potential rivals. The 1953 cease-fire corresponded with the death of Stalin, thus depriving the communist world of its most vaunted icon. When Nikita Khrushchev came to power in the Soviet Union in 1956, he began a shocking campaign of de-Stalinization that

[11] Halberstam, *Coldest Winter*, 359.

[12] Halberstam, *Coldest Winter*, 148–451.

[13] For a general history of South Korea, see Tudor, *Korea*.

reverberated throughout the communist bloc. Kim Il Sung despised the new Soviet leader's retreat from the Stalinist model as ideological weakening in the USSR, and in the place of Stalin, he began to rise to godlike status within his own country. His image was treated with reverence, and his regime enforced his deification by, for example, imprisoning someone who wrapped a piece of fish in a newspaper that bore Kim Il Sung's image. He began to rewrite history, claiming that he alone led the guerrilla campaign against the Japanese in Manchuria, and that the United States and South Korea had initiated the Korean War by attacking the North.[14]

Kim Il Sung, frustrated with what he viewed as liberal compromises to communist orthodoxy in China and the Soviet Union, began to isolate his regime from the rest of the world. He developed the notion of *Juche*—national self-reliance—as a bulwark against world trends he could not accept. At the same time, he began to plan for his son to succeed him, and his bloodline became part of the state's enforced worship of him.

Juche ideology initially appeared to be successful as, under Kim Il Sung's leadership, the country's economy grew. By the end of the 1950s, farms had been collectivized and industries had been nationalized, eliminating capitalism. The state-driven economy focused on developing heavy industry in the name of national security, and there were few consumer goods. The population was compelled to work harder and follow "The Great Leader" in defending the nation against supposed American aggression and Soviet backsliding. By the late 1960s, nearly all homes had electricity, there were numerous colleges and universities, and gross domestic product per capita was on an equal footing with the South. Nearly 70 percent of the population was urbanized, and one prominent British economist described Kim Il Sung's achievements as a "miracle."[15]

But the miracle turned sour in the 1970s. Kim's obsession with heavy industry and the military left the population largely without consumer goods. The nation's mineral export industry had peaked, and the 1973 oil crisis caused prices to decline, creating budget deficits for Pyongyang. Kim attempted to compensate with foreign loans, but in short order the DPRK was unable to repay the loans and lost its ability to borrow. Meanwhile, South Korea and the West had evolved to high-technology economies, which left North Korea even further behind. The Richard Nixon administration was exploring a détente with Moscow and a paradigm-shifting presidential visit to China. Kim's hard-line stance toward the South and the United States, coupled with the changing attitudes in Beijing and Moscow toward the Americans, left the regime completely isolated.[16]

[14] Oberdorfer and Carlin, *Two Koreas*, 6–9.

[15] Cumings, *Korea's Place*, 404.

[16] Ostermann and Person, *Rise and Fall*, 18, 19, 26–33.

While the United States struggled to extricate itself from the stagnating Vietnam War, Kim's regime used the opportunity to continually provoke and challenge the Americans. He refused to forgo his intention to eventually reunite the peninsula through war, and he instigated frequent border clashes, along with launching an attempted assassination of the ROK president. In 1968, DPRK warships seized the USS *Pueblo*, imprisoning its crew. Only after forcing US officials to apologize did the regime finally release the captive crew, who had been starved and tortured for nearly a year.[17] The following year DPRK fighter jets shot down two US EC-121 aircraft, killing thirty-one crewmen. The newly elected Nixon administration chose not to respond to the incident, in part because it was distracted by the Vietnam War and other world affairs. In 1976, DPRK soldiers murdered two American soldiers engaged in routine tree-clearing in the demilitarized zone, an incident that the regime later apologized for. Throughout the later phases of the Cold War, there were few attempts at negotiation between the United States and North Korea.

The Nuclear Notion

Toward the end of the Korean War, Dwight Eisenhower, campaigning for the presidency, mentioned that he would consider the use of nuclear weapons to end the conflict. His statement comported with MacArthur's desire to use the bomb to defeat Chinese and North Korean forces prior to his removal. Kim Il Sung took the threat seriously and was determined—in line with Juche ideology—to create his own nuclear deterrent, rather than rely on his wavering patrons in Moscow for nuclear protection.[18] Kim Il Sung, having seen American forces nearly conquer his country, was convinced that the United States was his chief obstacle in achieving reunification of the peninsula. He reasoned that the solution to the constant threat of American invasion would be to have his own nuclear arsenal. He applied to the Soviet Union for help in developing nuclear technology, and the Soviets agreed. Soviet scientists began to train their North Korean counterparts in 1956. Two years later, US forces began to deploy tactical nuclear weapons on the peninsula. In 1959, the Soviets and the DPRK responded with a formal agreement to cooperate in nuclear research. In 1963 the DPRK refused to sign the Nuclear Non-Proliferation Treaty (NPT), and two years later, it began to operate the Yongbyon Nuclear Scientific Research Center. It was a power plant, but Kim Il Sung's overriding interest was in expanding its use toward the creation of weapons.[19] The presence of American tactical nuclear warheads (artillery shells and air-delivered gravity bombs) on the Korean Peninsula seemed to provide clear evidence of Washington's aggressive intentions—a perspective that an increasing number

[17] Lerner, *Pueblo Incident*.

[18] Waxman, "How North Korea's Nuclear History Began."

[19] Oberdorfer and Carlin, *Two Koreas*, 196–198.

of American pundits agreed with. Consequently, American administrations began to roll back the number of weapons until George H. W. Bush removed them altogether in 1991.[20]

North Korea's development of nuclear weapons technology unfolded over four phases. From the nation's harrowing experience in the Korean War, Kim Il Sung concluded that he must lead his country's transformation into a fortress capable of deterring or defeating aggression from the United States. To that end, Kim asked both the USSR and China for help in developing nuclear weapons—a request that both at first refused. Instead, the Soviets offered help in building a nuclear industry for power generation. From 1956 through 1980, North Korea engaged in basic research and uranium mining. This initial phase included construction of the Yongbyon Nuclear Scientific Research Center. With Soviet help, Kim's regime constructed its first nuclear reactor in 1963—a small facility aimed at research that was later upgraded to eight-megawatt output in 1974.[21]

The second phase of the DPRK's nuclear weapons program began in 1980 and featured domestic production of plutonium. The regime opened a second nuclear plant at Yongbyon in 1980 and achieved a four-megawatt production level in 1986. From 1980 through 1994, the regime focused on producing weapons-grade plutonium. It signed the NPT in 1985 to gain leverage in forcing the Americans to withdraw nuclear weapons from the peninsula, but Pyongyang did not complete a safeguards agreement with the International Atomic Energy Agency (IAEA) until 1992.[22] When IAEA officials suspected noncompliant behavior and requested permission to inspect North Korean facilities, they were refused. In 1993, the DPRK threatened to withdraw from the NPT but later suspended the decision. Soon after, US intelligence reported that the regime was manufacturing nuclear weapons.

The third phase of nuclear weapons development commenced in 1994. Negotiations aimed at restoring IAEA access to North Korea stalled with the sudden death of Kim Il Sung in July of that year. His son, Kim Jong Il, resumed talks with the United States, while secretly intending to continue pursuing his father's dream of a nuclear deterrent.[23] The Bill Clinton administration negotiated the "Agreed Framework" under which North Korea would give up its nuclear weapons industry in exchange for energy imports and the construction of two light-water reactors. From 1994 to 2002, the DPRK froze plutonium development in accordance with the Agreed Framework, but, according to intelligence estimates, switched to uranium production and a secret bomb program.[24] While the

[20] Oberdorfer and Carlin, *Two Koreas*, 198–200.

[21] Jae-Bong, "US Deployment of Nuclear Weapon."

[22] Arms Control Association, "Chronology of U.S.-North Korean Nuclear and Missile Diplomacy."

[23] Arms Control Association, "Chronology of U.S.-North Korean Nuclear and Missile Diplomacy."

[24] Arms Control Association, "The U.S.-North Korean Agreed Framework at a Glance."

alleged uranium project was not specifically addressed in the Agreed Framework, it was clearly a violation of the Americans' intent for North Korea to cease nuclear weapons research. The George W. Bush administration thereby ended American aid programs and, in 2002, designated North Korea as part of an "Axis of Evil."

The fourth phase of Pyongyang's development of nuclear weapons corresponded with the end of the Agreed Framework. From 2002 through the present, North Korea established itself as a recognized nuclear power, withdrawing from the NPT and conducting its first underground explosion in 2006. The George W. Bush administration (2001–2009) participated in the six-party talks (including the United States, China, Russia, DPRK, ROK, and Japan) from 2003 through 2007, but there was little progress on the core issues of the DPRK's nuclear program and Pyongyang's demands for normalization and economic aid. But after the last of the six-party talks in 2007, the regime agreed to shut down nuclear weapons sites. It seemed for a brief period that a thaw might be in the offing, but in 2009, North Korea conducted a second underground test. The agreements that had proceeded from the six-party talks proved too brittle and subject to criticism from skeptics on all sides, including South Korea's new president, Lee Myung-bak, who was in no mood to engage the North.[25] Despite the arrival onstage of America's new president, Barack Obama, and his declared intention to work with former enemies if they would work with him, North Korea forged ahead with nuclear testing and ballistic missile development. A third underground test in February 2013 underscored the regime's determination to pursue their nuclear ambitions regardless of sanctions. From 2016 to 2017, it exploded three more nuclear warheads, claiming that it had produced a hydrogen bomb. While some experts disputed the claim, it gradually became clear that the regime had in fact produced credible nuclear warheads. The destructive power of their weapons had graduated from two kilotons in 2006 to over two hundred kilotons in 2017.[26]

Concurrent with warhead development, the North Koreans strove to build intercontinental ballistic missiles (ICBMs) that could carry their weapons to hit US and regional targets. The missile program faltered at first and experienced numerous failures in engineering, but by 2017, it appeared the regime had an ICBM capable of hitting the United States. While questions remain whether the missile could carry a nuclear warhead and survive reentry, the North Korean regime had achieved its goal of developing a credible nuclear deterrent.

[25] Oberdorfer and Carlin, *Two Koreas*, 427–431.

[26] Lee, "North Korea's Latest Nuclear Test."

The Logic of Stalemate

From his accession to power in 1994, Kim Jong Il continued to embrace Juche as a national ideology and added the concept of *Song-Un*—"military first"—to the narrative. Kim Jong Il continued to propagate the doctrine that the United States was determined to invade and destroy North Korea, thus justifying continued favoring of military developments over much-needed economic diversity. He ruled from 1994 through his death in 2011, during which period floods and famine killed about 3.5 million people. His son, Kim Jong Un, succeeded him in 2011 and spent the initial period of his reign brutally ridding himself of perceived threats to his rule. Kim Jong Un strengthened his grip on the government and society, and he parroted his father's paranoia about international American-led conspiracy threatening the nation.[27] His determination to achieve a nuclear deterrent was unwavering and ultimately successful. But the cost of his achievement was an ever-worsening regime of economic sanctions from the international community.

International concern over the DPRK's nuclear program dated back to the administration of president George H. W. Bush in 1989. American officials began to inform world leaders on both sides of the Iron Curtain of their suspicions about North Korea's intentions. Secretary of state James Baker set out to shape a diplomatic strategy aimed at coercing an end to Pyongyang's nuclear ambitions.[28] From the late 1990s through the present, these efforts have led to a multifaceted regime of international sanctions that would deepen as the standoff dragged on. The UN, in a series of Security Council resolutions, limited the import of military goods, luxuries, and oil. It later banned exports of minerals and seafood in 2017, thus cutting into the nation's budget. It also restricted financial dealings and authorized inspection of vessels trading with the regime.[29] The Obama administration widened sanctions by targeting entities that traded with North Korea, and the Trump administration targeted travel, trade, and finances. The ROK suspended trade and cultural exchanges and also banned DPRK vessels from ROK waters. Japan banned travel to the island nation, port visits, remittances, and financial activity. Australia imposed independent sanctions against North Korea's extractive industry and banned service to DPRK airlines. But China remained North Korea's chief trading partner and was constantly under pressure to join the international community in coercing Pyongyang. It responded—particularly after underground tests—by limiting or banning coal and textile imports. Despite the tightening restrictions, the North Koreans clung to Juche, Song-Un, and their nuclear weapons.

[27] Oberdorfer and Carlin, *Two Koreas*, 459–461.

[28] Oberdorfer and Carlin, *Two Koreas*, 459–461.

[29] UN Security Council, Resolution 1718, Non-Proliferation/Democratic People's Republic of Korea.

Rocket Man and the Dotard

With the election of a new and unconventional American administration in 2016, the course of US–DPRK diplomacy changed. President Donald Trump expressed his willingness to meet with Kim Jong Un, even in the middle of an alarming war of words between Washington and Pyongyang from 2017 to 2018. In the spring and summer of 2017, Kim Jong Un ordered a series of missile tests that threatened Japan and left Western analysts to speculate whether the weapons were a viable threat to the continental United States. Donald Trump responded with a series of tweets threatening Kim Jong Un with invasion and total destruction and also calling him derogatory names such as "rocket man," "short and fat," and "madman." Kim Jong Un responded by threatening nuclear retaliation against the United States and claiming that Trump was "bereft of reason," a "frightened dog," a "hideous criminal," and a "dotard." The two leaders went on to declare that their fingers were "on the button" as traditional pundits all over the world gasped in disbelief and concern.[30]

Then, suddenly, this combative exchange gave way to a remarkable diplomatic breakthrough, in part due to backchannel dealing that had continued unabated. In February 2018, North Korea sent a high-level diplomatic mission headed by Kim Yo-jong, sister of Kim Jong Un, and president Kim Yong-nam to the Winter Olympic Games in South Korea. The delegates from Pyongyang invited ROK president Moon Jae-in to visit North Korea. Soon afterward came headlines that a Trump–Kim summit was being planned. Following some diplomatic wavering on both sides, the two leaders met in Singapore in June. The event produced little of substance, but the symbolism was significant, because now the same two leaders who had been trading threats of nuclear devastation were praising each other and expressing hope for a solution. North Korea returned some remains of American soldiers left behind in the Korean War. Kim Jong Un followed that with a public destruction of the Punggye-ri nuclear test site. President Trump, perhaps impulsively, agreed to suspend US–ROK military exercises. Old hands from the national security arena were incredulous and warned that the Americans were giving up much and receiving little or nothing back.[31] After all, Kim had made no commitments concerning missiles, cyberwarfare, human rights, or nuclear proliferation. Although he theoretically agreed to work toward denuclearization of the peninsula, there has, so far, been little movement in that direction. Some intelligence analysts have claimed that nuclear research continued unabated. National security advisor John Bolton declared in August: "North Korea has not

[30] Yun, "Is a Deal With North Korea Really Possible?" See also Nielsen, "Donald Trump Says on Twitter"; and Keating, "Donald Trump Calls Kim Jong-un 'Little Rocket Man.'"

[31] See for example, Brewer, "Can the U.S. Reinstate 'Maximum Pressure' on North Korea?"

taken the steps we feel are necessary to denuclearize."[32] Pyongyang's position, particularly since the rise of Kim Jong Un, is that denuclearization must be global and not limited to its regime. As early as 2013, the DPRK government concluded "that the denuclearization of the Korean Peninsula is impossible unless the denuclearization of the world is realized as it has become clear now that the US policy hostile to the DPRK remains unchanged."[33] Thus, the two nations' positions remain wide apart.

Immediately following the Singapore summit, President Trump and Kim Jong Un changed their rhetoric from angry posturing to a lovefest. On June 12, Trump declared that he and Kim had "got along great" and that Kim is "a very talented guy . . . great personality . . . a worthy negotiator . . . very smart." He went on and, concerning Kim's intent, stated "I think he wants to do a great job for North Korea . . . I think he wants to de-nuke. I trust him, and he trusts me." Negotiations continued, but in July, in response to American demands for concrete steps toward "complete, verifiable, irreversible denuclearization," the North Korean Foreign Ministry insisted that America had resorted to "unilateral and robber-like denuclearization demands" and was violating the Singapore agreement.[34]

The negotiations and diplomatic intrigue continue as of this writing. Analysts are left with serious questions and a variety of possible outcomes. Questions include: Why would the Kim-family regime give up the very nuclear weapons that got them a seat at the table of international diplomacy? If it did give them up, what would a nonnuclear North Korea look like? Would sanctions go away, potentially stimulating economic growth in the North? Is political reunification of the peninsula a real possibility? If any of these breakthroughs are realized, will America's position have been enhanced or damaged in the region? Potential outcomes in the near term are likewise interesting. What will the next succession look like? Will the Kim family retain power or face a coup or civil war? Or will all the diplomatic achievements in the end be illusory, and war will come again to Korea? The history of this troubled peninsula promises continued human drama and teaches us to temper our hopes with a pragmatic understanding of the struggle for power in the Land of the Morning Calm.

References

Arms Control Association. "Chronology of U.S.-North Korean Nuclear and Missile Diplomacy." Updated March 2019. https://www.armscontrol.org/factsheets/dprkchron.

[32] Wong and Sanger, "Trump to Meet with Kim Jong-un."

[33] Oberdorfer and Carlin, *Two Koreas*, 459–461.

[34] Arms Control Association, "Chronology of U.S.-North Korean Nuclear and Missile Diplomacy."

———. "The U.S.-North Korean Agreed Framework at a Glance." Updated July 2018. https://www.armscontrol.org/factsheets/agreedframework.

Brewer, Eric. "Can the U.S. Reinstate 'Maximum Pressure' on North Korea?" *Foreign Affairs*, December 4, 2018. https://www.foreignaffairs.com/articles/north-korea/2018-12-04/can-us-reinstate-maximum-pressure-north-korea.

Cumings, Bruce. *Korea's Place in the Sun: A Modern History*. New York: W. W. Norton & Co., 2005.

Dower, John W. *Embracing Defeat: Japan in the Wake of World War II*. New York: W. W. Norton, 1999.

Fehrenbach, T. R. *This Kind of War: The Classic Korean War History*. 50th anniversary ed. Dulles, VA: Brassey's, 2001.

Halberstam, David. *The Coldest Winter: America and the Korean War*. New York: Hyperion, 2007.

Jae-Bong, Lee. "US Deployment of Nuclear Weapons in 1950s South Korea & North Korea's Nuclear Development: Toward Denuclearization of the Korean Peninsula." *Asia-Pacific Journal* 7, no. 8 (2009): 1–17.

Jansen, Marius B. *The Making of Modern Japan*. Belknap Press, 2000.

Keating, Fiona. "Donald Trump Calls Kim Jong-un 'Little Rocket Man' as He Again Threatens North Korea." *Independent*, September 23, 2017. https://www.independent.co.uk/news/world/americas/donald-trump-kim-jong-un-little-rocket-man-north-korea-alabama-senator-luther-strange-nuclear-a7962771.html.

Lee, Michelle Ye Hee. "North Korea's Latest Nuclear Test Was So Powerful It Reshaped the Mountain above It." *Washington Post*, September 14, 2017. https://www.washingtonpost.com/news/worldviews/wp/2017/09/14/orth-koreas-latest-nuclear-test-was-so-powerful-it-reshaped-the-mountain-above-it/?utm_term=.65b3f70ee6b7.

Lerner, Mitchell B. *The Pueblo Incident: A Spy Ship and the Failure of American Foreign Policy*. Lawrence, KS: University of Kansas Press, 2002.

New World Encyclopedia Online, s.v. "Jeju Uprising." http://www.newworldencyclopedia.org/p/index.php?title=Jeju_Uprising.

Nielsen, Kevin. "Donald Trump Says on Twitter He Would Never Call Kim Jong Un 'Short and Fat.'" *Global News*, November 11, 2017. https://globalnews.ca/news/3856729/donald-trump-twitter-kim-jong-un-short-and-fat/.

Oberdorfer, Don, and Robert Carlin. *The Two Koreas: A Contemporary History*. 3rd ed. Philadelphia: Basic Books, 2013.

Ostermann, Christian F., and James F. Person, eds. *The Rise and Fall of Détente on the Korean Peninsula, 1970–1974*. Washington, DC: Woodrow Wilson International Center for Scholars, 2011.

Rusk, Dean. *As I Saw It*. New York: W. W. Norton & Co. Inc., 1990.

Russel, Daniel R. "A Historic Breakthrough or a Historic Blunder in Singapore?" *Foreign Affairs*, June 12, 2018. https://www.foreignaffairs.com/articles/north-korea/2018-06-12/historic-breakthrough-or-historic-blunder-singapore.

Seth, Michael J. *North Korea: A History*. Red Globe Press, 2018.

———. *A Concise History of Korea: From Antiquity to the Present*. 2nd ed. Rowman & Littlefield Publishers, 2016.

Tudor, Daniel. *Korea: The Impossible Country*. North Clarendon, VT: Tuttle Publishing, 2012.

UN General Assembly. Resolution 112, The Problem of the Independence of Korea. A/RES/112(II). 1947. https://documents-dds-ny.un.org/doc/RESOLUTION/GEN/NR0/038/19/img/NR003819.pdf?OpenElement.

UN Security Council. Resolution 1718, Non-Proliferation/Democratic People's Republic of Korea. S/RES/1718. October 14, 2006. https://undocs.org/S/RES/1718(2006).

Waxman, Olivia B. "How North Korea's Nuclear History Began." *Time*, March 7, 2017. http://time.com/4692045/north-korea-nuclear-weapons-history.

Wilson, Ward. "The Bomb Didn't Beat Japan . . . Stalin Did." *Foreign Policy*, May 30, 2013. https://foreignpolicy.com/2013/05/30/the-bomb-didnt-beat-japan-stalin-did/.

Wong, Edward, and David E. Sanger. "Trump to Meet with Kim Jong-un, Despite North Korea's Lapses, Bolton Says." *New York Times*, December 4, 2018. https://www.nytimes.com/2018/12/04/us/politics/trump-kim-north-korea-summit-meeting.html.

Yun, Joseph. "Is a Deal With North Korea Really Possible? The Gap Between Expectations and Reality." *Foreign Affairs*, May 30, 2018. https://www.foreignaffairs.com/articles/north-korea/2018-05-30/deal-north-korea-really-possible.

National Objectives and Perspectives

A nuanced understanding of the objectives of the nations central to the North Korean nuclear crisis is a highly desirable, yet formidable, international relations challenge. Knowing the objectives of a nation has the potential to help in predicting which actions that nation will take in response to another nation's actions, and therefore an understanding of the objectives is critically important when applying game theoretic methods to scenarios related to the North Korean nuclear crisis. However, developing this understanding is an arduous task because the objectives of any nation are (often intentionally) imprecise and ambiguous, interconnected with one another, dynamic, and situationally dependent. The discussions at the workshop on the national objectives focused on the complexity and uncertainty in determining the objectives and the relationships between them. Several key points raised in these discussions are summarized here.

First, issues arise from looking at both long-term and short-term objectives simultaneously. A nation in a crisis is likely to focus primarily on the short-term objectives because (1) there are more obvious solutions to immediate problems; and (2) there is hope that over time some other unforeseen element in the situation will change, thereby avoiding pernicious long-term consequences.

Second, considering a single objective in a vacuum in unrealistic. Each objective has a complex relationship with multiple other objectives, and these relationships among the objectives, such as prioritization, trade-offs, and compromises, like the objectives themselves, are situationally dependent and evolve over time.

Additionally, subpriorities may be implicit but still need to be considered. An example given at the workshop was related to the US objective of deterring an attack on US territory. A participant raised the question: Would it matter whether North Korea used nuclear weapons against Hawaii (a US state with a lot of military forces), Guam (a US territory with a lot of military forces), or American Samoa (a US territory)? The participant noted that these subpriorities are important, but one cannot expect any political leader to subprioritize them for every situation.

Last, the differences in the prioritization of the objectives between allied nations will have operational implications. Elaine Bunn gave the example of the objective shared by the United States, South Korea, and Japan of preventing North Korea from using nuclear weapons against the United States, its allies, or its forces. One could imagine a scenario where the United States sees a North Korean nuclear missile on the launchpad. If the United States felt its homeland were threatened, it might decide to destroy the missile before launch, even while increasing the risk of North Korean retaliatory attacks on Japan and South Korea.

Central to determining the objectives of a nation is understanding how decisions are made in that particular nation. Because game theory requires a distinct set of players, and that the players act in their own best self-interests, this is especially important when considering how understanding national objectives can inform a game theoretic analysis. Several decision-making models were presented by Robert Leonhard, and this discussion is summarized on the following pages.

Decision-Making Models and National Objectives

When we consider the national objectives of the regional and world powers who constitute the major stakeholders in the North Korean nuclear problem, it is tempting to imagine that national objectives originate from the designated leaders of the respective countries. After all, presidents, prime ministers, and party chairs obtain their power to govern, and governing includes setting foreign policy objectives. However, history teaches us that the formulation of policy objectives is more nuanced than that. Historical experience, economics, bureaucratic process, and social dynamics also play a part. Still, when we think about what a given country might do or why, we most often default to what is popularly referred to as the "rational actor" model.[1]

The term is unfortunate because it is often misunderstood. When we say that a certain governing individual is a rational actor, it is not meant as a commentary on their sanity, wisdom, or intelligence. Instead, the term is pointing to a theory about decision-making. The rational actor model proposes that a nation's executive leader makes the relevant decisions regarding what his or her country will do. Hence, when President Trump, or Prime Minister Abe, or Kim Jong Un decide on a course of action, the rational actor model suggests that the leader in question made their decision independently as a duly appointed head of state, and their respective countries will carry out that decision.

Political scientists, however, have found that what appears to be a simple and clear-cut decision-making apparatus is often more complex, and the underlying dynamics are driven by more than a single actor. Graham T. Allison described the phenomenon this way:

> For some purposes, governmental behavior can be usefully summarized as action chosen by a unitary, rational decisionmaker: centrally controlled, completely informed, and value maximizing. But this simplification must not be allowed to conceal the fact that a "government" consists of a conglomerate of semi-feudal, loosely allied organizations, each with a substantial life of its own.[2]

Analysts have developed various decision-making models that describe these different conditions. Besides the rational actor model, there are the government bargaining model (also known as the bureaucratic model), the organizational process model, the self-aggrandizement model, the political process model, and the social constructivist model.

[1] Jackson and Sorensen, *Introduction to International Relations*.

[2] Allison, "Conceptual Models," 698.

A Game of Chess—Decision-Making Models in Action

To illustrate how the various models play out, let us consider a game of chess in which your opponent sits across the table, and the game is about to begin. Chess is very much a rational actor game—in other words, it is literally two opponents deciding and moving to checkmate the other. Suppose, though, that your opponent is not actually deciding how to move each turn. Instead, other factors bring about his actual moves.

We begin with the government bargaining model. According to this concept, it is your opponent's *pieces* that are deciding what move will come next. The pieces bargain among themselves, each faction desirous of increasing its own power. The king and the queen, for example, represent a royal faction trying to protect their prerogatives. The bishops represent the church, while the knights embody the nobility. The pawns organize into a union of factory workers and soldiers who demand a say in the game. The stolid castles symbolize heavy industry oligarchs. For the pieces in this anarchic confederation to act, they will have to bargain and form bureaucratic alliances with each other. No single faction is likely to prevail all the time, and so compromises must be made. On this turn, the pawn union will have their way and advance the king pawn. However, the nobility will agree to that only if on the next turn they get to advance a knight. Hence, your opponent's moves are the product of bureaucratic infighting, rather than rational, independent thought.

Closely related to the government bargaining model is the organizational process model. In this case, the emphasis is less on the goals of each faction of pieces and more on the bureaucratic process itself. We might imagine, for example, that each turn's move is decided according to an established standard operating procedure. First, the pawn union generates a proposed move, perhaps with input from the royal faction. Next, the proposal is passed to the knights, who can either approve or reject the proposal but may not alter it. Members of the church faction—the bishops—have an ecclesiastical veto power, should they choose to use it. No matter what decision is arrived at, only the oligarchs (castles) have the means to fund it. Hence, without their tacit approval, the move cannot be made. This kind of chaotic process might tend to rule out brilliant, unexpected moves in favor of safe, methodical plays.

Now suppose the opponent has to deal with a vain, egotistical king. Instead of being a passive, protected piece who hangs out in the back rank trying not to succumb to checkmate, the king instead insists on a more active role—to bring glory to himself. This would be an example of a self-aggrandizement model. The resulting moves may appear ridiculous and risky because the purpose behind them is personal, not strategic.

The political process model turns the focus to external influencers. Imagine that instead of you and your opponent playing a nice quiet game of chess in the library, you are playing in the city park. A group of annoying teenagers are watching the game and mocking your opponent for his lack of aggressiveness. Meanwhile, a wizened old man sits on a nearby bench, shaking his head each time your opponent begins to move a piece. In this model, the resulting decisions are strongly influenced by factors external to the government itself. As before, the actual moves might lack coherence.

In a social constructiveness model, the focus shifts to the power of ideas—ideology, religious belief, racial ideas, political theory, and so forth. As you face your opponent, you have set yourself the rational objective of achieving a checkmate against his king. However, he is operating according to a very different agenda. He may be following a strong ideology instead of playing in response to your moves. For example, he may believe strongly that in every game of chess, he must castle his king to the kingside before the eighth move. Hence, he vigorously clears his back rank instead of playing his pawns properly. Alternatively, he may be following a favorite opening in which he must fianchetto his bishops early in the game—regardless of whether such moves would actually be the best response to your moves.

Thus, when we think about the various regional and global states concerned with the North Korean nuclear crisis, we must consider not simply their goals and objectives but also *how* they are developed. Various actions taken by each state might derive not simply from a rational, independent head of state, but rather from factionalized bureaucratic processes, slavish obedience to an ideology, or overreaction to media coverage. By thinking hard about decision-making models, we can achieve a nuanced understanding of how and why decisions are made.

Although such decision-making models do not necessarily affect the mathematical underpinnings of game theory, they can affect the setup of each game. When constructing a game consisting of states as players, it is important to know how each state's decisions get made. Does the state have a unilateral decision-making leader? If so, is the main motivation of that leader self-preservation, glory, or preservation of his or her state, ideology, or something else? Do key advisors and/or funders play a role? Does the state have various factions, each of which have a say in decisions? Does rule of law prevail, or can even unlawful decisions get made?

All these questions can be important considerations in the setup of game theoretic games. Simply casting a player as some "State 1" can be misleading if there are actually competing interests and/or decision-makers present within that State 1. Similarly, considering only the head of state's personal self-interest may oversimplify how that head of state makes

decisions on behalf of his or her state. As with all game theory, the setup of any game is tremendously important in the game's ability to model reality and produce useful outcomes. Without truly understanding each state's goals and objectives, as well as how and why decisions get made in each state, any resulting game will likely be lacking in its applicability to the real world.

References

Allison, Graham T. "Conceptual Models and the Cuban Missile Crisis." *The American Political Science Review* 63, no. 3 (1969): 689–718.

Howard, Caroline. "The World's Most Powerful People 2013: No. 46: Kim Jong Un." *Forbes*, October 30, 2013. Archived from the original on December 26, 2015.

Jackson, Robert, and Georg Sorensen. *Introduction to International Relations: Theories and Approaches*. 5th ed. Oxford: Oxford University Press, 2013.

Waxman, Olivia B. "How North Korea's Nuclear History Began." *Time*, March 7, 2017. http://time.com/4692045/north-korea-nuclear-weapons-history.

Previous Game Theory Work on North Korea

Game theory, especially in military or defense contexts, is often associated with the Cold War. Certainly a great deal of academic literature about it was published during that time. Since the Cold War, the literature on game theory in political science has not been as prominent. However, there has recently been more limited work on applying game theory to questions of nuclear deterrence. Much of this work builds general models, not specifically meant to apply to any current real-world situation. This literature review will instead focus on the even more limited body of work that has applied game theoretic modeling to the North Korean nuclear crisis.

The scope of the following literature review is published, peer-reviewed papers using game theoretic methods applied to the North Korean nuclear crisis. This includes papers explicitly using game theory to model the North Korean nuclear crisis, papers including generic game theory models (e.g., Nation 1 versus Nation 2 or State 1 versus State 2) that mention the North Korean nuclear crisis as one possible application of their model (typically in either the introduction or discussion sections of the paper), and/or papers describing game theoretic thinking applied to the North Korean nuclear crisis. Papers in this latter category do not necessarily have to include math or describe the equations/model fully. Instead they could just describe a more generic way of thinking about the rationality of the North Korean nuclear crisis, possibly setting up a game, and thinking logically through to its conclusion.

Note that by no means is this literature review meant to be an exhaustive list of all game theory work that could *potentially* be applied to the North Korean nuclear crisis. While a multitude of general game theory papers could likely be applied to the North Korean nuclear crisis, this literature review is restricted to just those game theory papers that specifically mention the North Korean nuclear crisis somewhere in them. While many of the papers fully analyze (i.e., obtain equilibria, draw conclusions, etc.) each of the games that will be described here, this literature review will include only the setup of each of the proposed games. However, all papers are referenced here in case more details on any particular game are desired.

A couple of themes will be present throughout this literature review. First, good game theory requires a lot of work. Some papers using game theory to help explain the North Korean nuclear crisis are very thorough, well researched, and include empirical data used to validate their models. The majority of the papers, however, are lacking for a variety of reasons. For example, some papers include unrealistic actors and/or actions for the North Korean nuclear crisis; others include actors with choices that are not distinct from one

another; some describe games that need unrealistic assumptions; others oversimplify scenarios solely for mathematical convenience; and still others have outdated scenarios, making current actor choices unrealistic. Many games presented in this literature review suffer from one or more of these problems.

The various games discussed here, used to help explain the North Korean nuclear crisis, can be grouped into several general categories. Games include different numbers of players (for example, games can be either bilateral, with only two states or players included, or multilateral, with more than two states or players involved). Games also include different states as the main players. For example, these players can be generically named (typically as State 1 versus State 2, Nation 1 versus Nation 2, or Country 1 versus Country 2). Alternatively, the players can be explicitly named (for example, United States versus North Korea, China versus North Korea, South Korea versus North Korea, or United States versus China).

Games can have different evaluation methods—for example, using classical game theory versus Bayesian game theory versus agent-based modeling. There are also different game categories. For example, games applied to the North Korean nuclear crisis can resemble bargaining or negotiation games, inspection or verification games, signaling games, and/or escalation games. In addition, extensions of game theory have been used in modeling the North Korean nuclear crisis—for example papers using the graph model for conflict resolution.

Since having rational actors is a key assumption of game theory, it is important to acknowledge this rationality assumption in game theory work applied to the North Korean nuclear crisis. North Korean leadership is often described as being irrational. For example, Lohschelder writes of Condoleezza Rice, the secretary of state under former US President George W. Bush, recalling Bush's description of Kim Jong Il[1]: "He throws his food on the floor, and all the adults run to gather it up and put it back on the table. He waits a little while and throws his food on the floor again." Or for a South Korean perspective, Hwee Rhak Park of the Korea National Defense University[2] writes: "The fact that North Korea's policy has its own logic does not mean that the policymakers of North Korea are 'rational,' as an insane person's logic cannot make him rational."

These common viewpoints lead to the question: Is Kim Jong Un really irrational? Current available evidence would suggest that he *is* most likely rational. Kim Jong Un's actions are well in line with the preservation of both himself and his power/position as supreme leader, even if not always in line with what is best for the North Korean people. This

[1] Lohschelder, "Why North Korean Foreign Policy Is Rational," 3.

[2] Park, "Self-Entrapment of Rationality."

conclusion holds even for Kim Jong Un's seemingly "irrational" actions such as violently killing his relatives.

The following games regarding the North Korean nuclear crisis are organized by game structure. The first category is simple bilateral games. Within this category are many players, not just the United States and North Korea. First, the game could be played between North and South Korea. The game described by Kim and Choi[3] is one example with these players. The following is the payoff matrix used in the setup of this game. North Korea has the option to be sincere or deceptive when conducting its diplomacy. South Korea has the option to either accept or reject the North Korean diplomacy.

		North Korea	
		Sincere Diplomacy	Deceptive Diplomacy
South Korea	Accept North Korean Diplomacy	*Outcome 1*	*Outcome 2*
	Reject North Korean Diplomacy	*Outcome 3*	*Outcome 4*

Outcome 1 can be viewed as reconciliation between North and South Korea, and outcome 2 as the exploitation of South Korea by North Korea. Note this game can be played repeatedly over time. This is an older game, from 2002, assuming an outdated notion of Korean reconciliation being feasible or even easy to accomplish. In most of the later games, reconciliation of the Korean Peninsula is not even considered an option.

Second, the game could be played between China and the United States, two world powers with much at stake in the North Korean nuclear crisis. An *Asian Survey* article offers an example.[4] China has the options of punishing North Korea or not punishing North Korea. The United States has the options of conducting a military response to North Korea or not conducting any military response to North Korea.

[3] Kim and Choi, "Uncertainty in Foreign Policy Making."

[4] Song, "Understanding China's Response."

		China	
		Punish North Korea	Do Not Punish North Korea
United States	Military Response to North Korea	*Outcome 1*	*Outcome 2*
	No Military Response to North Korea	*Outcome 3*	*Outcome 4*

The authors claim that China's motivation in making its decision is based on the US decision, as well as whether or not there is a threat to stability in North Korea. China wants a stable North Korea, since China's main focus is further economic development and stability in the region, something an unstable North Korea would greatly disrupt.

Third, the game could also be played between North Korea and China. In a *Journal of Modern Science* article,[5] the game is North Korea versus China. North Korea has the options to either stop its nuclear research or continue its nuclear research. China has the options to either attack North Korea or not attack North Korea (note: potentially unrealistic strategy choices from China's perspective).

		North Korea	
		Stop Nuclear Research	Continue Nuclear Research
China	Attack North Korea	*Outcome 1*	*Outcome 2*
	Do Not Attack North Korea	*Outcome 3*	*Outcome 4*

In addition to bilateral games being played between two specific opponents, there are also games where North Korea instead plays some general "opponent." For instance Park[6] and

[5] Levi, "Applying Game Theory to North Korea-China Relations."

[6] Park, "Application of Risk."

Park, Nho, and Yoon[7] presented games where North Korea is playing an "opponent" that the authors state could be South Korea, the United States, or China. In these games, both North Korea and its opponent can take either a hard-line policy strategy toward the other or a more moderate policy strategy toward the other (or appeasement). A strategy that is moderate or with appeasement would be one where the country is willing to compromise and more highly values peace or coming to a deal/agreement. Such games can also be extended to games with imperfect information (namely, whether or not North Korea already possesses a functional hydrogen bomb).

		North Korea	
		Hard-line Policy	Appeasement
Opponent	Hard-line Policy	*Outcome 1*	*Outcome 2*
	Appeasement	*Outcome 3*	*Outcome 4*

		North Korea	
		Hard-line Policy	Moderate Policy
Opponent	Hard-line Policy	*Outcome 1*	*Outcome 2*
	Moderate Policy	*Outcome 3*	*Outcome 4*

Many of the bilateral games, however, are between the United States and North Korea. Some of these involve larger games, where players have more than just two strategy options. For example, Obeidi, Hipel, and Kilgour[8] created a game between North Korea and the United States. North Korea has two options: either stop its nuclear weapons program or reduce its conventional weapons program. The idea is that North Korea only has a limited

[7] Park, Nho, and Yoon, "Quantitatively Modified Two-Level Game Theory."

[8] Obeidi, Hipel, and Kilgour, "Role of Emotions."

amount of money to spend on its military/defense. This means that North Korea will ultimately have to choose whether to invest that money in its nuclear weapons program or its conventional weapons program. In other words, North Korea's investment in its nuclear weapons program implies a reduction in its conventional weapons program.

The United States, on the other hand, has three choices. First, it could engage in direct negotiations with North Korea. Second, it could adopt an aggressive military posture. Or third, the United States could choose to influence North Korea through sanctions and isolation. This creates a game with six possible outcomes, instead of only four possible outcomes as in all the games previously discussed.

		North Korea	
		Stop Its Nuclear Weapon Program	Reduce Its Conventional Weapons Program
United States	Engage in Direct Negotiations with North Korea	*Outcome 1*	*Outcome 2*
	Adopt an Aggressive Military Posture	*Outcome 3*	*Outcome 4*
	Influence North Korea through Sanctions and Isolation	*Outcome 5*	*Outcome 6*

Jelnov, Tauman, and Zeckhauser[9] created a game between a generic Nation 1 and Nation 2. Nation 1 is some nuclear proliferator or a country currently developing nuclear weapons, while Nation 2 is a country overseeing Nation 1's nuclear capabilities, or in other words, acting as an observer. While the authors keep this game general, they do mention North Korea and the United States as possible Nation 1 and Nation 2 players, respectively.

In this game, Nation 1 (North Korea) has the options of (1) stopping its nuclear weapons research and opening up all its nuclear facilities for inspection, (2) stopping its nuclear weapons research, or (3) continuing its nuclear weapons research. Nation 2 (the United States) has the options of attacking North Korea or not attacking North Korea. Note that this again creates a game with six possible outcomes.

[9] Jelnov, Tauman, and Zeckhauser, "Attacking the Unknown Weapons."

		Nation 1 (North Korea)		
		Stop Nuclear Weapon Research and Open Facilities for Inspection	Stop Nuclear Weapon Research	Continue Nuclear Weapon Research
Nation 2 (United States)	Attack North Korea	*Outcome 1*	*Outcome 2*	*Outcome 3*
	Do Not Attack North Korea	*Outcome 4*	*Outcome 5*	*Outcome 6*

The authors also allow the game to be expanded to include an imperfect intelligence system. This system will advise the United States on the progression of North Korea's nuclear weapons program. Since the system is assumed to be imperfect, it has only some probability of communicating the truth back to the United States as the observer. The authors also allow for a cost to North Korea in opening up its nuclear weapons facilities and allowing inspections. The idea is that North Korea may choose not to allow inspections simply because of the cost that such inspections would impose on it.

Feaver and Niou[10] also included a game between the United States and North Korea. In this game, the United States will have an outlook that is either pragmatic or purist. A pragmatic outlook would mean the United States accepts that North Korea is going to develop nuclear weapons anyway, so the United States should help North Korea to ensure both that North Korea is developing the weapons safely and that the United States can be confident in how far North Korea has advanced in the process. A purist outlook would instead mean the United States believes that no country should be developing nuclear weapons, and it is the responsibility of countries like the United States to stop others from doing so.

North Korea, on the other hand, can have an outlook that is either risk acceptant or risk averse. A risk-acceptant North Korea would understand that nuclear war could very well happen, but that is a risk that North Korea is willing to accept to get what it wants. A risk-averse North Korea would instead say that nuclear war is a very real possibility and not a risk that it is willing to accept for any possible benefit. The main caveat to these outlooks, though, is that neither country necessarily knows the outlook of its opponent.

[10] Feaver and Niou, "Managing Nuclear Proliferation."

For example, just because North Korea is publicly suggesting that it is risk acceptant does not mean that its internal decision-making is not instead risk averse. Similarly, just because the United States is putting up a very pragmatic public front does not mean that its internal decision-making is not instead taking a purist outlook.

Given these outlook possibilities, North Korea has the options to deploy nuclear weapons or not deploy nuclear weapons. The United States has the options of assisting North Korea in developing a safe nuclear arsenal or not assisting North Korea in developing its nuclear arsenal. Since this model is from 1996, it is understandable that some of these actions or strategy options may be outdated/unrealistic in the current day.

Avenhaus et al.[11] present an example of an inspection (or verification) game applied to North Korea. Inspection games involve the verification aspect that is often a part of any nuclear deal. Avenhaus et al. specify two players: the inspector (United States) and the inspectee (North Korea); they wrote this model in general form but use the North Korean nuclear program as one possible example of their game. In the game, North Korea can either comply with all the regulations being placed on its nuclear program as a result of a nuclear deal or it can violate one or more of those regulations. The United States can either thoroughly inspect a site or not thoroughly inspect a site. In other words, how likely is it, if North Korea violates its regulations, that the United States would find the violation? This creates a game with four possible outcomes.

		Inspectee (North Korea)	
		Comply with Regulations	Violate Regulations
Inspector (United States)	Thoroughly Inspect Site	*Outcome 1*	*Outcome 2*
	Do Not Thoroughly Inspect Site	*Outcome 3*	*Outcome 4*

While the majority of the published work on applying game theory to the North Korean nuclear crisis involves relatively simple game setups, some games have been extended to multiple stages, where more than one decision is being made and/or more than one game is being played sequentially within a larger overall game.

[11] Avenhaus et al., "Inspection Games in Arms Control."

One example of a simple two-stage game can be found in Kim and Choi,[12] who describe a game between the United States and North Korea. In the first stage, North Korea chooses either sincere or deceptive diplomacy toward the United States, and the United States chooses to adopt either hawkish or dovish policies toward North Korea. Hawkish policies are more aggressive, military focused, and/or less motivated by compromising to reach a deal. Dovish policies are more focused on peaceful resolutions, compromising, and working toward a deal above all else. In the second stage of the game, North Korea and the United States each independently choose to either work toward Korean reconciliation or reject Korean reconciliation. It is apparent that the decisions made in the first stage will greatly affect the choices and outcomes in the second stage.

First Stage		North Korea	
		Sincere Diplomacy	Deceptive Diplomacy
United States	Adopt Hawkish Policies toward North Korea	*Outcome 1a*	*Outcome 2a*
	Adopt Dovish Policies toward North Korea	*Outcome 3a*	*Outcome 4a*

Second Stage		North Korea	
		Work toward Korean Reconciliation	Reject Korean Reconciliation
United States	Work toward Korean Reconciliation	*Outcome 1b*	*Outcome 2b*
	Reject Korean Reconciliation	*Outcome 3b*	*Outcome 4b*

Another simple two-stage game applied to the North Korean nuclear crisis can be found in the work of Kraig.[13] The first stage of this game is depicted below in the red-outlined box, while the second stage of the game is depicted in the green-outlined box. This game

[12] Kim and Choi, "Uncertainty in Foreign Policy Making."

[13] Kraig, "Nuclear Deterrence."

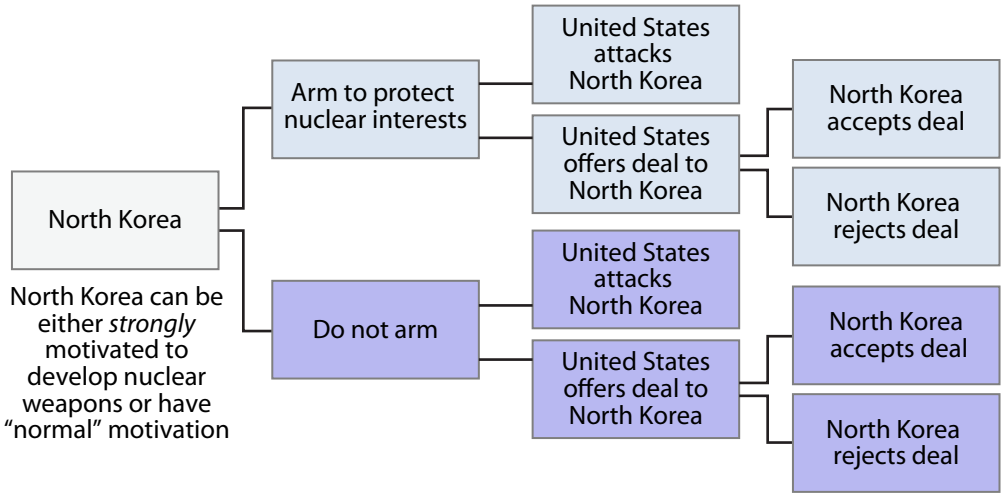
was written generically, between Adversary 1 and Adversary 2. However, the North Korean nuclear crisis was used as an example, with North Korea being Adversary 1 and the United States being Adversary 2.

The general setup of this game requires that the countries first choose between cooperating and defecting, and then only if the country defects will it choose between defect and escalate. In other words, in the first stage, North Korea and the United States will each choose between cooperating (which can also be thought of as conceding) or defecting, which in this case means the country initiates or responds with conventional weapons. In the second stage (which can only happen for a country that has chosen to defect in the first stage), the defecting country or countries will then choose between defect and escalate (here defined as responding with nuclear weapons).

		Adversary 1 (North Korea)		
		Cooperate/ Concede	Defect (Initiate/ Respond with Conventional Weapons)	
Adversary 2 (United States)	Cooperate/ Concede	*Outcome 1*	*Outcome 2*	Escalate (Respond with Nuclear Weapons)
	Defect (Initiate/ Respond with Conventional Weapons)	*Outcome 3*	*Outcome 4*	*Outcome 5*
		Escalate (Respond with Nuclear Weapons)	*Outcome 6*	*Outcome 7*

Games more complex than these two-stage games have also been applied to the North Korean nuclear crisis. Benson and Wen[14] discuss one example. They wrote this game in general form (about Country 1 and Country 2) but state that it applies to North Korea and the United States. The game begins with North Korea having either strong motivation to develop nuclear weapons or "normal" motivation. A country with strong motivation will go out of its way more to successfully proliferate weapons. North Korea can then choose to arm to protect its nuclear interests or not arm. After that, the United States can either attack North Korea or offer a deal to North Korea. If North Korea is offered a

[14] Benson and Wen, "A Bargaining Model."



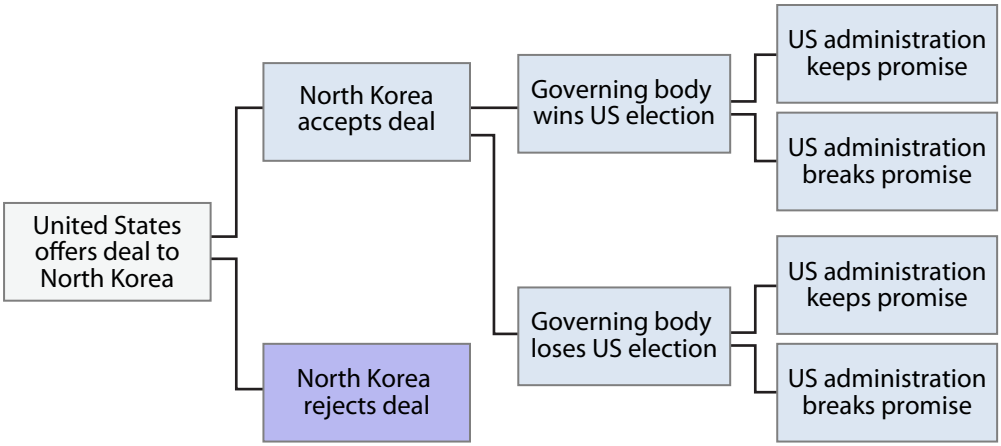
deal, it can either accept or reject this deal. Note that this game is now in extensive (tree) form, rather than the payoff-matrix form seen previously.

Another extended-form game applied to North Korea and the United States can be found in an article by Park and Hirose.[15] Again, the authors wrote this game in general form but stated that it could be applied to North Korea and the United States. The game starts as a negotiation game, where the United States offers a deal to North Korea. North Korea can then either accept or reject the deal. Taking into account the political situation of a country like the United States, if North Korea accepts the deal, the US governing body (i.e., the one that brokered the deal with North Korea) can either win the next US election or lose the next US election. Then, regardless of the election's results, the current governing body/US administration can either keep the promises made in the deal with North Korea or break the promises made in the deal with North Korea. Historically, the recent Iranian nuclear deal shows why these dynamics may be important for a country like North Korea to take into consideration. North Korea would not want to make a deal that will not ultimately be honored.

A final example of an extended-form game applied to North Korea and the United States comes from Bas and Coe.[16] In this game setup, three different games are played in sequential game stages. The authors wrote the game in general form but mentioned North Korea and the United States as possible players.

[15] Park and Hirose, "Domestic Politics."

[16] Bas and Coe, "A Dynamic Theory."

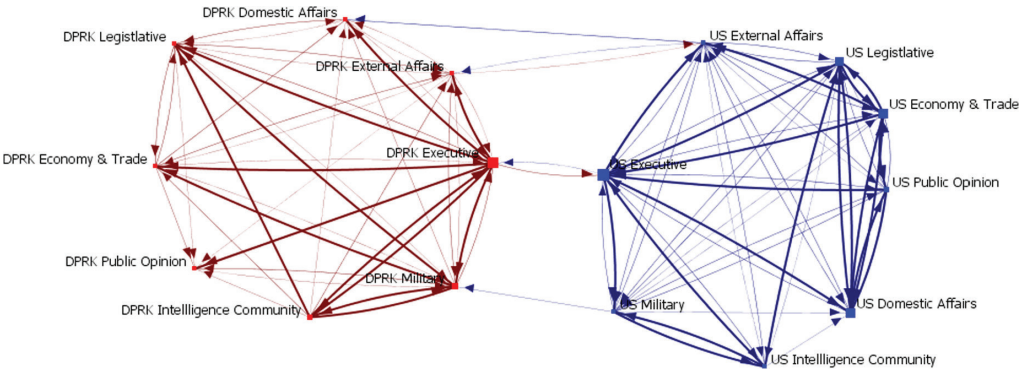


In the game's first stage, a signaling game is played, where North Korea will signal whether or not North Korea plans on investing in its nuclear weapons program. For example, signaling could be in the form of tweets or public news media. The United States will then receive a (possibly inaccurate) signal on North Korea's progress in its nuclear weapons program. This means the United States will have some sense of where North Korea is in its nuclear program, but only with some probability of accuracy. In the second stage, a bargaining game is played, where the United States offers North Korea a deal to stop its nuclear weapons development. North Korea can then either accept or reject this deal. Finally, in the third stage, an investment game is played. Here, if North Korea accepts the offered deal, it can either fully stop development of its nuclear weapons or deceitfully continue with its nuclear weapons program.

All the games discussed so far are bilateral games, meaning they involve only two players. While much can often be learned from these simpler games, the North Korean nuclear crisis certainly involves more than two key players. In particular, many experts believe that any successful North Korean nuclear deal will require the involvement of at least six different countries: the United States, North Korea, China, Russia, Japan, and South Korea. While many of these players have overlapping interests, on the global stage they each function as independent players.

There has been limited work to date in increasing the number of players in game theory models. However, Carley, Morgan, and Lanham[17] present one example. This work uses the graph model for conflict resolution, a newer game theory–based method, implemented using agent-based modeling. In particular, instead of viewing the United States and North

[17] Carley, Morgan, and Lanham, "Deterring the Development."



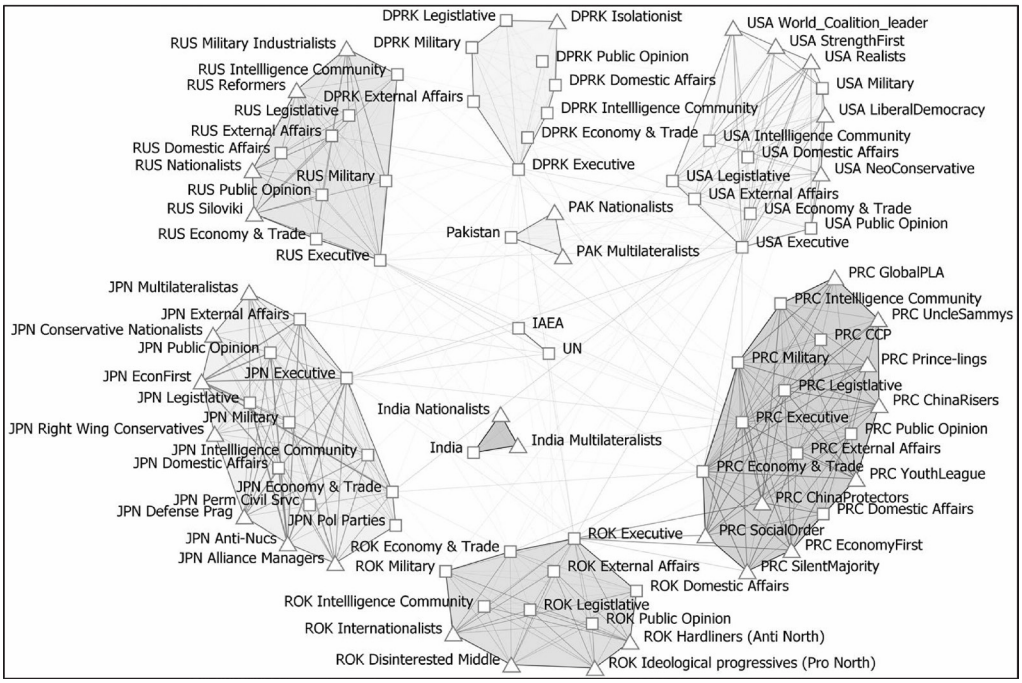
Korea each as one single respective actor, they asked: Who are all the players, within each of the United States and North Korea, that make up the game? This creates an entire network of influence between the many different players.

The figure above depicts the influence network among these various stakeholders in North Korea (the red network) and the United States (the blue network), obtained via much input from subject matter experts. For example, comparing public opinion between the United States and North Korea, US public opinion is an important node with many connections, whereas the North Korean public opinion node is a much sparser, less influential node of the network.

These same authors extended their work in the article by Morgan et al.[18] to include more countries' networks, including all countries with a possible stake in the North Korean nuclear crisis. The subsequent influence network is shown in the figure on the following page, where squares represent stakeholders, and triangles represent national narratives.

Several main takeaways stand out in this literature review. First, game theory will not always be successfully applied to problems in political science or international relations. However, it will be at its best when it can increase understanding by using deductive reasoning to highlight potentially complicated interactions between players. Game theory also excels at tracing the causal effects of (or logic behind) any strategic interactions. In other words, game theory can be used for evaluating the basis for causation, rather than only the correlation that statistical analyses can provide. Game theory also does well at generating specific, empirically testable hypotheses. Even better are the game theoretic papers that then empirically test their own hypotheses, providing some level of verification to their game theory models.

[18] Morgan et al., "Sociocultural Models of Nuclear Deterrence."



In addition, game theory can be useful in explaining unique and/or singular cases, illustrating the logic behind a hypothesis (even when only a limited amount of data is available), and testing hypotheses with qualitative data only. Statistics often fails when only qualitative data and/or a small sample size of data are available. Game theory can be useful as an alternative method in these cases.

Game theory is frequently misunderstood in terms of when it should be applied. Game theory will never be able to successfully answer the question "What should I do?," which is often an important question asked by policy makers. Game theory models are regularly criticized for failing to produce unique empirical insights, even if that is not a priori what a model was intended to do. Finally, when the game theory community is too inward-looking, not engaging enough with the qualitative researchers who may be able to use their modeling results, game theory is often not useful.

In terms of game theory applied to the North Korean nuclear crisis, there has not been much work yet on specifically applying game theory to the North Korean nuclear crisis. In addition, much of the work that does exist is of questionable quality, published in low-impact journals.

Most of the current work in game theory for political science and international relations is in general theory development, rather than games developed to analyze specific

real-world international events like the North Korean nuclear crisis. However, the latter is what policy makers are most interested in: how the general theory can be applied to help in specific situations.

This could potentially be partly a result of publication biases. More good journals are needed for true interdisciplinary work. The length of time it takes to prepare and publish a journal article can make a non-general-theory paper obsolete by the time it gets published. General theory papers also typically get more citations and have higher impact because they have many different applications. This could result in editors preferring (subconsciously or overtly) general theory papers.

Finally, the vast majority of work published on game theory applied to the North Korean nuclear crisis is bilateral, a fact that may hinder the realism of the models. This focus on bilateral games persists despite the situation being very much multilateral (e.g., United States, North Korea, South Korea, China, Russia, and Japan).

Alexandre Debs[19] succinctly summarizes these issues with game theory, noting that while many political science onlookers recognize game theory's benefits, they remain resoundingly skeptical: "Political science has long experienced a love/hate relationship toward game theory, admiring the rigor, deductive logic, and sophistication of game theoretic arguments but questioning its empirical purchase."

References

Avenhaus, Rudolf, Morton Canty, D. Marc Kilgour, Bernhard Von Stengel, and Shmuel Zamir. "Inspection Games in Arms Control." *European Journal of Operational Research* 90, no. 3 (1996): 383–394.

Bas, Muhammet A., and Andrew J. Coe. "A Dynamic Theory of Nuclear Proliferation and Preventive War." *International Organization* 70, no. 4 (2016): 655–685.

Benson, Brett, and Quan Wen. "A Bargaining Model of Nuclear Weapons Development and Disarmament." In *Causes and Consequences of Nuclear Proliferation: A Quantitative-Analysis Approach*, edited by Robert Rauchhaus, Matthew Kroenig, and Erik Gartzke, 45–62. Oxford: Taylor and Francis, 2011.

Carley, Kathleen M., Geoffrey P. Morgan, and Michael J. Lanham. "Deterring the Development and Use of Nuclear Weapons: A Multi-Level Modeling Approach." *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 14, no. 1 (2017): 95–105.

[19] Debs, "The Empirical Promise."

Debs, Alexandre. "The Empirical Promise of Game Theory." *Oxford Research Encyclopedia of Politics*. http://politics.oxfordre.com/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-515.

Feaver, Peter D., and Emerson M. S. Niou. "Managing Nuclear Proliferation: Condemn, Strike, Or Assist?" *International Studies Quarterly* 40, no. 2 (1996): 209–233.

Jelnov, Artyom, Yair Tauman, and Richard Zeckhauser. "Attacking the Unknown Weapons of a Potential Bomb Builder: The Impact of Intelligence on the Strategic Interaction." *Games and Economic Behavior* 104 (2017): 177–189.

Kim, Hee Min, and Jun Y. Choi. "Uncertainty in Foreign Policy Making: A Bayesian Game Analysis of Korea." *Global Economic Review* 31, no. 3 (2002): 25–40.

Kraig, Michael R. "Nuclear Deterrence in the Developing World: A Game-Theoretic Treatment." *Journal of Peace Research* 36, no. 2 (1999): 141–167.

Levi, Nicolas. "Applying Game Theory to North Korea-China Relations." *Journal of Modern Science* 2, no. 33 (2017): 355–366.

Lohschelder, Sarah. "Why North Korean Foreign Policy Is Rational: An Application of Rationality Theories." *Yonsei Journal of International Studies* 9, no. 1 (2017): 56–83.

Morgan, Geoffrey P., Michael J. Lanham, William Frankenstein, and Kathleen M. Carley. "Sociocultural Models of Nuclear Deterrence." *IEEE Transactions on Computational Social Systems* 4, no. 3 (2017): 121–134.

Obeidi, Amer, Keith W. Hipel, and D. Marc Kilgour. "The Role of Emotions in Envisioning Outcomes in Conflict Analysis." *Group Decision and Negotiation* 14, no. 6 (2005): 481–500.

Park, Hwee Rhak. "The Self-Entrapment of Rationality in Dealing with North Korea." *Korean Journal of Defense Analysis* 20, no. 4 (2008): 353–365.

Park, Jeongho. "Application of Risk Dominance Concept and Bayesian Nash Equilibrium for Analysis of Recent Geopolitical Tension between North and South Korea." *International Journal of Game Theory and Technology* 2, no. 2/3 (2016): 1–8.

Park, Jong Hee, and Kentaro Hirose. "Domestic Politics, Reputational Sanctions, and International Compliance." *International Theory* 5, no. 2 (2013): 300–320.

Park, Seongho, Joo Hyun Nho, and Taeseon Yoon. "Quantitatively Modified Two-Level Game Theory and Its Application to Reality: Modified Two-Level Game Theory and the Possibility of Its Application to the North East Asian International Affairs." *Advanced Science and Technology Letters* 84 (2015): 83–87.

Song, Jooyoung. "Understanding China's Response to North Korea's Provocations." *Asian Survey* 51, no. 6 (2011): 1134–1155.

3

GAME THEORY AND THE NORTH KOREAN NUCLEAR CRISIS

During the workshop, eleven speakers presented their work on game theoretic methods applied to the North Korean nuclear crisis. All but three speakers provided Johns Hopkins University Applied Physics Laboratory (JHU/APL) with written summaries of the presented work, and those summaries are included in this chapter. JHU/APL participants summarized two of the presentations because the speakers did not provide summaries. One presentation is not included per the author's request.

Nuclear Crisis Bargaining and Escalation Revisited

Robert Powell[1]

While the theory of brinkmanship has long been widely applied, increasingly new examples of nuclear crises can be studied using this theory, including the North Korean nuclear crisis. A brief intellectual history of nuclear deterrence theory will be discussed, as well as how game theory fits into these studies. Then, the importance of the concepts of incomplete information and credibility will be discussed, including their applicability in game theory being applied to nuclear deterrence. In general, nuclear deterrence theory helps us understand the dynamics of nuclear crises, although not necessarily any one country's specific outcome. A game is then developed and analyzed to help answer the question: What risk is a player willing to run when there is a severely detrimental outcome at stake? Conclusions and implications from this model are discussed, as well as some still-open questions in nuclear deterrence theory.

Presentation Summary

Although the study of brinkmanship is often associated with the Cold War and studying the United States versus the Soviet Union, it can be applied more widely than that—for example, Republicans versus Democrats in Congress debating a governmental shutdown or Germany versus Greece debating the Euro. However, more current nuclear examples can once again be used to study brinkmanship—for example the North Korean nuclear crisis.

First, a brief intellectual history of nuclear deterrence theory will be discussed, along with the game theory revolution of the 1980s. The "golden age" of nuclear deterrence theory was around 1956–1966. During this time, game theory played an often-misunderstood role. Thomas Schelling, Bernard Brodie, Herman Kahn, and many others worked on game theory as it applies to nuclear deterrence theory. Because of all this work, and the prominence of Schelling's work in particular, many people assumed that game theory was the main driver of, and foundational work for, nuclear deterrence theory. In other words, there would be no nuclear deterrence theory without game theory.

However, this viewpoint is wrong. Game theory was not a driver of nuclear deterrence theory. While there was certainly coherent thinking about nuclear deterrence during this time, it was not inherently driven by game theory. Thomas Schelling categorized himself as *not* a technical game theorist. He simply used 2×2 games to illustrate and explain the concepts and insights he was coming up with from his own logic and reasoning. Such

[1] This summary was written by Kelly Rooker (JHU/APL) based on the slides and audio recording of Robert Powell's presentation.

game theory was quite simple, and it was not until later that more sophisticated methods were used to answer nuclear deterrence questions.

When thinking about most qualitative discussions about nuclear crises, one key point is the fundamental role of uncertainty—for instance, the uncertainty of resolve. Another key point is credibility: are threats credible or not? In 1966–1967, the tail end of the golden age, John Harsanyi was the first to develop the tools needed to address the issue of incomplete information. Such incomplete information could be about the other player's motivations, strategy options, resolve, beliefs about the other player, etc. Most escalation models say there will not be escalation unless uncertainty is involved, but during the golden age, game theory was not developed enough to handle games dealing with uncertainty. As such, uncertainty was ignored, leading to relatively unrealistic games.

Harsanyi's work, however, did not have much of an impact until the late 1970s and early 1980s. During this time, scholars developed ways to formalize credibility in games, and Harsanyi's work on incomplete information could be joined with this work. Formalizing credibility included refinements of Nash equilibrium (e.g., perfect Bayesian equilibria, sequential equilibria) and subgame perfection, but the basic idea is it became possible to say when a threat (at least game theoretically) was credible: a player's threat or promise will be credible when it is in that player's own self-interest to actually carry it out. Simply stating a threat or promise can be cheap talk or just blowing smoke, but looking at when it is actually *logical* to carry out the threat or promise holds much more substantive meaning. The game theoretic tools are just formalizing these concepts.

Being able to combine the game theoretic concepts of incomplete information and credibility allowed dynamic interactions to be studied. Players can start with great uncertainty (for example, regarding levels of resolve), and then learn over time to lessen the uncertainty. This work began revolutionizing economics during the 1980s. For example, in graduate-level microeconomics courses in the late 1970s, game theory or equilibria would have been mentioned only a handful of times; just ten years later, 30–50 percent of these entire courses focused on game theory.

By the mid-1980s, this same revolution was being felt in political science, and in nuclear deterrence theory in particular. The first such paper came in 1986 (Barry Nalebuff, "Brinkmanship and Nuclear Deterrence"), with subsequent papers following in the *American Political Science Review* in the late 1980s. In total, these works created at least three distinct models. Of course, in 1989 the Berlin Wall came down and the Cold War ended. Interest in, and funding for, work in nuclear deterrence quickly faded, so there are still very few models or empirical examples of intense nuclear crises. This also resulted in many fundamental questions remaining unanswered. Only in the last few years has

there been renewed interest in nuclear deterrence questions. (Note that there has still been work on nuclear proliferation questions.)

The acquisition of nuclear weapons did not eliminate political conflicts of interest, but people were interested in how having nuclear weapons *affected* any such political contests. Nuclear deterrence theory was one key effort used in answering this question. In particular, nuclear deterrence theory was used to try to explain how political conflicts of interest play out when there are nuclear weapons. When will mutual assured destruction (MAD) happen or be believed? What makes one state more likely to prevail? What makes crises more or less dangerous?

Nuclear deterrence theory is *not* about how a particular state (for instance, the United States) gets its way. Conceptually, this makes sense, because if such a theory did exist, every country would always know what to do to get what it wants. Nuclear deterrence theory instead works by helping us understand the dynamics of nuclear crises, regardless of whether any one country (for instance, the United States) prevails/successfully deters an adversary. This means that nuclear deterrence can work as a framework for thinking about how political conflicts of interest play out, even it does *not* answer one question policy makers may care more about: can a specific country (e.g., the United States) successfully deter a specific adversary (e.g., North Korea)? These are two very different goals.

One important point on rationality is the question of whether North Korea is rational. Claims of irrationality are largely based on the idea that any country willing to stand up to the United States (or be more resolute or take more risks than the United States) must not be rational. Perhaps a rogue state is simply one that is willing to take more risk than the United States, but that does not make it inherently irrational. Rather, this is more a statement on risk propensity.

Given that nuclear deterrence theory is about explaining how political conflicts of interest play out in the shadow of nuclear weapons, what is the fundamental strategic problem inherent in the "shadow of nuclear weapons"? At least regarding MAD, the fundamental problem centers on credibility. If a state can make an adversary's costs to standing firm outweigh the benefits, that state can compel its adversary to back down if the state can make its threat to impose these costs sufficiently credible. However, if both states can make the costs outweigh the benefits, neither state can credibly threaten to deliberately impose these costs on the other. This symmetry is similar to what happens when states achieve second-strike capability. Anytime a state is *certain* that its only options are to suffer some major cost or to give in to its opponent, the state will choose to give in. But *certainty* here is crucial, and how can a state be certain? This means nuclear crises are as much about the players' *credibility* as they are about the nuclear weapons.

For states more or less equal in their capabilities, do their respective nuclear arsenals just cancel each other out, and do the nuclear crises unfold as if there were no nuclear weapons involved? How can fully rational actors use their adversaries' fears of nuclear destruction to exert coercive pressure on that state? Is MAD (and its inherent credibility problem) a useful approximation to the post–Cold War world (for example, with the North Korean nuclear crisis)? While MAD implies all-out nuclear war, the theory would say that a state must raise the costs for its adversary only *so high* that the adversary has no choice but to give in. Such costs could be all-out nuclear war, but in the modern day, the costs are more interpretable as other (smaller) costs.

Thomas Schelling suggested a way that states could exert coercive pressure on each other even though neither could credibly threaten to deliberately launch a massive nuclear attack. First, there must always be something left to chance within the threat—in other words, allowing for the risk that something will become out of control. Second, crises are simply contests of resolve, not necessarily of military strength. While the theory would say that military strength has no impact, this may or may not be the case. Third, opponents then exert coercive pressure by increasing the risk that things will become out of control (i.e., "walking toward the brink"). The result is either the risk leading to catastrophe or the risk becoming so large that one of the players will back down. The uncertainty, though, is what creates this game, since neither player knows exactly when the other player will decide to back down. The two critical elements of this game then are resolve (including what it is) and the role of uncertainty.

Resolve can be thought of as players' "valuations," or in an auction analogy, each player's maximum bid. This means that resolve is a function of the stakes of the crisis: what risk is a player willing to take when there is a severely detrimental outcome at stake?

To formalize these ideas in a game, let there be three possible outcomes for a player: the player could back down, the player could prevail, or things could spiral out of control and end disastrously. Let s represent the payoff of backing down, w the payoff to prevailing, d the payoff if the events end disastrously, and r the risk that things will become out of control if the escalation goes one step further. Note that $w > s > d$.

Analysis of this model shows that a state is willing to take the risk r that a crisis will spin out of control if the payoff to backing down is greater than or equal to the payoff to escalating if the player is sure to prevail. In other words:

$$s \leq rd + (1 - r)w$$

$$r \leq \frac{w - s}{w - d} = R_{max},$$

where R_{max} represents the resolve. Analyzing this further shows that the resolve will increase with any of increased w, increasing d, or decreased s. Intuitively, if a player values winning more, the player is willing to take a bigger risk. If the disaster becomes less costly, the player is also willing to take a bigger risk of it happening. Finally, if giving in becomes more costly, the player is willing to take a bigger risk.

Let R_1, R_2 be two resolves, but of unknown value. If $R_2 < R_1$ and there is no uncertainty in this inequality, there is no risk of war or escalation because the balance of resolves is clear. If there is uncertainty about resolve, but not about the *balance* of resolves, then there is still no risk of war or escalation. As long as the uncertainty surrounding R_2 is still less than R_1, the balance that $R_2 < R_1$ is still clear. However, if the balance of resolves is uncertain, meaning there is some small chance that $R_2 > R_1$, there will be a small risk of war or disaster. If there is instead a large chance that $R_2 > R_1$, there will be a large risk of war or disaster.

There are post–Cold War implications from simple, straightforward models such as these. For example, nuclear deterrence theory can help frame the thinking about possible nuclear crises like the North Korean nuclear crisis. However, the North Korea/United States nuclear dyad does not have the stability that was present between the United States and Soviet Union during the second half of the Cold War.

Finally, some questions in nuclear deterrence theory alluded to here are still open. First, relative military strength plays no role in brinkmanship theory, but is this really how the world works? How much would a country acquiring nuclear weapons embolden a militarily weak state? Second, is the autonomous risk of all-out nuclear war truly the major risk in nuclear crises? What specifically is exerting the coercive pressure in a nuclear crisis? Third, what is the "nuclear revolution"? Is it that MAD is common knowledge? What happens if that assumption is relaxed? Do the models of conventional war now also encompass nuclear confrontations?

Extended Deterrence of North Korea: A Three-Party Game

Stephen Quackenbush

To analyze the ability of the United States and allies to deter North Korea, I apply the Three-Party Extended Deterrence Game. This game was developed in previous work[1] to provide a general model of extended deterrence that considers all three actors in extended deterrence situations: Challenger, Defender, and Protégé. This allows conclusions to be drawn not only regarding deterrence but also about the related areas of alliance reliability and war expansion.

Introduction

Since the end of the Korean War in 1953, fears of a renewed North Korean assault on South Korea have continued unabated. The United States has maintained sizable military forces in South Korea for decades to deter North Korea. As North Korea has developed nuclear weapons and ballistic missiles, concerns about the fragility of deterrence on the Korean Peninsula have grown.

Classical deterrence theory, with its focus on the game of Chicken and brinksmanship, constitutes the conventional wisdom regarding deterrence. Nonetheless, classical deterrence theory is badly flawed.[2] The assumption that conflict is always the worst possible outcome, which is fundamental to classical deterrence theory, needs to be discarded. It has not proven useful for developing logically consistent and empirically accurate theory. Unfortunately, policy discussions regarding deterrence in academia and government have generally been based on classical deterrence theory.

Zagare and Kilgour[3] introduced perfect deterrence theory, which provides a logically consistent alternative to understand the dynamics of deterrence. Furthermore, it is empirically supported.[4] Therefore, perfect deterrence theory provides the most appropriate basis for further theoretical development, empirical testing, and application to policy.[5]

[1] Quackenbush, "Not Only Whether but Whom."

[2] For reviews of classical deterrence theory, see Quackenbush, "Deterrence Theory"; and Zagare, "Classical Deterrence Theory."

[3] Zagare and Kilgour, *Perfect Deterrence*.

[4] Quackenbush, "General Deterrence"; Quackenbush, *Understanding General Deterrence*; and Quackenbush, "Empirical Analyses of Deterrence."

[5] Quackenbush and Zagare,"Modern Deterrence Theory."

To analyze the ability of the United States and allies to deter North Korea, I apply the Three-Party Extended Deterrence Game. This game was developed in previous work[6] to provide a general model of extended deterrence that considers all three actors in extended deterrence situations: Challenger, Defender, and Protégé.

This paper continues as follows. First, I discuss North Korea and international relations in Northeast Asia over recent decades to provide historical context. Then, I lay out the Third-Party Extended Deterrence Game and discuss application of the general game to the specific case of deterring North Korea. I then discuss equilibria of the game, which paves the way for conclusions regarding policy to be drawn.

Three-Party Extended Deterrence Game

To analyze extended deterrence of North Korea, I apply the Three-Party Extended Deterrence Game, which I developed in previous research.[7] The game, shown in the figure on the following page, involves three players: Challenger, Defender, and Protégé. At node 1, Challenger chooses whether to cooperate, attack Protégé, or attack Defender. If Challenger cooperates, the game ends with the *Status Quo* (SQ). Otherwise, the state confronted by Challenger has an opportunity to respond.

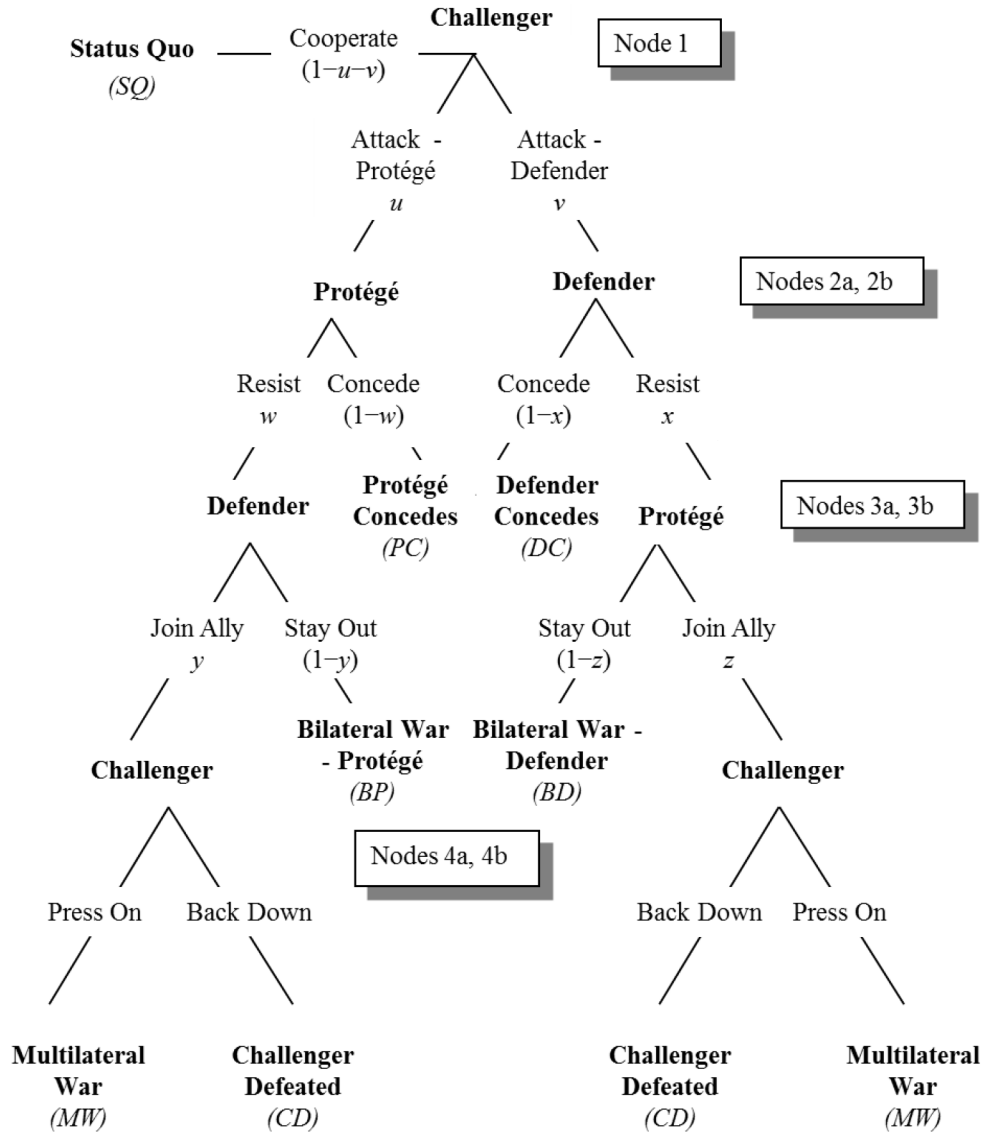
At node 2a, Protégé can either concede or resist. If she concedes, the game ends and the outcome is *Protégé Concedes* (PC). However, if she resists Challenger, Defender chooses next at node 3a. Similarly, Defender may either concede or resist at node 2b; concession leads to *Defender Concedes* (DC); otherwise Protégé moves next at node 3b.

At node 3a, Defender has the option of standing behind Protégé by choosing to join ally or abandoning her by choosing to stay out. If Defender joins Protégé, Challenger moves next at node 4a, but if she stays out, the game ends with *Bilateral War - Protégé* (BP). Should Challenger attack Defender, then Defender and Protégé's roles are reversed. Following Defender's resistance, Protégé has an opportunity to either stay out or join ally at node 3b. If Protégé stays out, the game ends with *Bilateral War - Defender* (BD). But if Protégé chooses to join Defender instead, Challenger chooses next at node 4b.

Nodes 4a and 4b both offer Challenger a choice between press on and back down at the terminal node of the game. If Challenger continues to press on despite Defender's intervention (at node 4a) or Protégé's intervention (at node 4b), a *Multilateral War* (MW) results. However, if Challenger backs down, the game ends with *Challenger Defeated* (CD).

[6] Quackenbush, "Not Only Whether but Whom."

[7] Quackenbush, "Not Only Whether but Whom."



Three-Party Extended Deterrence Game

Each actor's preference ordering over the seven outcomes of the game is fully discussed in previous work.[8] In summary, the players' preferences are as follows:

[8] Quackenbush, "Not Only Whether but Whom."

Challenger: $DC >_{Ch} PC >_{Ch} SQ >_{Ch} BP >_{Ch} BD >_{Ch} [MW, CD]$ (1)

Defender: $CD >_{Def} SQ >_{Def} PC >_{Def} [(BP >_{Def} BD), MW, DC]$ (2)

Protégé: $CD >_{Pro} SQ >_{Pro} DC >_{Pro} [(BD >_{Pro} BP), MW, PC]$, (3)

where the preferences between outcomes enclosed in brackets remain open.[9] Thus, Equation 1 indicates that Challenger's preference between MW and CD is unspecified. Since more preference orderings are left open for Defender and Protégé, Equation 2 and Equation 3 may need further explanation. While the preference ordering for the two types of bilateral war is known—Defender and Protégé each prefer the other to fight (and thus, Defender prefers BP to BD)—the preference between a bilateral war, multilateral war, and concession is determined by the player's type.

Applying the Model to North Korea

The Three-Party Extended Deterrence Game developed by Quackenbush[10] is a general model of deterrence relationships, and is therefore geared toward explaining deterrence in general rather than specific cases. However, we can apply it to specific cases—in particular, deterrence of North Korea—by determining which countries fit which actors in the game model, what their credibility is, and so on. In applying the game to Northeast Asia, it is important to think about how it relates to both extended deterrence and direct deterrence.

The six states involved with Northeast Asian security—those involved in the six-party talks mostly from 2003 through 2007—are North Korea, South Korea, the United States, Japan, China, and Russia. So how do these countries fit in the Three-Party Extended Deterrence Game?

The primary concern that states have with Northeast Asian security is deterring North Korea. Therefore, it is straightforward to equate North Korea as the Challenger, because Challenger is the actor in the game that is being deterred.

As the main target of a potential North Korean attack, South Korea fits well within the model as Protégé. Similarly, the United States fits naturally as Defender since it is engaged in extended deterrence in Northeast Asia, attempting to deter North Korea from attacking South Korea.

[9] For outcome X, Challenger's utility is c_X, Defender's utility is d_X, and Protégé's utility is g_X.

[10] Quackenbush, "Not Only Whether but Whom."

Three of the six countries most relevant to regional security in Northeast Asia are easy to assign roles in the Three-Party Extended Deterrence Game. However, determining appropriate roles within the game for the other three countries is not as straightforward.

Japan is aligned with both South Korea and the United States. As South Korea's ally, Japan partially fits the role of Defender. However, as an American ally, Japan also fits the role of Protégé.

The other states involved in the six-party talks, China and Russia, do not fit in with the game very well, at least with North Korea as the Challenger and South Korea as the Protégé. However, they could fit in the game with the United States and South Korea as the Challenger and North Korea as the Protégé.

Equilibria

Quackenbush[11] fully examines the equilibria of the Three-Party Extended Deterrence Game under conditions of both complete information, when each actor knows with certainty the credibility of the other actors' threats, and incomplete information, when states are uncertain of the other states' intentions. Here, I provide just a summary of the three types of perfect Bayesian equilibria (PBE) that emerge from the incomplete information analysis.

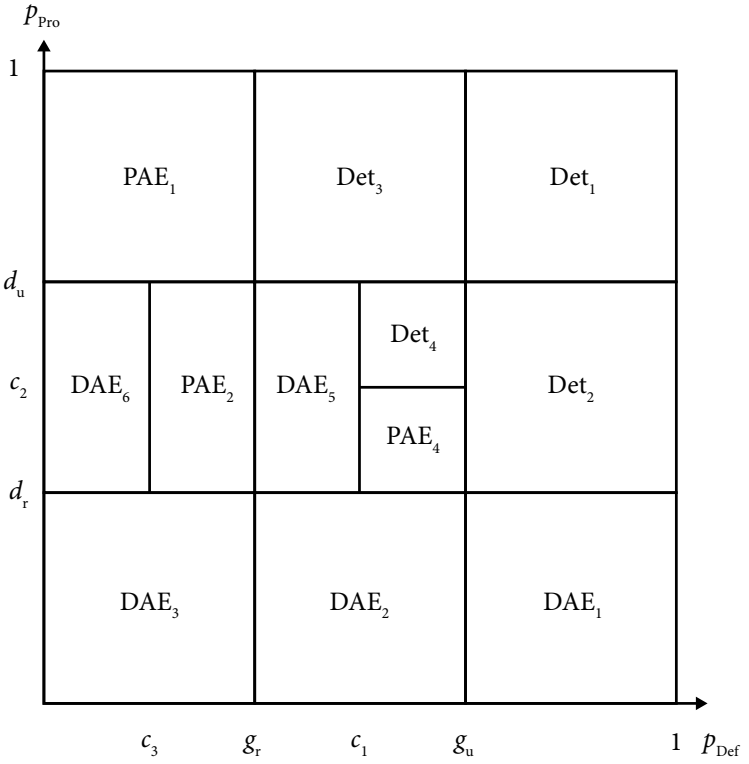
For these analyses, Challenger is assumed to have a credible threat, preferring *Multilateral War* to *Challenger Defeated*, while the others' types are uncertain. There are fourteen perfect Bayesian equilibria that result, separated into three classes: Deterrence Equilibria, Defender Attack Equilibria, and Protégé Attack Equilibria. The locations of equilibria are shown in the figure on the following page.

At a Deterrence Equilibrium (Det), Challenger always chooses to cooperate at node 1. That is, Challenger is always deterred, so he never makes any challenge. Deterrence Equilibria only occur at relatively high levels of alliance reliability. The more reliable Defender and Protégé are, the more likely deterrence is to succeed.

At a Defender Attack Equilibrium (DAE), Challenger always chooses Attack - Defender at node 1. Not only is deterrence assured of failure, but Challenger's target is also assured: it will always be Defender. These equilibria more fully delineate the conditions under which a Challenger will attack the stronger of two parties/allies.

The final class of PBE is called Protégé Attack Equilibria (PAE), where Challenger always chooses Attack - Protégé at node 1. Again, not only will deterrence certainly fail, but there

[11] Quackenbush, "Not Only Whether but Whom."



is also no uncertainty about the target of a challenge: it will always be Protégé. These equilibria delineate when Challenger will confront the weaker of the two allies.

Policy Conclusions

The preceding discussion of the equilibria and their characteristics allows several general conclusions about the dynamics of deterrence to be drawn. First, deterrence is more likely to succeed with a more reliable alliance between Defender and Protégé. Two parameters, d_u and g_u, are crucially important to the success of deterrence because Deterrence Equilibria generally only exist when either Defender or Protégé's credibility parameter (or both) exceeds these thresholds. These parameters are functions of Defender and Protégé's utilities for a concession, a bilateral war, and a multilateral war when unreliable, leading to several observations.[12]

[12] Formally, $d_u = \frac{d_{DC} - d_{BD}}{d_{MW-} - d_{BD}}$ and $g_u = \frac{g_{PC} - g_{BP}}{g_{MW-} - g_{BP}}$.

See Quackenbush, "Not Only Whether but Whom," for a complete discussion of all parameters.

- *The more highly Defender and Protégé value a bilateral war, the more likely deterrence is to be successful.*

This occurs because as the values of d_{BD} and g_{BP} increase, the threshold values d_u and g_u decrease. Accordingly, the region of a deterrence equilibrium becomes larger, making this equilibrium more likely to exist and deterrence more likely to succeed. Thus, while the United States and South Korea would each prefer to fight alongside the other, the more willing each side is to engage North Korea alone, the more likely deterrence is to succeed.

- *The less Defender and Protégé value a concession to Challenger, the more likely deterrence is to be successful.*

This occurs because decreasing values of d_{DC} and g_{PC} lead to decreases in the thresholds d_u and g_u. If either potential target is too eager to concede, this increases the likelihood of their being challenged. But by having lower utility for a concession, each state is more willing to stand up to Challenger, and so deterrence is more likely to succeed.

This lowering of the utility for concession could come through tying one's credibility to a particular issue. Either the United States or South Korea could do this. By reducing the utility of backing down, the United States (or South Korea) could make it less likely that North Korea would initiate a challenge.

- *The more highly Defender and Protégé value fighting a multilateral war when unreliable, the more likely deterrence is to be successful.*

As the values of d_{MW-} and g_{MW-} increase, the threshold values d_u and g_u decrease.[13] By definition, an unreliable ally values a multilateral war less highly than a reliable ally. But the smaller this decrease in utility for multilateral war is (i.e., the smaller the difference between d_{MW+} and d_{MW-} and between g_{MW+} and g_{MW-}), the less reliable each state requires its ally to be before it is willing to resist Challenger. And, the more willing each state is to resist Challenger, the less willing Challenger is to attack, and so the more stable is deterrence.

The parameters c_1 and c_2 are also important to the success of deterrence (equilibria Det_2, Det_3, and Det_4).[14] Note that each of these threshold values depends greatly on Challenger's utility for the status quo (c_{SQ}). Thus, one can observe that:

- *The more highly Challenger values the status quo, the more likely deterrence will succeed.*

[13] The negative sign in these utilities indicates the state's utility from *Multilateral War* when it is an unreliable ally, while a positive sign indicates their payoff when reliable.

[14] Formally, $c_1 = \frac{c_{DC} - c_{SQ}}{p_{Pro}(c_{BD} - c_{MW+}) + (c_{DC} - c_{BD})}$ and $c_2 = \frac{c_{PC} - c_{SQ}}{p_{Def}(c_{BP} - c_{MW+}) + (c_{PC} - c_{BP})}$.

This is because increases in the value of c_{SQ} lead to decreases in the threshold values c_1 and c_2. With a higher utility for the status quo, Challenger is deterred at lower levels of alliance reliability. Notice, however that Defender and Protégé's valuations of the status quo have no bearing on the likelihood of deterrence. This suggests that the United States and South Korea should take steps to ensure that North Korea is more satisfied with the status quo to prevent deterrence failures.

Conclusions about the target of a challenge when deterrence fails can be made in a similar manner. Defender Attack Equilibria are most likely when Protégé's reliability is low but Defender's reliability is high, and conversely Protégé attack equilibria are most likely when Protégé's reliability is high but Defender's reliability is low. Thus, one can conclude that:

- *If there is an imbalance in the reliability of the alliance, the more reliable state is disadvantaged.*

This counterintuitive result is driven by Challenger's desire to avoid a multilateral war. Since a highly reliable state is likely to be involved in any conflict, while an unreliable state will only become involved if attacked, Challenger attacks the more reliable ally.

For example, consider the case where Protégé is highly reliable, but Defender is not. If Challenger attacks Defender, then Protégé is likely to come to Defender's side and a multilateral war ensues. However, if Challenger attacks Protégé, then Defender is likely to step aside. Seeing the second possibility as more favorable, Challenger chooses to attack Protégé, the more reliable of the two allies. Challenger's choice of whom to attack depends importantly on the reliability of the alliance between Defender and Protégé, not just on their relative strength. Therefore, the United States and South Korea should each seek to maximize the other's reliability at the same time as maximizing their own.

- *If Challenger attacks, his target is always certain.*

This conclusion comes from the surprising observation that there are no mixed strategies in equilibrium; either Challenger never attacks, always attacks Defender, or always attacks Protégé. Therefore, given the conditions assumed in this analysis, Challenger is never uncertain about whom to attack.

General Conclusions

Perfect deterrence theory has made great strides in helping us understand the dynamics of deterrence.[15] To examine deterrence of North Korea, I apply the Three-Party Extended

[15] Quackenbush, "Deterrence Theory."

Deterrence Game.[16] This model considers the strategic interactions of all three actors in extended deterrence: Challenger, Defender, and Protégé. This accomplishes three important tasks. First, it provides a formal model that matches the basic informal logic of extended deterrence. Second, it explicitly models the simultaneous conduct of extended and direct deterrence. And third, the inclusion of all three players in the analysis allows examination of other issues such as the target of a challenge and alliance reliability.

Although space constraints prevent a more complete analysis of the subject, this basic application shows one way in which general game theoretic models of international conflict can be applied to consider particular important cases.

References

Quackenbush, Stephen L. "Deterrence Theory: Where Do We Stand?" *Review of International Studies* 37, no. 2 (2011): 741–762.

———. "Empirical Analyses of Deterrence." In *Encyclopedia of Empirical International Relations*, edited by William Thompson. New York: Oxford University Press, 2017.

———. "General Deterrence and International Conflict: Testing Perfect Deterrence Theory." *International Interactions* 36, no. 1 (2010): 60–85.

———. "Not Only Whether but Whom: Three-Party Extended Deterrence." *Journal of Conflict Resolution* 50, no. 4 (2006): 562–583.

———. *Understanding General Deterrence: Theory and Application*. New York: Palgrave Macmillan, 2011.

Quackenbush, Stephen L., and Frank Zagare. "Modern Deterrence Theory: Research Trends, Policy Debates, and Methodological Controversies." In *Oxford Handbooks Online*, edited by Desmond King. New York: Oxford University Press, 2016.

Zagare, Frank C. "Classical Deterrence Theory: A Critical Assessment." *International Interactions* 21, no. 4 (1996): 365–387.

———. *The Games of July: Explaining the Great War*. Ann Arbor: University of Michigan Press, 2011.

Zagare, Frank C., and D. Marc Kilgour. *Perfect Deterrence*. Cambridge: Cambridge University Press, 2000.

[16] Quackenbush, "Not Only Whether but Whom."

Stabilizing Cooperative Outcomes in Nuclear Conflicts: Theory and Cases

Steven Brams and Mehmet Ismail

We describe an alternative notion of stability in normal-form games, which can be represented by a payoff matrix. This notion, called *nonmyopic equilibrium* (NME), differs from the standard notion of *Nash equilibrium* (NE) by assuming that players look ahead to all possible moves and countermoves—and their consequences—when deciding whether to move from an outcome. In such well-known games as Prisoners' Dilemma and Chicken, cooperation is not an NE, but it is an NME.

To illustrate the relevance of NMEs in nuclear conflicts, we analyze games that plausibly model the choices of players in two cases: (1) no first use of nuclear weapons, a policy that has been adopted by some nuclear powers; and, in more detail, (2) the July 2015 agreement between Iran and a coalition of the United States and other countries that forestalled Iran's possible development of nuclear weapons, though it was abrogated by the United States in May 2018. These cases seem applicable to stabilizing cooperation between the United States and North Korea.

Introduction

The standard solution concept in noncooperative game theory is that of Nash equilibrium (NE). However, what might be considered a "cooperative outcome" in a significant number of games is not an NE. The best-known examples of such games are Prisoners' Dilemma and Chicken.

In this paper, we show that cooperative outcomes that are not NEs in almost all 2×2 strict ordinal normal-form games can be stabilized as *nonmyopic equilibria* (NMEs), a dynamic equilibrium concept in which players are assumed to be farsighted when making their choices. It is based on rules of play wherein players start at an outcome (or *initial state*)—rather than with the choice of strategies—and can move or countermove from that state in light of the possible future choices of other players.

If players would not move from an initial state, anticipating all possible moves and countermoves in a game of complete information, that state is an NME. (A state may also be an NME if players would move to it from another state, not just stay at it if they start there.) In Prisoners' Dilemma and Chicken, the cooperative outcome in each game is an NME when play commences at it.

In a few games, however, cooperative outcomes that are not NEs are also not NMEs if play starts there. Fortunately, cooperation in them can be induced by one player's credible

threat of a *Pareto-inferior outcome*, which is worse for all players, if its opponent does not comply with the threat. Thereby we show that in all 2 × 2 ordinal games that have cooperative outcomes that are not NEs, either (1) the nonmyopic stability of a cooperative outcome or (2) one player's credible threat of a worse outcome (for all players) stabilizes cooperation in these games.

More generally, we prove that in all normal-form games (not just two-person), independent of the number of strategies that the *n* players have, there is at least one NME that is *Pareto-optimal*—there is no better outcome for all players—starting from some initial state. Although not all initial states in a game may lead to Pareto-optimal outcomes, this is always the case from at least one state, which almost always leads to a cooperative outcome.

If it does not, credible threats can be used to induce cooperative outcomes in these games. Two cases of nuclear conflict in international relations illustrate the choice of cooperative NMEs that are not NEs.

Two Cases of Nuclear Conflict

No First Use of Nuclear Weapons

The United States was the first to develop, and the only country ever to use, nuclear weapons in war. The two atomic bombs dropped on Hiroshima and Nagasaki in August 1945 brought World War II to a close when Japan surrendered after the second bomb was dropped.

Since the Soviet Union (now Russia) developed nuclear weapons in 1949, seven other countries are now known to possess such weapons (China, France, India, Israel, North Korea, Pakistan, and the United Kingdom). Each has threatened their use if attacked, but three countries (China, India, and Israel) have gone further by declaring that they will not be the first to introduce them into a conflict.[1]

With the exception of China, whose declaration is unqualified, India and Israel have indicated minor qualifications in their declarations. Taken at face value, however, they have pledged no first use of nuclear weapons. All the other nuclear powers have said they would use their weapons only defensively—in retaliation against a nuclear attack—and some have said they would never use such weapons against an attack by a country that did not possess nuclear weapons.

The most serious threat of a nuclear attack occurred during the Cuban missile crisis in October 1962. This confrontation between the Soviet Union and the United States is

[1] Wikipedia, s.v. "No First Use."

sometimes modeled as a game of Chicken, though Brams[2] argues that a different game 2 × 2 is a more accurate representation of choices in the crisis.

In both Chicken and this game, the *cooperative outcome*—at least next best for both players—is not an NE, but it is an NME from itself. Whether this game or Chicken offers a more realistic model of the Cuban missile crisis—or any future nuclear confrontation—the cooperative outcome in both games is an NME from itself. Therefore, if neither player during a confrontation initiates a strike against its foe, cooperation is in the long-term interest of both players, taking into account rational moves and countermoves away from it. Thereby both games offer at least a partial explanation of why antagonisms between nuclear powers, including India and Pakistan in recent decades and the United States and North Korea today, have not escalated to the nuclear level.

The 2015 Iran Agreement on Nuclear Weapons

In 2012, fifty years after the Cuban missile crisis, several countries and the International Atomic Energy Agency feared that Iran might be attempting to develop a nuclear capability that could be used for military purposes. Israel, in particular, believed that Iran was enriching uranium to develop nuclear weapons that could be used against it. It had suspected such surreptitious activities earlier, but in 2012 it claimed they posed an imminent threat to its existence.

Iran denied that developing nuclear weapons was its intention, despite the discovery of previously hidden nuclear-production facilities. It said that it desired to enrich uranium only as an alternative energy source to be used for civilian purposes.

Israeli prime minister Benjamin Netanyahu threatened to attack Iran and destroy its nuclear capability unless there was proof, based on the rigorous inspection of its suspected nuclear facilities, that Iran was not developing nuclear weapons. (A number of Israeli leaders opposed such an attack, arguing that at best it might delay but not stop Iran's acquisition of nuclear weapons.) Israel and Iran were at an impasse, with Iran denying international inspectors access to the facilities in question.

Because of its refusal, Iran suffered ever more severe economic sanctions imposed by the United States, the European Union, and other countries. But a carrot was held out, with the sanctioners offering to relax or lift the sanctions if Iran agreed to allow inspections and credibly commit to halting any efforts that could lead to the production of nuclear weapons. However, a number of countries, including China and Russia, opposed the use of sanctions.

[2] Brams, *Game Theory and the Humanities*; and Brams, "If Trump Doesn't Want a Nuclear War."

The most immediate danger of armed conflict arose from Israel's threat to attack Iran's nuclear-production facilities. More specifically, Israel's position was that, failing an agreement, it would attack Iran's facilities before a point of no return—called a "zone of immunity" by Israeli defense minister Ehud Barak—was reached. That point that would set off an attack would be the time just before these facilities became sufficiently hardened (they were inside a mountain) to be effectively impregnable.

Whether the United States would actively participate in such an attack, or covertly facilitate it, was unclear. On March 8, 2012, president Barack Obama said the United States "will always have Israel's back," which signaled that he was supportive of Israel's concern but did not spell out exactly what the United States would do to aid Israel.

Worth noting is that in the 1970s, Israel began producing, but never publicly acknowledged possessing, nuclear weapons, though it is now presumed to have about eighty nuclear warheads.[3] It did say, however, that it would not be the first party to introduce them into a conflict.

In claiming that Iran's acquisition of nuclear weapons threatened its existence, Israel implied that it would use every means short of nuclear weapons to arrest Iran's development of them if economic sanctions or covert actions failed. The latter actions had included assassinations and cyberwarfare, which had disrupted Iran's enrichment of uranium.

Unlike the superpowers during the Cold War, Israel was unwilling to rely on its own nuclear deterrent—that is, the threat of MAD (mutual assured destruction)—perhaps in part because it feared that terrorists could gain control of Iranian nuclear weapons and act "crazily," without concern for what Israel's response might be. Also, Israel's small physical size made its survival an issue, even after retaliation from an attack, whereas Iran's ability to absorb a retaliatory strike was greater, possibly giving it an incentive to preempt with nuclear weapons.

We present in the following figure a game to model the conflict between Iran and Israel. In this game, Iran chooses between developing (D) or not developing ($\overline{\text{D}}$) nuclear weapons, and Israel chooses between attacking (A) or not attacking ($\overline{\text{A}}$) Iran's nuclear facilities.

We assume Israel's ranking to be $\overline{\text{D}}\,\overline{\text{A}} > \text{DA} > \overline{\text{D}}\,\text{A} > \text{D}\,\overline{\text{A}}$, where the payoffs to the players in the figure are *ordinal* (i.e., order the outcomes from best to worst). As justification, there is little doubt that Israel would most prefer a cooperative solution ($\overline{\text{D}}\,\overline{\text{A}}$), in which Iran does not develop nuclear weapons so no attack is required, and would least prefer that Iran develop nuclear weapons without Israel's making an effort to stop their production

[3] Arms Control Association, "Nuclear Weapons: Who Has What at a Glance."

		Israel	
		Don't Attack: $\overline{A}$	Attack: A
Iran	Don't Develop: $\overline{D}$	Peaceful settlement **(3,4)**	Attack unwarranted (1,2)
	Develop: D	Attack justified but not carried out (4,1)	Attack justified and carried out <u>(2,3)</u>

(x, y) = (ordinal payoff to Iran, ordinal payoff to Israel)
4 = best; 3 = next best; 2 = next worst; 1 = worst
Nash equilibrium (NE) underlined
Nonmyopic equilibrium (NME) in boldface

Iran–Israel Conflict

($D\overline{A}$). Between attacking weapons that are being developed (DA) and mistakenly attacking weapons that are not being developed ($\overline{D}A$), we assume that Israel would prefer the former: an attack on weapons being developed would certainly create a major crisis, but it would be seen by Israel as a measure that was essential to its survival.

As for Iran, we assume that its most preferred outcome is to develop nuclear weapons without being attacked ($D\overline{A}$), and its least preferred outcome is not to develop nuclear weapons and be attacked anyway ($\overline{D}A$). In between, we assume that Iran prefers the cooperative outcome ($\overline{D}\,\overline{A}$) to the noncooperative outcome (DA), which could lead to a major conflict and even war after the attack.

D is a *dominant strategy* for Iran—better for Iran whatever strategy Israel chooses—and yields the unique NE in the figure game (DA). Unfortunately for both countries, this outcome, (2,3), is Pareto-inferior to (3,4), but the strategies that yield (3,4), $\overline{D}\,\overline{A}$, are not an NE.

In July 2015, P5+1—the five permanent members of the UN Security Council plus Germany and the European Union—reached an agreement with Iran for robust inspections of its nuclear facilities that would prevent the significant enrichment of uranium, which could produce nuclear weapons, for fifteen years, as well as several other measures to inhibit Iran's development of nuclear weapons. Although Israel opposed this agreement, the agreement defused a volatile situation that could have led to an Israeli attack, perhaps implicitly if not explicitly supported by the United States.

Why didn't Israel attack earlier, as it continually threatened to do from 2012 to 2015? We suggest that it probably had good intelligence that Iran was not approaching the zone of immunity and, as well, that Iran did not have the capability, or even the intention, of producing nuclear weapons. In that case, Israel would prefer $\overline{A}$ to A.

Iran, we presume, knew that Israel, as well as the United States, could closely track its progress in its nuclear program. While not knowing exactly what these countries knew about its activities, it could predict that Israel, with a high probability, would choose $\overline{A}$. Furthermore, because the Joint Comprehensive Plan of Action provided for the gradual lifting of sanctions if Iran verifiably abided by its commitment not to develop nuclear weapons, there would be benefits in its choice of $\overline{D}$.

In summary, we have suggested that the conflict between Iran and Israel over Iran's possible development of nuclear weapons can plausibly be represented by the game shown in the figure. The cooperative outcome in this game is not an NE, but it is an NME, wherever play starts.

Although it appears that Iran attempted to enrich its supply of uranium before Israel's threat of attacking its nuclear facilities became imminent, as the conflict escalated, it became in Iran's interest to choose not to develop nuclear weapons for two reasons: (1) its fear of an attack on its production facilities; and (2) the continued tightening of economic sanctions, which is not in our game but which was certainly a significant factor in inducing Iran to reach an settlement. When agreement was finally achieved in 2015 after long and arduous negotiations that are detailed in the work of Parsi,[4] it became no longer in Israel's interest to attack or threaten to attack Iran. True, both before and after the 2016 presidential election, Donald Trump disparaged the agreement and has now abrogated it, but so far it appears that it has not been seriously violated.

Conclusions

All normal-form games have at least one Pareto-optimal NME, starting from some state. When play does not start at such a state, or the resulting NME is Pareto-optimal but not a cooperative outcome, credible threats can be used to induce such an outcome.

We discussed two examples in international relations in which the cooperative outcome in games that model recent conflicts were not NEs but NMEs from themselves. To model no first use, we suggested Chicken and another game, both of which have been used to model the Cuban missile crisis, as models of confrontation between nuclear powers in which the cooperative outcome is not an NE but is an NME.

[4] Parsi, *Losing an Enemy*.

Similarly, in the Iran–Israel conflict over the former's possible development of nuclear weapons, the agreement reached in 2015 was a cooperative outcome that we modeled by the game shown in the figure. Before the agreement was reached, however, there was temptation on both sides to escalate the conflict. But Israel's threat of attack as well as economic sanctions made it in the long-term interest of both sides to defuse the confrontation, which nevertheless required difficult negotiations over many months before a compromise was hammered out.

The possibility of nuclear conflict between the United States and North Korea has diminished in the past year, following a serious confrontation between the two sides in October 2017 and especially after a summit meeting between the leaders of the two countries in June 2018.

The strategic logic of no first use seems at least partially responsible for defusing the crisis, whose amelioration could be reinforced by a more explicit declaration.[5] By encouraging farsighted thinking about rational choices in nuclear conflicts, as illustrated by the Iran–Israel game, NMEs offer a path for reaching cooperative, if elusive, settlements.

References

Arms Control Association. "Nuclear Weapons: Who Has What at a Glance." 2016. https://www.armscontrol.org/factsheets/Nuclearweaponswhohaswhat.

Brams, Steven J. *Game Theory and the Humanities*. Cambridge, MA: MIT Press, 2011.

———. "If Trump Doesn't Want a Nuclear War with North Korea, a 'No First Use' Pledge Might Work Better Than Threats." *Washington Post*, October 16, 2017.

———. *Theory of Moves.* Cambridge: Cambridge University Press, 1994.

Parsi, Trita. *Losing an Enemy: Obama, Iran, and the Triumph of Diplomacy.* New Haven, CT: Yale University Press, 2017.

Wikipedia, s.v. "No First Use." Last edited April 17, 2019. 12:11. https://en.wikipedia.org/wiki/No_first_use.

[5] Brams, "If Trump Doesn't Want a Nuclear War."

Bargaining and North Korea

William Spaniel

The North Korean regime faced enormous technical challenges on its path toward building a nuclear weapon. Given that, why did the United States struggle to reach a nonproliferation agreement with North Korea? Although common arguments emphasize lack of trust between the parties, this paper argues that the North Korea's technical incompetency provides an alternative explanation for bargaining failure. In particular, longer development times make the relative patience between the actors more important. When a proliferator is more patient than a nonproliferator and those development times are long, the most a nonproliferator is willing to give up can be less than the minimum that a proliferator must receive. The observable features of US–North Korean negotiations indicate that these constraints were true, thus helping explain why the parties never reached an agreement.

Introduction

From a technical standpoint, North Korea's development of nuclear weapons is a surprise. During the nuclear era, many other countries have faced similar security environments and have had superior nuclear proficiency. Yet, despite North Korea's difficulty in mastering nuclear technology, it has become the tenth country to build a nuclear arsenal, whereas more competent countries have opted against proliferating.

In this paper, I provide an explanation for why the United States struggled to reach an agreement to end North Korea's program. One might initially suspect that longer development times ought to cause less proliferation. After all, a greater delay makes the investment look less attractive. However, the model I create shows the opposite relationship: slower proliferation speeds can cause *more* proliferation.

The key to understanding this counterintuitive result is to consider how patience matters to nuclear negotiations. For an agreement to work, the most an opponent is willing to give up must be greater than the minimum that a potential proliferator requires to accept a deal. If a proliferator is patient, the time delay between the initial investment and the reaped benefits of proliferation matters little. It would therefore need great concessions to forgo developing a bomb. Meanwhile, if its opponent is impatient, that time delay means everything. It does not care very much about the future and therefore would be unwilling to make great concessions to stop proliferation. This can make deals impossible.

Note that the development time is critical here. If development is fast, then the patience gap becomes irrelevant—proliferation will happen soon in the future, meaning that even

an impatient actor still cares about the consequences. Thus, deals only fail when the proliferator is more patient *and* development speeds are slow.

Bringing this theory back to North Korea, the observable features of the situation suggest that the aforementioned bargaining problem was in full effect. Political scientists believe that autocratic governments are generally more patient than democratic governments—they do not face election constraints every few years, and the leaders themselves reign over a longer time period. And while North Korea had the capacity to develop nuclear weapons, their proficiency levels were still relatively low. Combined, these two features indicate that the United States would struggle to reach an agreement with North Korea.

Developing the Model

I now construct a simple model of nuclear negotiations to demonstrate the argument. Consider two actors, *P* and *N*. *P* symbolizes a potential proliferator, or North Korea for this substantive discussion. *N* represents a nonproliferator or a group of nonproliferators, in this case the United States and/or its coalition of partners.

The outcome of the game can end in one of two ways: a negotiated settlement or proliferation. Payoffs for the negotiated settlement are simpler. If the parties reach a deal, a transfer $x \geq 0$ flows from *N* to *P*. *P*'s payoff for this outcome is therefore x. *N*'s payoff is $-x$, reflecting how it pays the transfer. To make this more concrete, one could imagine x representing the basket of inducements the United States might offer to entice North Korea to end its program, ranging from security assurances, to improved diplomatic relations, to trade partnerships, to economic aid.

The proliferation outcome is more complicated. Four components comprise *P*'s payoff: b, δ_P, t, and c. First, $b > 0$ captures all the benefits a proliferator acquires from nuclear acquisition. This ranges from additional coercive ability, increased international prestige, and appeasement of possible domestic factions. However, a country's ability to proliferate is not instantaneous. To capture that, let $\delta_P \in (0, 1)$ represent how much *P* discounts its payoffs from one year to the next and $t \geq 1$ be number of years to successful development of a weapon. Combined with the b value from earlier, the present value of proliferation for *P* equals $\delta_P^t b$. Note that $\delta_P = 1$ means that *P* does not mind waiting at all, whereas $\delta_P = 0$ means that *P* is completely unwilling to wait. Values in between give *P* a moderate level of patience. But also note that for any value of δ_P, increasing t further undermines *P*'s value for proliferation.

The final component of *P*'s proliferation payoff is $c > 0$. This value captures the capital costs of building nuclear weapons and all the manpower that must be sunk into the effort. Combining c with the previous considerations, *P*'s overall payoff equals $\delta_P^t b - c$.

N's proliferation payoff is more straightforward. Following *P*'s development of nuclear weapons, *N* suffers an externality $e > 0$. This value captures any security loss *N* incurs, any damage to the nonproliferation regime as a whole, environmental externalities from testing, and the additional risk of catastrophic nuclear accident. Like *P*'s benefit, *N* does not suffer these externalities immediately. Rather, it takes *t* periods. *N* has its own discount weighting $\delta_N \in (0, 1)$. Combining these elements together, *N*'s proliferation payoff equals $-\delta_N^t e$. Note that if $\delta_P > \delta_N$ *P* is more patient than *N*.

When Are Deals Possible?

Obtaining agreements requires two straightforward conditions: *P* must prefer taking the transfer to developing nuclear weapons, and *N* must prefer giving that transfer to suffering the consequences of proliferation. Formally, the first constraint is:

$$x \geq \delta_P^t b - c. \tag{1}$$

Note that greater values of *t* decrease the right-hand side of Equation 1. That is, longer times to proliferation mean that *P* is willing to accept fewer concessions to give up its program. This captures the central intuition of why North Korea seemed to be an unlikely proliferator.

Meanwhile, the second constraint is:

$$-x \geq -\delta_N^t e$$

$$x \leq \delta_N^t e. \tag{2}$$

Here, however, increasing development times has a countervailing effect. As *t* increases, the right-hand side of Equation 2 decreases. In words, the most *N* is willing to pay goes down. This is also intuitive: when the externalities of proliferation do manifest themselves long into the future, *N* has no need to pay a large settlement to avoid suffering proliferation.

It is not immediately clear which of these effects dominates, and so I press forward. Combining Equations 1 and 2 together, a mutually acceptable *x* requires:

$$\delta_P^t b - c \leq x \leq \delta_N^t e. \tag{3}$$

In words, a settlement requires that the transfer *x* simultaneously be greater than *P*'s minimal demands and less than *N*'s maximal tolerable payment. Such a value exists as long as the minimum is smaller than the maximum. Formally, then, deals exist if:

$$\delta_P^t b - c \leq \delta_N^t e. \tag{4}$$

Proficiency, Patience, and Proliferation

The central proliferation question is how increasing *t* changes whether Equation 4 holds or not. In fact, as the following claims demonstrate, the answer is conditional on the relative patience between the two actors.

Claim 1. *If* N *is more patient than* P, *nonproliferation agreements exist regardless of* t.

From a technical standpoint, claim 1 simply states that Equation 4 always holds if $\delta_N > \delta_P$. A simple intuition explains this. Proliferation is inefficient, because of both the costs of weapons and the externalities they impose. If *N* is more patient than *P*, *N* internalizes those externalities at a higher rate than *P* does. As such, *N* prefers brokering a deal to standing firm and forcing *P* to proliferate. Altering the time to proliferation does not change whether they can reach a deal, just the terms by which they settle.

In contrast, the relationship is more complicated in the opposite case:

Claim 2. *If* P *is more patient than* N, *then the existence of nonproliferation agreements can be nonmonotonic in* t, *with deals impossible for middling development times.*

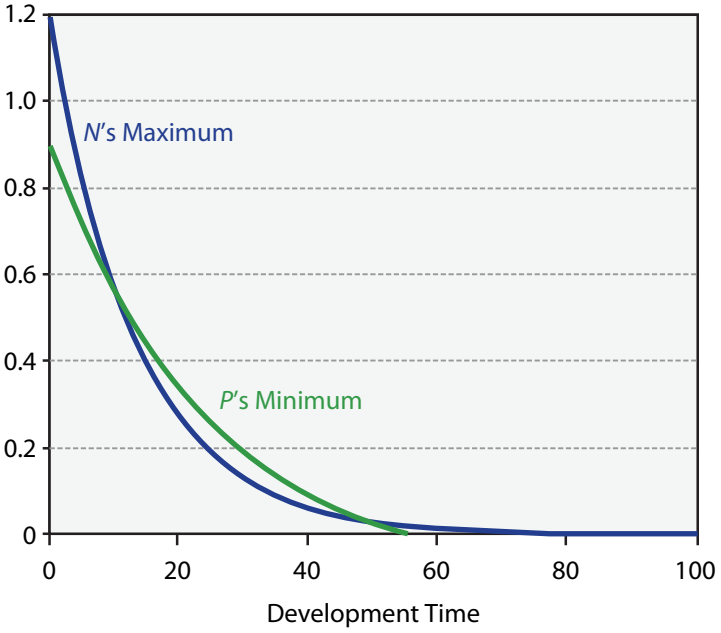
Why does a more patient *P* sabotage the existence of a deal? To halt progress toward a bomb, *N* must make a transfer today to stop a problem it will not actually suffer until later. Thus, when *P* is more patient than *N*, it may view the time-adjusted benefits of proliferation at a greater rate than *N* views the externalities. Indeed, if that patience gap is wide enough, *N*'s best offer may be insufficient despite the costs of nuclear weapons.

Note, however, that claim 2 is conditional on the development time. The two extreme cases clarify why. Suppose that proliferation were instantaneous. Then the patience gap becomes irrelevant, as the consequences of proliferation occur immediately. Meanwhile, when proliferation takes forever, the up-front capital costs eventually override any benefit. *P* becomes willing to accept even the smallest of agreements.

The figure on the following page illustrates the central finding. Deals only work when *N*'s maximum acceptable settlement exceeds *P*'s minimum necessary transfer. This is true for both short and long development times, but not in the middle range. This is where proliferation occurs.

Conclusion

This paper provided an explanation for why North Korea developed nuclear weapons and why American attempts to negotiate a solution ultimately failed. I began with a simple premise: to reach a deal, the most an opponent is willing to give up must be greater than the minimum a potential proliferator needs to forgo nuclear weapons. One would initially



The Constraints on a Negotiated Settlement as a Function of Development Time

suspect that faster speeds to develop a nuclear weapon would make this harder to fulfill because the proliferator's minimum value goes up. But it also increases the opponent's maximum. Thus, developing a model helps adjudicate which effect prevails.

Doing just that, I showed how the relative patience of the actors matters. When the proliferator is more patient than its opponent, longer delays exacerbate the patience gap. Thus, it may be harder to strike a deal with less competent would-be proliferators. These factors provide an explanation for why North Korea proliferated, as autocracies tend to be more patient, and North Korea has limited nuclear capacity.

As compared to other explanations for why negotiations failed, this mechanism strikes a hopeful tone for the future bilateral relationship. In the recent post-proliferation negotiations, a central theme is trust between the two countries. If trust issues caused bargaining failure over proliferation as conventional wisdom may suggest, this indicates that the United States faces an ongoing problem in that domain. But if no deal were possible, then the absence of a proliferation deal says little about the possibility of agreements in other domains.

A Game Theory Analysis of the Stability–Instability Paradox

Barry O'Neill[1]

While the stability–instability paradox is not well defined, its underlying idea is very important: if a conventional war is less likely to lead to nuclear war, states are more likely to get into smaller conflicts. Indeed, there are several historical examples demonstrating the idea behind this paradox, including the India–Pakistan conflict, the Cold War, and—more recently—with North Korea. Despite such real-world applicability, there are still many open questions related to the stability–instability paradox, and game theory could help address them. In particular, game theory is useful in clarifying the premises of a claim or the language of the discussion and/or discovering new strategic phenomena. Using these ideas, a game representing brinkmanship is developed and analyzed, as well as its extension to modeling a war of attrition with caps and/or gaps. In general, there can be inherent issues in applying conclusions from games to the real world. For example, it is often unclear how to determine necessary probability values in game theory models. Probabilities are subjective and open to biases, and policy makers are inherently bad at estimating risk. All these are possible limitations in using game theory to aid policy makers.

Presentation Summary

The majority of game theory work in international relations focuses on generic models rather than specific situations. This workshop is instead asking about applying game theory to the specific situation of the North Korean nuclear crisis. The most admirable attempt to use game theory in modeling specific situations is Nobel Prize winner Reinhard Selten's *The Schwaghof Papers*.[2] Selten uses game theory in a very practical but ingenious way. For instance, Selten allows two countries to form alliances, or generally allows any two players to become one player in a game.

The stability–instability paradox is not particularly well defined, so its general topic is more important than its exact statement. The general stability/balance metaphor can be misleading. Such a metaphor implies that there is some physical activity happening, and all this activity must somehow be equal. However, its general idea is that if a conventional war is less likely to lead to nuclear war, states are more likely to get into smaller conflicts.

[1] This summary was written by Kelly Rooker (JHU/APL) based on the slides and audio recording of Barry O'Neill's presentation.

[2] Abridged in Selten, *Game Theory and Economic Behaviour*.

In other words, stability at the nuclear strategic level can cause instability at lower levels. Think in terms of country dyads (although this could get generalized to include more than two states), with each country in one such pair. Each country is going to be more confident in its ability to deter an attack on its *strategic* interests, which means it will be more ready to fight for its *peripheral* interests.

Practical results of this idea can be found throughout nuclear history. For example, multiple case studies look at the history of India and Pakistan and how these countries fought each other more conventionally *after* obtaining nuclear weapons. Pakistan, for instance, may be more reckless in conventional war approaches because after gaining nuclear weapons, it may believe it is able to deter any nuclear threat. In addition, during the Cold War, there were calls to increase NATO's conventional forces. There were also calls to achieve "escalation dominance," meaning a country having dominance at *every* step of the escalation ladder. In other words, a country cannot let its enemies get comfortable at any lower rung on the escalation ladder, since that would take away the credible threat of escalating to nuclear weapons use, or strategic nuclear war. Historically, this can be seen in the drive toward intermediate-range nuclear force missiles, or any otherwise intermediate nuclear weapon (i.e., in between tactical missiles and strategic intercontinental ballistic missiles). In addition to case studies, there are also a few game theory models, as well as limited statistical analysis studies, regarding these ideas.

Following this logic for the more recent North Korean nuclear crisis leads to a relatively pessimistic conclusion: North Korea will *not* denuclearize. Or even more pessimistic: Japan and/or South Korea will also get nuclear weapons. Either way, the stability–instability paradox would say that any of these actions will cause *more* conflict in Northeast Asia, rather than less, despite the new conflicts not directly involving nuclear weapons. Although there would be no nuclear conflict, there would be an increase in conventional conflict.

The stability–instability paradox has been referenced for decades. Robert Jervis claimed that much of the United States' strategy is governed by the thinking of the stability–instability paradox, something he viewed as largely illogical.[3] The context of the following quote was the Soviet Union had just acquired an H-bomb, and while this may lessen the threat of a World War III between the United States and Soviet Union, it may also put Third World countries at increased risk of more limited conflicts occurring there: "To the extent that the H-bomb reduces the likelihood of full-scale war, it increases the possibility of limited war pursued by widespread local aggression."[4]

[3] Jervis, *Illogic of American Nuclear Strategy*.

[4] Liddell Hart, *Deterrent or Defense*, 23.

Glenn Snyder is often credited with naming the stability–instability paradox and being the first to write it down. (This is most likely not true.) Snyder defined "stability" as the propensity of war happening (i.e., how likely it is for nuclear or conventional war to happen). Indeed, while Snyder acknowledged the stability–instability paradox as being an already widely accepted point of view, he would argue that the *opposite* of the stability–instability paradox was true: "The point is often made in the strategic literature that the greater the stability of the 'strategic' balance of terror, the lower the stability of the overall balance at its lower levels of violence . . . but one could argue precisely the opposite."[5]

In general, there are various ways to advance from peace to strategic war. For example, it could happen accidentally. It could also happen because of emotions (something that should not get overlooked). It could be a more gradual contest of attrition, where both parties stubbornly refuse to drop out or give in to the other, until strategic war is the ultimate result. Finally, it could be the result of a social trap, or the temptation of each country to strike first. Despite two opponents both not wanting to advance to strategic war, strategic war may result because of each country being so convinced that the other is about strike first, but each also strongly desiring first-strike advantage and not being willing to give that to its opponent.

Thinking of the transitions as going from peace to peripheral war to strategic war, is it a paradox or a truism that if going from peripheral to strategic war is *less* likely, going from peace to peripheral war is *more* likely? If it is true, does the nuclear threat have to be made *credible*? Or is it enough to just claim that extended deterrence has been achieved? Why do states even *need* the step of peripheral war? Why do they not just go straight from peace to strategic war? What is the logic behind avoiding peripheral war out of fear of strategic war?

There are many unanswered questions here that game theory models could help address. Note that while strategic war could be nuclear war, it does not necessarily have to be. Strategic war is more generally saying that an involved country has little to gain but very much to lose. Also note that there can be a connection between making strategic war less *likely* versus less *intense*. These can be equated in terms of what countries want: countries first want strategic war to be less likely, but if that does not work, they want strategic war to be less intense.

Models using game theory provide several notable benefits. First, such models can help clarify the premises of a claim. They can help clarify the language or make sure that any assumptions are in fact what they should be to get to some conclusion. Doing so alleviates

[5] Snyder, "The Balance of Power and the Balance of Terror."

the impulse to just state a conclusion as factual, when in reality it does not logically follow. Second, this also helps to clarify the language of the discussion, or even suggest other ways to frame the discussion. Third, game theory models can help discover new strategic phenomena—for example, those regarding escalation discussed below.

One such model involves brinkmanship, or a competition in risk-taking. In the game, two parties each value a certain prize, v and w, respectively. Each player (country) knows its own values and must estimate the other's values (somewhere between 0 and 1). At some unpredictable time, somewhere between 0 and 1, the brink happens and both countries will go over the brink. Each player's decision, then, is how long to stay in the game. Note that the longer a player stays in the game, the more likely it is that the brink will happen. On the other hand, both players want to win by staying in the game longer than their opponent.

From here, payoffs can be assigned. If the brink happens, both players receive $-b$. If both players drop out simultaneously, they receive $v/2$ and $w/2$, respectively. If only one player drops out, whichever drops out first will receive a payoff of 0, while the other player receives v or w, respectively. This is relevant to the stability–instability paradox by studying how changing the payoff $-b$ will influence players' persistence.

A strategy will advise a player on when the player should drop out for any possible value of the prize and regardless of the other player's actions. A Nash equilibrium would then be a pair of strategies such that neither player would be incentivized to change its strategy, even after knowing its opponent's strategy.

Intuitively, the model shows that players will stay in the game longer when the value of their prize is higher (v, w) and/or when the cost of going over the brink is smaller (b). Both of these effects are nonlinear in nature. Relating this back to the stability–instability paradox, the model shows that if a strategic war's expected costs rise, players will spend less time in limited wars, the *opposite* of what the paradox says.

However, one interesting result from this model is that the expected benefit in all cases is equal, meaning the expected benefit over any drop-out time, regardless of the value of b, is the same. The intuitive idea behind this is that humans are able to regulate their actions. For larger potential cost, they stay in for a shorter amount of time; for larger potential benefit, they stay in longer. In other words, the "auction" participants will get the same benefit on average, regardless of the specific costs and benefits of playing in the auction. One extension to this game would be allowing the brink to never happen with some probability. How would the strategies and Nash equilibria change if there were no guarantee of eventually going over the brink?

It is also important to take into account the *likelihood* of going over the brink, and not just the consequences were one to go over the brink. Stability is often defined as the propensity for (or probability of) war, but in practice, the possible severity of war often has just as much impact as the raw probability. Really, "stability" often means the subjective probability of some conditional event. The consequences could be enormous for a nonnuclear power to attack a nuclear power, but if the nonnuclear power is confident enough that the nuclear power would never use nuclear weapons against it, the nonnuclear power can confidently ignore those potentially dire consequences. However, b can be interpreted as the probability of going over the brink *times* the consequences should one go over the brink, solving this issue.

This game can get extended to a war of attrition with caps and/or gaps. In this game, there is some penalty t for staying in the game longer. Two players will still value their prizes v and w, respectively. Each player will stay in the game for some time t, less than or equal to the cap c. Intuitively, this can be thought of as escalating up to higher levels of violence or destruction, which can be quite costly and why the cap exists. Payoffs can be assigned as before. If both players drop out simultaneously at time t, they receive $v/2 - t$ and $w/2 - t$, respectively. If only one player drops out, whichever drops out first will receive a payoff of $-t$, while the other player receives $v - t$ or $w - t$, respectively. This now means that the losing player will not get any benefit, and it will also lose all the resources already invested in the war.

With no cap (meaning c is equal to infinity, or there can be unlimited violence), the Nash equilibrium says players should drop out at time $-v - ln(1 - v)$. Ironically, players will stay in the game for longer than the amount of time at which they value the prize. Because they have already invested so much in it, they will continue playing past where they would have otherwise dropped out, in the hopes of ultimately winning while waiting for the other player to drop out. This illustrates one way in which escalation can be viewed as a trap.

In addition to the regular costs of war, there can also be political costs of war. Politicians may lose a lot by losing a war (in terms of public opinion, party support, etc.). While war is never viewed favorably, it will almost always be viewed *more* favorably after winning than after losing. This also relates to prospect theory, where humans are more willing to not gain something rather than lose something. All these costs help justify the penalty t.

When adding a finite cap c, or saying there is some cap on the level of violence, the stability–instability paradox would suggest that both sides will fight shorter wars. Indeed, this is not the case. Rather, players will rise up to the cap instead of dropping out, assuming other players will be dropping out soon once they reach their cap. This means that adding a cap will actually lengthen wars, rather than shorten them.

Another extension of this game is to also add a gap, where there is some gap in the amount of violence possible. Again, this results in longer wars, where there is more violence than there would be if no such gap existed. In other words, conventional wars are fought harder.

In conclusion, models are able to clarify the logic, but it can still be unclear how to determine the necessary probabilities of such models. Probabilities are not objective, but rather subjective, perceptions. This is in contrast to the assumptions of the stability–instability paradox. Subjective probabilities are very much open to biases. Policy makers are inherently bad at estimating risk. This is not a limitation in the game necessarily, but possibly a limitation in using game theory to aid policy makers.

One possible solution to help defuse the North Korean nuclear crisis would be to promote a new nuclear power, like North Korea, having knowledge of nuclear weapons, sharing weapons safety technology, and helping to understand each other's true probabilities in order to lessen the likelihood of nuclear war. Such a solution could help us avoid some of the mistakes of the Cold War.

References

Jervis, Robert. *The Illogic of American Nuclear Strategy*. 1st ed. Ithaca, NY: Cornell University Press, 1984.

Liddell Hart, B. H. *Deterrent or Defense: A Fresh Look at the West's Military Position*. New York: Praeger, 1960.

Selten, Reinhard. *Game Theory and Economic Behaviour: Selected Essays*. Two volumes. Cheltenham, UK: Edward Elgar Publishing, 1999.

Snyder, Glenn H. "The Balance of Power and the Balance of Terror." In *The Balance of Power*, edited by Paul Seabury, 184–201. San Francisco: Chandler, 1965.

Strategic Consequences of Psychological Factors and Emotional Misrepresentation in Negotiation

Alexandra Mislin

The success of a negotiated agreement depends on implementation and implications for future exchange between the parties. This presentation examines structural, affective, and contractual factors that influence implementation behavior and cooperation after a deal. A series of laboratory studies involving negotiation simulations between participants found that a long-standing notion that feigning anger is an effective bargaining tactic is flawed: feigning anger actually jeopardizes post-negotiation deal implementation and subsequent exchange. Misrepresented anger creates an action–reaction cycle that results in genuine anger and diminishes trust in both the negotiator and the counterpart. If negotiators are to fully understand the economic consequences of costly contract implementation and post-negotiation cooperation, they must look beyond tactical advantages associated with concession patterns to consider the strategic implications of their bargaining behavior after the deal.

Presentation Summary

Negotiated agreements provide the basis for international relations, but the execution of agreements often generates dissatisfaction, disputes, and enmity. Failing to anticipate risks associated with the implementation of a negotiated agreement can turn apparent "wins" at the bargaining table into profound losses after the deal. Formal theories of bargaining[1] have largely ignored these concerns. Although contract theory is a notable exception that has focused on how the terms of an agreement motivate implementation,[2] an implicit underlying assumption of this body of theory is that "the final contract the parties end up signing is independent of the bargaining process leading up to the signature of the contract" and that "the main determinants of contracts are parties' objectives, technological constraints, and outside options."[3]

Yet decades of research have established that psychological factors arising during the bargaining process impact the negotiated agreement. The creation and distribution of value in negotiated agreements is affected not only by the negotiator's objectives and

[1] Kalai and Smorodinsky, "Other Solutions"; Nash, "The Bargaining Problem"; and Rubinstein, "Perfect Equilibrium."

[2] Milgrom and Roberts, "Adaptive and Sophisticated Learning"; Ross, "Economic Theory of Agency"; and Salanié, *Economics of Contracts*.

[3] Bolton and Dewatripont, *Contract Theory*, 7.

constraints but also by cognitive biases,[4] emotions,[5] trust,[6] social motives,[7] and the framing of information.[8] Psychological factors that emerge during the bargaining process even motivate a negotiator's commitment to follow through with the deal and implement the agreement.[9]

Skilled negotiators can employ negotiation tactics that trigger psychological factors that help them claim value at the expense of their counterpart. Some theorists[10] and practitioners[11] believe one such tactic to be emotional misrepresentations—the deliberate expression of an emotion that is different from the one genuinely felt by the negotiator. A stream of recent research has established that sending angry expressions generates, under certain circumstances, increased concession-making from the recipient.[12] When emotions are conveyed either by a computer program or by a confederate, results appear to affirm a long-standing notion that feigning anger is an effective bargaining tactic. The deliberate expression of negative emotion that is different from that genuinely felt by the negotiator offers tactical advantages by making the actor appear tough[13] and gain concessions.[14]

The apparent advantages derived from the strategic expression of anger, however, may disappear or even reverse because of implementation challenges after the negotiating parties leave the bargaining table. If a negotiated agreement cannot be fully specified and implementation behavior cannot be fully monitored, the anger tactic may backfire[15] and

[4] Bazerman and Neale, *Negotiating Rationally*.

[5] Van Kleef, De Dreu, and Manstead, "Interpersonal Effects"; and Carnevale and Isen, "Influence of Positive Affect and Visual Access."

[6] Kong et al., "Interpersonal Trust within Negotiations."

[7] De Dreu et al., "Influence of Social Motives."

[8] Bottom and Studt, "Framing Effects."

[9] Mislin, Campagna, and Bottom, "After the Deal"; and Campagna et al., "Strategic Consequences."

[10] Frank, "If *Homo economicus* Could Choose His Own Utility Function."

[11] For example, Hutson, "The Rationality of Rage"; Machiavelli, *Discourses*; Pacelle and Schmitt, "Last Chapter"; and Sagan and Suri, "Madman Nuclear Alert."

[12] Van Kleef, De Dreu, and Manstead, "Interpersonal Effects"; Van Dijk et al., "A Social Functional Approach"; and Wang, Northcraft, and Van Kleef, "Beyond Negotiated Outcomes."

[13] Van Kleef, De Dreu, and Manstead, "Interpersonal Effects."

[14] Van Dijk et al., "A Social Functional Approach"; Van Kleef, De Dreu, and Manstead, "Interpersonal Effects;" and Wang, Northcraft, and Van Kleef, "Beyond Negotiated Outcomes."

[15] Van Dijk et al., "A Social Functional Approach."

the counterpart may become angry[16] and less trusting.[17] Such tough tactics could thus jeopardize post-negotiation deal implementation and subsequent exchange.

In a paper written with Campagna and colleagues,[18] we directly test both tactical and strategic consequences of emotional misrepresentation. We propose and find that the strategic consequences are due to the reciprocal interdependence between negotiators, which represents an action–reaction cycle with the output of one party becoming the input of the other.[19] This cycle creates a blowback effect when the negotiator's misrepresented anger causes both parties to become genuinely angry and less trusting of each other. Exposure to angry messages prompts the counterpart to react, quite possibly making him or her truly angry[20] and less trusting.[21] Extending the cycle one additional step, the counterpart's reaction to the negotiator's misrepresentation will also affect the negotiator who initially misrepresented.

In a series of four studies,[22] we empirically examined the strategic consequences of misrepresenting emotions. Negotiation simulations were conducted in laboratory studies, with paired participants randomly assigned to a role of employer or candidate, to negotiate over an employment opportunity. Studies involved two-task designs where the first task was a negotiation between an employer and employee, and the second task included a post-negotiation action made by the candidate. Participants assigned to the employer role were randomly assigned to a control condition or a strategic emotion condition where they were coached on the alleged benefits of strategic expressions of anger during a negotiation and incentivized to employ the tactic during their negotiation.

Our studies revealed that negotiators expressing anger both angered and diminished the counterpart's trust. Reflecting counterparty risk, angry subjects tended to renege on their agreement. To better examine the blowback effect, we established reciprocal interdependence through two-sided designs in the subsequent studies. Providing financial incentives induced negotiators to convey emotion-laden messages to their counterparts. The initially false expressions again triggered genuine anger in the counterpart, but also led them to return affect-laden messages to the misrepresenting party. Through this cycle the negotiator misrepresenting anger became genuinely angry in a spiral of diminishing

[16] Van Kleef, De Dreu, and Manstead, "Interpersonal Approach."

[17] Van Kleef and De Dreu, "Longer-Term Consequences."

[18] Campagna et al., "Strategic Consequences."

[19] Thompson, *Organizations in Action*.

[20] Van Kleef, De Dreu, and Manstead, "Interpersonal Approach."

[21] Van Kleef and De Dreu, "Longer-Term Consequences."

[22] Campagna et al., "Strategic Consequences."

trust. The loss of trust impaired agreement implementation and increased rates of outright reneging. These dampening effects persisted even after a time delay, demonstrating that untrustworthy conduct is not quickly forgotten.

Various practitioners and theorists[23] have asserted that misrepresenting emotions can yield material benefits for negotiators. Although recent experiments show that expressing anger increases a counterpart's concession-making, across four studies we found little evidence that such expressions translated into improved terms of agreement.[24] Rather, we found that expressing anger generated significant and consistent strategic disadvantages. Our studies incorporated strategic risk and reciprocal interdependence absent from the majority of prior empirical work on emotional expressions in negotiation. Adding these two elements enabled us to separate short-term tactical consequences from longer-term strategic ones.

Practitioners considering the tactical benefits of anger misrepresentation should think carefully about the wider strategic disadvantages. Losses may encompass much more than foregone future gains.[25] Our research found that the misrepresentation of anger yielded few discernible tactical benefits in negotiation but generated clear and persistent strategic disadvantages.[26] While a negotiator's objectives, constraints, and outside options certainly matter in determining a negotiated agreement, understanding psychological factors arising during the bargaining process is essential to predicting implementation behaviors and the economic consequences of negotiation behavior. Our results suggest the need for great caution in employing tough tactics in negotiations over nuclear stability in Northeast Asia. Advantages that accrue during the negotiation process may disappear or even reverse after the parties leave the negotiation table. Without genuine trust, winning battles may not translate into winning wars.

References

Bazerman, Max H., and Margaret A. Neale. *Negotiating Rationally*. New York: Simon and Schuster, 1993.

Bolton, Patrick, and Mathias Dewatripont. *Contract Theory*. Cambridge, MA: MIT Press, 2005.

[23] For example, Machiavelli, *Discourses*; Pacelle and Schmitt, "Last Chapter"; and Hutson, "The Rationality of Rage."

[24] Campagna et al., "Strategic Consequences."

[25] For example, Pacelle and Schmitt, "Last Chapter"; and Sagan and Suri, "Madman Nuclear Alert."

[26] Campagna et al., "Strategic Consequences."

Bottom, William P., and Amy Studt. "Framing Effects and the Distributive Aspect of Integrative Bargaining." *Organizational Behavior and Human Decision Processes* 56, no. 3 (1993): 459–474.

Campagna, Rachel L., Alexandra A. Mislin, Dejun Tony Kong, and William P. Bottom. "Strategic Consequences of Emotional Misrepresentation in Negotiation: The Blowback Effect." *Journal of Applied Psychology* 101, no. 5 (2016): 605–624.

Carnevale, Peter J. D., and Alice M. Isen. "The Influence of Positive Affect and Visual Access on the Discovery of Integrative Solutions in Bilateral Negotiation." *Organizational Behavior and Human Decision Processes* 37, no. 1 (1986): 1–13.

De Dreu, Carsten K. W., Laurie R. Weingart, and Seungwoo Kwon. "Influence of Social Motives on Integrative Negotiation: A Meta-Analytic Review and Test of Two Theories." *Journal of Personality and Social Psychology* 78, no. 5 (2000): 889–905.

Frank, Robert H. "If *Homo economicus* Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American Economic Review* 77, no. 4 (1987): 593–604.

Hutson, Matthew. "The Rationality of Rage." *New York Times*, September 20, 2015, SR9. http://nyti.ms/1iCP2hR.

Kalai, Ehud, and Meir Smorodinsky. "Other Solutions to Nash's Bargaining Problem." *Econometrica* 34, no. 3 (1975): 513–518.

Kong, Dejun Tony, Kurt T. Dirks, and Donald L. Ferrin. "Interpersonal Trust within Negotiations: Meta-Analytic Evidence, Critical Contingencies, and Directions for Future Research." *Academy of Management Journal* 57, no. 5 (2014): 1235–1255.

Machiavelli, Niccolò. *Discourses*. New York: Oxford University Press, 1987 (originally published in 1519).

Milgrom, Paul, and John Roberts.. "Adaptive and Sophisticated Learning in Normal Form Games." *Games and Economic Behavior* 3, no. 1 (1991): 82–100.

Mislin, Alexandra A., Rachel L. Campagna, and William P. Bottom. "After the Deal: Talk, Trust Building and the Implementation of Negotiated Agreements." *Organizational Behavior and Human Decision Processes* 115, no. 1 (2011): 55–68.

Nash, John F., Jr. "The Bargaining Problem." *Econometrica* 18, no. 2 (1950): 155–162.

Pacelle, M., and Schmitt, R. B. "Last Chapter: A Bankruptcy Lawyer Famed for Theatrics Sees His Job Go Bust." *Wall Street Journal*, February 27, 2002, A1.

Ross, Stephen A. "The Economic Theory of Agency: The Principal's Problem." *American Economic Review* 63, no. 2 (1973): 134–139.

Rubinstein, Ariel. "Perfect Equilibrium in a Bargaining Model." *Econometrica* 50, no. 1 (1982): 97–109.

Sagan, Scott D., and Jeremi Suri. "The Madman Nuclear Alert: Secrecy, Signaling, and Safety in October 1969." *International Security* 27, no. 4 (2003): 150–183.

Salanié, Bernard. *The Economics of Contracts: A Primer*. 2nd ed. Cambridge, MA: MIT Press, 1997.

Thompson, James D. *Organizations in Action: Social Science Bases of Administrative Theory*. New York: McGraw-Hill, 1967.

Van Dijk, Eric, Gerben A. Van Kleef, Wolfgang Steinel, and Ilja Van Beest. "A Social Functional Approach to Emotions in Bargaining: When Communicating Anger Pays and When It Backfires." *Journal of Personality and Social Psychology* 94, no. 4 (2008): 600–614.

Van Kleef, Gerben A., and Carsten K. W. De Dreu. "Longer-Term Consequences of Anger Expression in Negotiation: Retaliation or Spillover?" *Journal of Experimental Social Psychology* 46, no. 5 (2010): 753–760.

Van Kleef, Gerben A., Carsten K. W. De Dreu, and Antony S. R. Manstead. "An Interpersonal Approach to Emotion in Social Decision Making: The Emotions as Social Information Model." *Advances in Experimental Social Psychology* 42 (2010): 45–96.

———. "The Interpersonal Effects of Anger and Happiness in Negotiations." *Journal of Personality and Social Psychology* 86, no. 1 (2004): 57–76.

Wang, Lu, Gregory B. Northcraft, and Gerben A. Van Kleef. "Beyond Negotiated Outcomes: The Hidden Costs of Anger Expression in Dyadic Negotiation." *Organizational Behavior and Human Decision Processes* 119, no. 1 (2012): 54–63.

Nuclear Weapons and Nuclear Risk on the Korean Peninsula: Two Game Theoretic Takes

James Fearon

This paper briefly develops two arguments concerning the conflict over the North Korean nuclear program. Although I will not use explicit game theoretic tools here, the arguments are based on strategic analysis that is facilitated and formalized using game theory.

What Is the Conflict *About*?

After almost seven decades of hot and then cold war between North Korea and its allies, and South Korea and its allies, it is easy to see this conflict as completely normal. Not strange or odd in the least. But if you think about it a bit, it *is* a strange conflict. Or at least these days it is rather strange.

What is strange about it? The North Koreans say, in so many words, that they just want security against attack by the United States. And multiple foreign policy officials of the Trump administration, just as in previous administrations, have said that they have no program for regime change in North Korea. They just want the United States not to be threatened by North Korean nuclear missiles.

In principle, these are *completely compatible positions*. There is no real issue at stake.

Contrast to the Cold War in Europe. There, nuclear risk was seen as arising ultimately from substantive policy conflicts. The Soviets wanted to control all of Berlin, and initially the NATO allies feared that they might want to attack and absorb West Germany (or more). Globally, the United States and Soviets engaged in an ideological competition over allies in the newly independent states.

In the case of North Korea, while it is true that the North and South Korean governments have conflicting preferences over which should rule the whole peninsula, it is highly unlikely that either side thinks that it has much chance of successful military conquest in a ground war. Further, this would be true even if there were no nuclear weapons on the peninsula.[1]

So what explains the North Korean drive to get nuclear weapons, and why is there not an obvious diplomatic deal here? Given that both sides basically just want security against attack by the other, the obvious deal is for the United States to promise not to attack or

[1] As is often noted, South Korea and the United States can be effectively deterred from invading North Korea by the North's conventional missile threat to Seoul.

destabilize Kim Jong Un's regime—including removing economic sanctions—in exchange for Kim Jong Un disarming.

But this is where the core problem driving this conflict, and the nuclear risk associated with it, lies. If Kim Jong Un denuclearizes, the United States, and South Korea, *cannot credibly commit not to "Gaddafi"* him if there is domestic turmoil in North Korea. That is, they cannot commit that they will not support or foment a domestic opposition, or coup makers, if there is some unanticipated shock that occasions mobilization in the North Korean police state. Kim Jong Un sees nuclear weapons as security against external pressure and coercion in this event. The threat to actually use nuclear weapons starts to become credible when a government is facing catastrophe. A nuclear-armed Kim Jong Un facing a domestic uprising would not be treated the way that the United States and NATO treated the Gaddafi regime (providing air support for rebels).

It is probably not only the United States and South Korea that Kim Jong Un seeks insurance against. If not for North Korean nuclear weapons—which are already capable of striking Beijing—China could probably more easily depose Kim Jong Un than the United States and South Korea could. China may be a critical ally today, but not necessarily in the future. It could happen, for instance, that political disorder in Pyongyang could lead China to want to force replacement of Kim Jong Un in favor of a new Chinese ally, to preempt reunification under South Korea. Just as in the last scenario, it is not unreasonable to see Kim Jong Un's nukes as making this less likely than if he could not threaten Beijing with a nuclear strike.[2]

In game theoretic terms, the heart of the US versus North Korean nuclear dispute is this *commitment problem*. Because the United States and South Korea (or China) cannot commit that they will not take advantage of conventional military superiority in a range of future scenarios related to domestic turmoil in North Korea, Kim Jong Un (and his father before him) seek and acquire nuclear capability. This makes *both* sides worse off than in principle they could be if the United States (China, etc.) could somehow commit not to exploit their conventional advantage in the future. But unlike in domestic contexts where parties can sign contracts that are enforceable by a court system (and the police), there is no higher authority that could enforce a promise by, say, the United States to not Gaddafi Kim Jong Un.

In addition to this implication of international "anarchy" (lack of an effective third-party guarantor of agreements between states), the other key component of the commitment

[2] I don't mean to say that these are the only motivations for the North Korean program. There surely is some aspect of wanting to signal strength and nationalist achievement to Kim Jong Un's domestic audience.

problem is *US revisionist preferences*. If the US government truly did not care who ruled in Pyongyang, Kim Jong Un would have no reason to fear that the United States would back rebels against his government. But successive US administrations have *despised* the Kim-family dictatorship—for truly excellent reasons—and Kim Jong Un knows this perfectly well, as did his father.[3] Given that underlying preference, based on dislike of a brutal dictatorship, the United States can't credibly commit against using its extraordinary power to undermine the regime if promising circumstances arise.

What to do then? If US policy continues from the premise that North Korea as a nuclear weapons state is unacceptable, the underlying strategic situation leaves the US government with an impossible policy dilemma. I see basically three options, none of which are likely to succeed in the short or medium term if the objective is genuine denuclearization.

First, the United States can try coercion, escalating militarily and so raising nuclear risk (see below for discussion of how this works). This seemed to be the Trump administration's approach in the fall of 2017, perhaps in part to try to scare China into putting more serious pressure on Kim Jong Un. But the problem here is the "ask" (denuclearization). The more we threaten and raise risk of attack, the more we simply convince Kim Jong Un that he is right, that his only true security lies in being a nuclear weapons state.

Second, if the United States tries to make nice and commit not to Gaddafi him—sign a peace treaty, remove sanctions, remove troops—Kim Jong Un will just keep his weapons and program. Why not, and why believe that the United States would actually support his rule if he became domestically threatened?

It should be stressed that essentially the same commitment problem will rear its ugly head if current negotiations manage to proceed to the point of agreement on first steps toward, say, mothballing a reactor. The United States will (I hope!) want to be able to inspect and monitor to see that any baby steps toward real denuclearization are actually being taken. But the Kim regime will be extremely resistant to any intrusive inspection or monitoring regime, for fear that the intelligence collected could be used against it in various ways. Arms control talks between the United States and Soviets in the 1950s failed in part for this reason. As Khrushchev said to ambassador Averell Harriman when Harriman professed that the United States would not use on-site inspections for spying: "You're trying to tell me that if there is a piece of cheese in the room and a mouse comes in the room the mouse won't go and take the cheese. You can't stop the mouse from going for the cheese."[4]

[3] Recent professions of love notwithstanding.

[4] Fearon, "Bargaining, Enforcement, and International Cooperation."

The first approach—military escalation to raise nuclear risk—is bad because it is pointless, self-defeating, and dangerous. The second approach—acquiescing to North Korea as a nuclear weapons state while pretending not to—is bad because it undermines the core nonproliferation objective, weakens the Non-Proliferation Treaty framework, and ignores some legitimate reasons to prefer that this vicious regime not have nuclear weapons.[5]

A third US foreign policy approach would be to try to return to the "maximum pressure" policy of organizing a united international front that publicly rejects North Korea as a nuclear weapons state, implemented by the strongest sanctions regime that diplomacy can manage. The goal here would not be to compel the regime to denuclearize—that would be nice, but it is not realistic—but rather to keep the pressure on it to increase the odds that it will collapse. Unfortunately that could take a long time. Moreover, after the Singapore meetings, it is not at all clear that the Trump administration could successfully return to a policy of isolating the North Korean regime even if it wanted to. Chinese, South Korean, and Russian policies would probably not allow this (or, the maximum pressure you could get would not be that much).

Where Does Risk of Nuclear Escalation Come from in This Case?

One of the goals of the JHU/APL workshop was to discuss how game theory might help elucidate risks of nuclear escalation in this case. In this section I very briefly sketch how classical nuclear deterrence, with some updating, can be applied to this question.

How might it happen that a nuclear weapon would actually be launched, by North Korea or the United States, if relations were again to turn sour? There are of course many hypothetical possibilities, including a pure technical accident. But probably the most likely scenarios involve what Thomas Schelling famously described as "the reciprocal fear of surprise attack."[6]

Schelling and others worried about the following mechanism. Suppose that two states have large nuclear forces with secure second-strike capabilities. It could still be that each would have the following preference order:

No nuclear war > first strike by me >
simultaneous attacks > second strike by me

Schelling speculated about the following possibility: if I worry that you are worried that I might go first, this increases my incentive to actually launch, to avoid getting the worst

[5] My concern would be less with North Korea using nuclear weapons out of the blue, and more with the possibility that Kim Jong Un comes to see his nuclear capability as a possible covert profit center.

[6] Schelling, *Strategy of Conflict*, chap. 9.

payoff of second strike in a nuclear war. He was concerned that even if it is an equilibrium for us each to expect that the other will *not* launch, in which case peace is assured, it is also a Nash equilibrium for each of us to expect that the other will launch, in which case each wants to launch to avoid going second. Schelling thought that it might happen that in a tense crisis, the fear that the other side might get an erroneous signal that the other side was preparing to launch would lead a state to launch preemptively. And further that the fear of this fear would somehow compound on itself, making preemptive launch a rational decision.

He illustrated the proposed mechanism with the following famous example:

> If I go downstairs to investigate a noise at night, with a gun in my hand, and find myself face to face with a burglar who has a gun in his hand, there is a danger of an outcome that neither of us desires. Even if he prefers to just leave quietly, and I wish him to, there is danger that he may *think* I want to shoot, and shoot first. Worse, there is danger that he may think that *I* think *he* wants to shoot. Or he may think that *I* think *he* thinks *I* want to shoot. And so on. "Self-defense" is ambiguous, when one is only trying to preclude being shot in self-defense.[7]

These days, this kind of problem is arguably in play in the issue of police shootings in the United States. It could be, for example, that police shootings are more likely in districts where gun ownership is more common (other things equal), so that police are more on edge in random encounters, and more apt to misinterpret subtle actions and movements as dangerous because they are on more of a "hair trigger." (Not to say that this is the only thing going on!)

In a 1989 article, Robert Powell used game theoretic methods developed in the 1980s to try to formalize Schelling's argument about this reciprocal fear of surprise attack as a possible generator of nuclear risk in a nuclear crisis.[8] His conclusion was that if each side could end the crisis by ceding whatever was at stake, any reciprocal fear dynamic would be short-circuited. To illustrate, if there is a clear signal of giving up in the police or burglar encounter situations—like putting your hands in the air—this is better than launching. This is true under the assumption that, because we are in a secure second-strike situation, nuclear war/exchange is known to be worse than ceding any little issue at stake.

[7] Schelling, *Strategy of Conflict*, chap. 9.

[8] Powell, "Crisis Stability in the Nuclear Age."

This is where I would argue that the US/North Korea nuclear situation may be different from the US/Soviet scenarios that Schelling, Albert Wohlstetter, and other classical nuclear strategists were understandably preoccupied with. Not that there is some big issue at stake—to the contrary, as argued above, it is not clear that there is any substantive issue at stake in the US/North Korea/etc. conflict where use of force would be a plausible way for either side to change the status quo in the direction it desires.

Rather, what is different is that it is not clear yet if North Korea has a reliable second-strike capability that it could deploy against the US homeland. As a result, it could be reasonable, if very risky, for the US military to attempt a damage-limiting first strike if it received intelligence that it interpreted as indicating that the North Koreans were preparing for a launch. And the North Koreans understand this, that the United States could consider trying to take their nuclear capability out in a major preemptive attack. If so, then we have a situation more like the police shooting or burglar situation: the North Korean military could misinterpret intelligence/signals as indicating US preparation for attack, or an actual attack, in which case it would be in a use-it-or-lose-it situation. Knowing that the North Koreans might be on something of a hair trigger, US intelligence would rationally be more likely to interpret ambiguous intelligence or signals as indicating possible preparation for an attack. Because signals and intelligence can always be fallible, this means that in an intense political crisis—or deliberate manipulation of military moves to try to increase pressure—the danger that innocuous actions are interpreted as attack preparation goes up.

I do not think that this scenario is likely. Nuclear use clearly has very negative payoffs for both sides in this conflict, and both sides know this. So it is probably still the case that caution would prevail, even in the event of some level of false warning or ambiguous intelligence.

But in this relatively "stake-less" conflict, this is the most likely path I can see that might lead to nuclear use.[9] This is in contrast to the "classical" nuclear crisis (like Cuba), where one side is making a specific demand and threatening escalation if the other doesn't comply.

The reciprocal fears problem without secure second-strike forces can be modeled formally with game theoretic tools, and doing so is instructive (I think). I will note in closing that in a model of the problem described above, the difference between going first and second has to be *very* large to get any significant equilibrium risk of nuclear escalation.

[9] Apart from flat-out technical accident or unauthorized launch, which is probably more likely on the North Korean side. Of course, the higher this probability is assessed by the United States, the greater the reciprocal fears dynamics are as well.

References

Fearon, James D. "Bargaining, Enforcement, and International Cooperation." *International Organization* 52, no. 2 (1998): 269–305.

Powell, Robert. "Crisis Stability in the Nuclear Age." *American Political Science Review* 83, no. 1 (1989): 61–76.

Schelling, Thomas. *The Strategy of Conflict*. New Haven, CT: Yale University Press, 1960.

Denuclearization or Not? A Multiple-Player Sequential Game Model

Puyu Ye and Jun Zhuang

The Democratic People's Republic of Korea (DPRK) nuclear crisis is a complex international issue that relates to the security and geopolitics of many countries such as North Korea, the United States, China, Russia, Japan, and South Korea. In this paper, we use a multiple-player sequential game model to analyze the strategic interactions of these players. A multi-attribute (politics, economy, military) utility model is used for each player. To our best knowledge, this is the first game theoretic study for modeling denuclearization decisions in a complex multiple-player scenario. Based on inputs from subject matter experts, we find that denuclearization is an optimal choice for North Korea, followed by providing aid from the United States, Japan, South Korea, China, and Russia. The developed framework provides novel insights to decision-making in this complex and important multilateral issue of denuclearization.

Introduction

The Democratic People's Republic of Korea (DPRK) nuclear crisis is one of the major issues affecting the stability of Northeast Asia. In March 1993, North Korea announced its withdrawal from the treaty on the nonproliferation of nuclear weapons (NPT), which started the crisis. In 2017, with the success of North Korea's sixth nuclear test, DPRK officially became a nuclear power. The problem is no longer whether to proliferate or not, but whether to denuclearize or not. Several countries (the United States, Japan, South Korea, China, Russia) have different positions on the issue. We consider three groups of players:

- From North Korea's perspective, North Korea has the right to safeguard national autonomy and subsistence rights and has the right to develop nuclear weapons for self-defense in the face of the United States' hostile policy toward it. The development of nuclear weapons not only brings security interests but also develops economy.
- From the United States', Japan's, and South Korea's perspectives, North Korea's nuclear weapons pose great threats to the world and its courage to challenge the authority of the international nuclear nonproliferation regime. Their desired solution lies in resolutely achieving North Korea's denuclearization.
- From China's and Russia's perspectives, active involvement in the Korean nuclear crisis could maintain their international influences and the stability of East Asia.

There exists a lot of literature on the causes and consequences of nuclear proliferation.[1] Recently, Levi[2] used Nash equilibrium to study the relations between China and North Korea after North Korea started regular nuclear tests. He and Zhuang[3] studied a sequential game between one government and one terrorist group, where the government moves first by deciding the defense effort and rent offered to the terrorist, and then the terrorist decides attack effort. Different from the literature studying whether to proliferate or not, this paper considers the decision on whether North Korea would denuclearize or not.

On the other hand, few researchers have considered such a game consisting of more than two players. Previous work mainly studies games with two players. Nevertheless, Cimbala[4] states that "the most frequently cited way in which the US and its partners can effectively contest DPRK operations is with multilateral engagement." In one exception, Shan and Zhuang[5] developed a sequential four-player game consisting of a proliferation subgame between two terrorist groups and a subsidization subgame between two governments. To our best knowledge, there is no paper in the literature studying a multiple-player game in the context of denuclearization. To fill the gap, in this paper we develop a novel model among three groups of players. The rest of the paper is organized as follows. The following section describes a sequential game model between three groups of players (North Korea; United States/South Korea/Japan; and China/Russia). The paper progresses to describe how to solve the game and analyze the optimal decisions of each group. This paper then concludes with some future research directions.

The Model

Notation

The following table documents the notation used in this paper, including players, decisions, and parameters.

[1] For example, Crane and Schelling, "Arms and Influence"; Betts, *Nuclear Blackmail*; Powell, *Nuclear Deterrence Theory*; Sagan and Waltz, *Spread of Nuclear Weapons*; Sagan, "Causes of Nuclear Proliferation"; Sagan, "Letter to the Editor on 'Proliferation Pessimism'"; Singh and Way, "Correlates of Nuclear Proliferation"; Hymans, *Psychology of Nuclear Proliferation*; Solingen, *Nuclear Logics*; and Shan and Zhuang, "Subsidizing to Disrupt."

[2] Levi, "Applying Game Theory."

[3] He and Zhuang, "Modelling 'Contracts.'"

[4] Cimbala, in Popp, *How the US Can Work With Its Partners*, 5.

[5] Shan and Zhuang, "Modeling Credible Retaliation Threats."

Notation Used in This Paper

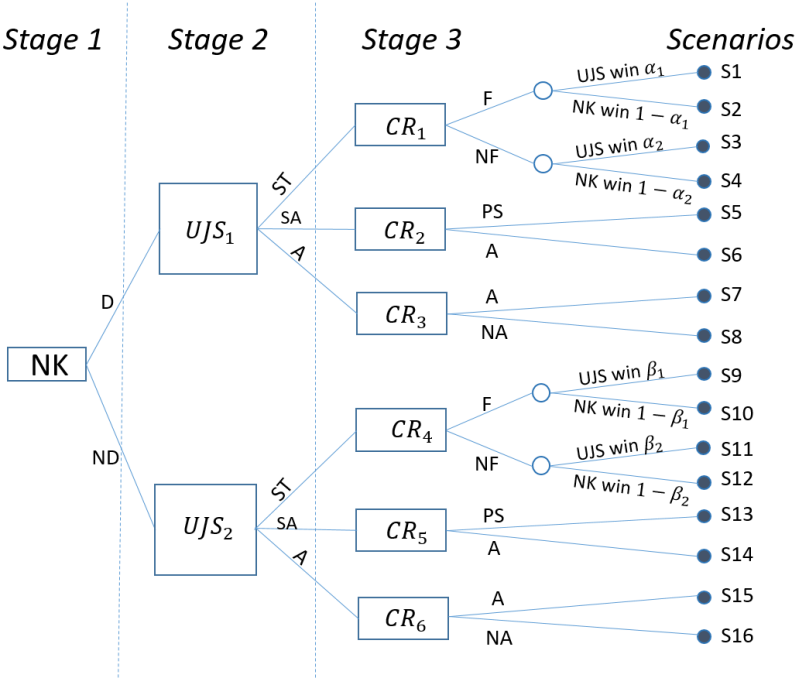
Notation		Explanation
Players	NK	North Korea
	UJS	United States, Japan, and South Korea
	CR	China and Russia
Decision Variables	{D, ND}	NK denuclearizes, or does not denuclearize
	{ST, SA, A}	UJS strikes NK, sanctions NK, or aids NK
	{F, NF}	CR fights or not when UJS strikes NK
	{PS, A}	CR participates in sanction, or aids NK when UJS sanctions NK
	{A, NA}	CR aids or not when UJS aids NK
Parameters	$\alpha_1, \alpha_2, \beta_1, \beta_2$	The probabilities that UJS wins under four circumstances
	$s_1, \ldots, s_{16}$	Sixteen scenarios
	N	The number of attributes for player utility model
	v_{ijs}	The value of attribute j for player i in scenario s
	w_{ij}	The weights of attribute j for player i, $0 \leq w_{ij} \leq 1$, $\sum_{j=1}^{N} w_{ij} = 1, \forall i$
	$u_{i,s}$	$= \sum_{j=1}^{N} w_{ij} v_{ijs}$, the total utility of player i in scenario s
	$U_{*\_*\_*}$	The payoffs under decision chain *-*-*

Model Formulation

The following figure depicts a three-stage game tree. In stage 1, North Korea (NK) decides whether to denuclearize (D) or not (ND). In stage 2, the United States, Japan, and South Korea (UJS) have the options of military strikes (ST), economic sanctions (SA), and economic aid (A). (Although it may be against international morality for UJS to conduct military strikes after NK chooses denuclearization, security and stability might be the key consideration for NK and that scenario may not be ignored.)

In stage 3, if UJS conducts a military attack on NK, CR could choose to fight (F) or not fight (NF). If UJS imposes economic sanctions on NK, CR could either participate in the sanctions (PS) or still offer some financial aid (A). If UJS opts to provide economic aid to NK, CR could either aid (A) or not (NA). Moreover, when UJS conducts a military strike against NK, there are two outcomes: either UJS wins or NK wins. (For simplicity, we do not consider tie or other middle grounds.) We use α_1 to indicate the probability of

UJS winning in the case of NK denuclearizing, UJS imposing military strike on NK, and CR choosing to fight against UJS. Similarly, we use β_1 to indicate the probability of UJS winning in the case of NK not denuclearizing, UJS imposing military strike on NK, and CR choosing to fight against UJS. We use α_2 to indicate the probability of the UJS winning in the case of NK denuclearizing, UJS imposing military strike on NK, and CR choosing not to fight against UJS. Finally, we use β_2 to indicate the probability of UJS winning in the case of NK not denuclearizing, UJS imposing military strike on NK, and CR choosing not to fight against UJS. It is reasonable to assume $\alpha_2 > \beta_2 > \alpha_1 > \beta_1$. The probabilities of NK winning are $1 - \alpha_2 < 1 - \beta_2 < 1 - \alpha_1 < 1 - \beta_1$, in those four circumstances, respectively. (In the following calculations, let $\alpha_1 = 0.5$, $\beta_1 = 0.2$, $\alpha_2 = 0.9$, and $\beta_2 = 0.6$.)



D, Denuclearize; ND, Not Denuclearize; ST, Strike; SA, Sanction; A, Aid; NA, Not Aid; F, Fight; NF, Not Fight; PS, Participate in Sanction.

Overall Illustration of the Game Tree Between Players

The previous figure summarizes all sixteen scenarios. The current state (as of summer 2018) is scenario s_{13}, where NK does not denuclearize, UJS gives economic sanctions, and CR participates in the sanctions. We assume that the payoffs at the current state are zero for each player. That is $u_{NK,s_{13}} = u_{UJS,s_{13}} = u_{CR,s_{13}} = 0$.

Multi-Attribute Utility Model and Data

Based on the inputs from subject matter experts, the following table shows the payoffs for three groups of players in sixteen scenarios, together with the corresponding attribute values for each of the sixteen scenarios. We assume that all attribute values range from −10 to 10.

Player Payoffs $u_{i,s}$ for Each of the Sixteen Scenarios

Scenarios	NK	UJS	CR
s_1	−10 (−10, −10, −10)	4 (3, 3, 6)	−1 (−4, 3, −2)
s_2	−4 (−9, −6, −3)	−7 (−6, −6, 9)	7 (3, 9, 9)
s_3	−10 (−10, −10, −10)	6 (3, 9, 6)	−1 (−3, 3, −3)
s_4	−5 (−5, −9, −3)	−7 (−6, −6, 9)	5 (3, 9, 3)
s_5	−2 (0, −4, −2)	2 (0, 3, 3)	2 (0, 0, 6)
s_6	−1 (3, −4, −2)	2 (0, 3, 3)	3 (0, 0, 9)
s_7	4 (10, −4, 6)	6 (3, 10, 5)	6 (3, 10, 5)
s_8	3 (7, −4, 6)	7 (3, 10, 8)	5 (3, 9, 3)
s_9	−9 (−10, −7, −10)	4 (3, 0, 9)	−5 (−4, −4, 7)
s_{10}	0 (−9, 3, 6)	−7 (−6, −9, −6)	3 (3, 3, 3)
s_{11}	−9 (−10, −7, −10)	7 (3, 9, 9)	−2 (−3, 0, −3)
s_{12}	1 (−9, 3, 9)	−7 (−6, −9, −6)	1 (3, 3, −3)
s_{13}	0 (0, 0, 0)	0 (0, 0, 0)	0 (0, 0, 0)
s_{14}	2 (3, 3, 0)	−1 (0, 0, 3)	1 (0, 0, 3)
s_{15}	6 (10, 0, 8)	−2 (0, 0, −6)	−1 (0, 0, −3)
s_{16}	5 (8, 0, 7)	−0.5 (0, 0, −1.5)	−2 (0, 0, −6)

(with three attributes' values v_{ijs}: economy, military, and politics)

Taking scenario 1 as an example (i.e., NK decides to denuclearize, UJS strikes NK, CR fights against UJS, and ends up with the victory of UJS), the first row of the table shows the attribute values for each of the three groups of players in scenario 1 that are used to calculate the payoffs. Because s_1 is the worst possible scenario for NK, we put $v_{NK,Eco,s_1} = v_{NK,Mil,s_1} = v_{NK,Pol,s_1} = -10$. For each of the players, we use a linear multi-attribute utility model.

$$u_{is} = \sum_{j=1}^{N} w_{ij} v_{ijs}, \ \forall i,s. \tag{1}$$

As the baseline, we set $w_{ij} = \frac{1}{3}, \ \forall i,j$. In particular, for scenario 1, using the attribute values from the table, we have:

$u_{NK,s_1} = \frac{1}{3}(-10-10-10) = -10$,

$u_{UJS,s_1} = \frac{1}{3}(3+3+6) = 4$, and

$u_{CR,s_1} = \frac{1}{3}(-4+3-2) = -1$.

Analysis

To solve the game specified in the overall illustration figure, we use backward induction to solve for the stage 3 decision first, and then stages 2 and 1, respectively.

Stage 3 Decision Analysis for China and Russia

First, we analyze the decisions made by China and Russia (CR) in stage 3. The following figure shows that there are six decision nodes (denoted as CR_1, ... , CR_6). Each decision node has two options.

Considering the first branch (D-ST-F; that is, NK denuclearizes, UJS military strikes NK, and CR fights against the UJS), there will be two potential outcomes. Either UJS wins with a probability of α_1 (scenario 1, CR's payoff is: $u_{CR,s_1} = -1$) or NK wins with a probability of $1 - \alpha_1$ (scenario 2, CR's payoff is: $u_{CR,s_2} = 7$). Therefore, the expected payoff of the first branch (D-ST-F) for CR is:

$U_{D-ST-F} = \alpha_1 \times u_{CR,s_1} + (1-\alpha_1) \times u_{CR,s_2} = -1 \times 0.5 + 7 \times 0.5 = 3$.

Similarly, we have:

$U_{D-ST-NF} = \alpha_2 \times u_{CR,s_3} + (1-\alpha_2) \times u_{CR,s_4} = -1 \times 0.9 + 5 \times 0.1 = -0.4$,

$U_{D-SA-PS} = u_{CR,s_5} = 2$,

$U_{D-SA-A} = u_{CR,s_6} = 3$,

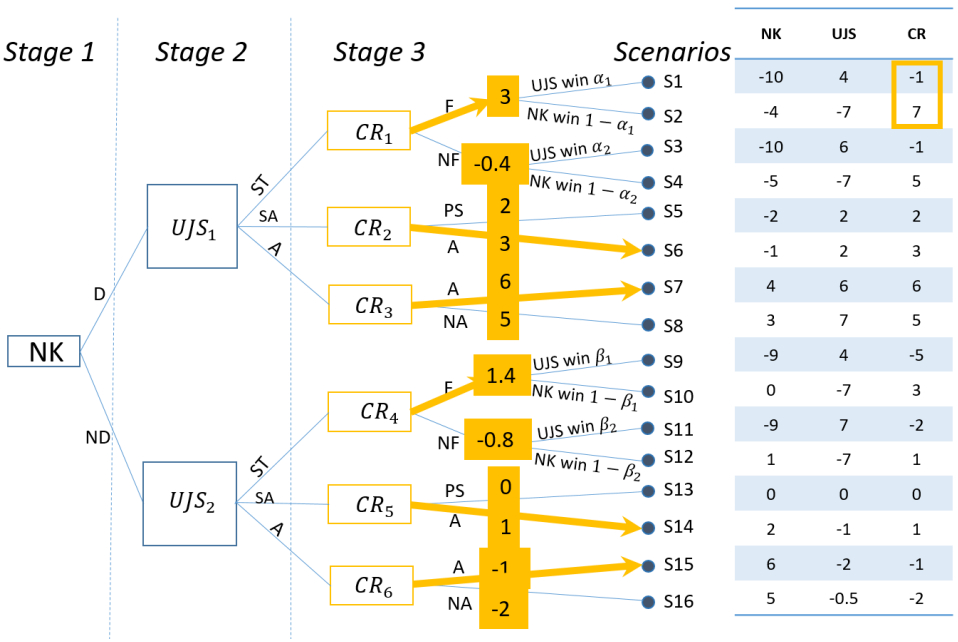
$U_{D-A-A} = u_{CR,s_7} = 6$,

$U_{D-A-NA} = u_{CR,s_8} = 5$,

$U_{ND-ST-F} = \beta_1 \times u_{CR,s_9} + (1-\beta_1) \times u_{CR,s_{10}} = -5 \times 0.2 + 3 \times 0.8 = 1.4$,

$U_{ND-ST-NF} = \beta_2 \times u_{CR,s_{11}} + (1-\beta_2) \times u_{CR,s_{12}} = -2 \times 0.6 + 1 \times 0.4 = -0.8$,

$U_{ND-SA-PS} = u_{CR,s_{13}} = 0$,



NK	UJS	CR
-10	4	-1
-4	-7	7
-10	6	-1
-5	-7	5
-2	2	2
-1	2	3
4	6	6
3	7	5
-9	4	-5
0	-7	3
-9	7	-2
1	-7	1
0	0	0
2	-1	1
6	-2	-1
5	-0.5	-2

D, Denuclearize; ND, Not Denuclearize; ST, Strike; SA, Sanction; A, Aid; NA, Not Aid; F, Fight; NF, Not Fight; PS, Participate in Sanction.

Payoffs and Optimal Decisions in Stage 3 for China and Russia

$U_{ND\text{–}SA\text{–}A} = u_{CR,s_{14}} = 1$,

$U_{ND\text{–}A\text{–}A} = u_{CR,s_{15}} = -1$, and

$U_{ND\text{–}A\text{–}NA} = u_{CR,s_{16}} = -2$.

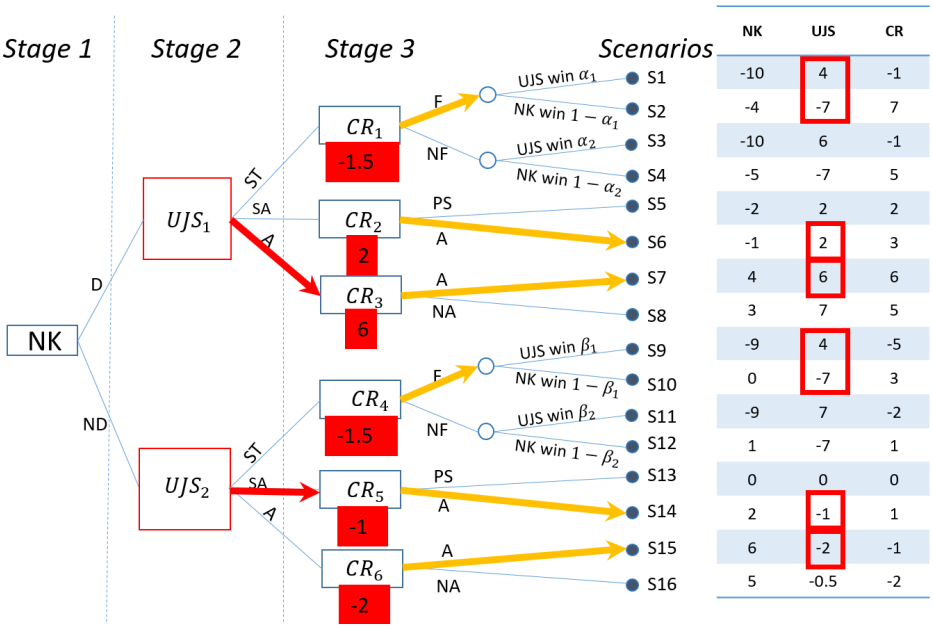
Comparing the two expected payoffs at each of the six decision nodes ($CR_1, \ldots, CR_6$), we get the optimal choices, which are marked in bold yellow in the previous figure and as below:

- At CR_1, the optimal payoff is $u_{CR_1} = \max\{U_{D\text{–}ST\text{–}F}, U_{D\text{–}ST\text{–}NF}\} = 3$, and the optimal decision is Fight (F).
- At CR_2, the optimal payoff is $u_{CR_2} = \max\{U_{D\text{–}SA\text{–}A}, U_{D\text{–}SA\text{–}PS}\} = 3$, and the optimal decision is Aid (A).
- At CR_3, the optimal payoff is $u_{CR_3} = \max\{U_{D\text{–}A\text{–}A}, U_{D\text{–}A\text{–}NA}\} = 6$, and the optimal decision is Aid (A).
- At CR_4, the optimal payoff is $u_{CR_4} = \max\{U_{ND\text{–}ST\text{–}F}, U_{ND\text{–}ST\text{–}NF}\} = 1.4$, and the optimal decision is Fight (F).

- At CR_5, the optimal payoff is $u_{CR_5} = \max\{U_{ND\text{-}SA\text{-}A}, U_{ND\text{-}SA\text{-}PS}\} = 1$, and the optimal decision is Aid (A).
- At CR_6, the optimal payoff is $u_{CR_6} = \max\{U_{ND\text{-}A\text{-}A}, U_{ND\text{-}A\text{-}NA}\} = -1$, and the optimal decision is Aid (A).

Stage 2 Decision Analysis for United States, Japan, and South Korea

Next, we analyze the decisions made by the United States, Japan, and South Korea (UJS) in stage 2, by considering the optimal decision of China and Russia in stage 3. The following figure shows that there are two decision nodes related to UJS (denoted as UJS_1, UJS_2), and each decision node has two options.



NK	UJS	CR
-10	4	-1
-4	-7	7
-10	6	-1
-5	-7	5
-2	2	2
-1	2	3
4	6	6
3	7	5
-9	4	-5
0	-7	3
-9	7	-2
1	-7	1
0	0	0
2	-1	1
6	-2	-1
5	-0.5	-2

D, Denuclearize; ND, Not Denuclearize; ST, Strike; SA, Sanction; A, Aid; NA, Not Aid; F, Fight; NF, Not Fight; PS, Participate in Sanction.

Payoffs and Optimal Decisions in Stage 2 for United States, Japan, and South Korea

Consider the first branch (D-ST; that is, NK denuclearizes and UJS conducts a military strike on NK). According to the analysis of decision node CR_1 in the Stage 3 Decision Analysis for China and Russia section, it is the best choice for CR to fight against UJS. Then the complete decision chain is D-ST-F. There will be two outcomes: UJS wins with a probability of α_1 (scenario 1, UJS's payoff is: $u_{UJS,s_1} = 4$) or NK wins with a probability of

$1-\alpha_1$ (scenario 2, UJS's payoff is: $u_{UJS,s_2}=7$). So the expected payoff of the first branch (D-ST) for UJS is: $U_{D-ST}=\alpha_1\times u_{UJS,s_1}+(1-\alpha_1)\times u_{UJS,s_2}=4\times 0.5-7\times 0.5=-1.5$.

Similarly, we have:

$U_{D-SA}=u_{UJS,s_6}=2$,

$U_{D-A}=u_{UJS,s_7}=6$,

$U_{ND-ST}=\beta_1\times u_{UJS,s_9}+(1-\beta_1)\times u_{UJS,s_{10}}=4\times 0.2-7\times 0.8=-4.8$,

$U_{ND-SA}=u_{UJS,s_{14}}=-1$, and

$U_{ND-A}=u_{UJS,s_{15}}=-2$.

Comparing the expected payoffs for United States, Japan, and South Korea at each of the two decision nodes UJS_1 and UJS_2, we get the optimal choices, which are marked in bold red in the previous figure and as below:

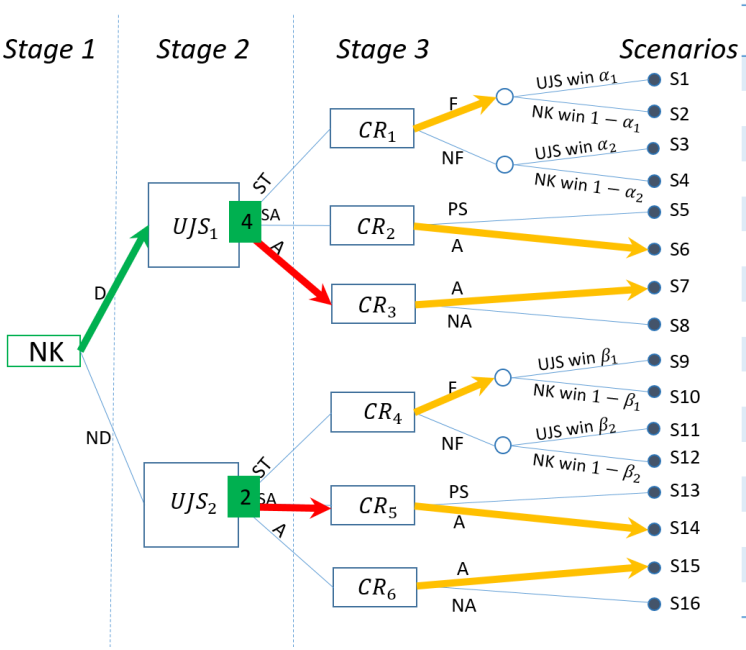
- At UJS_1, the optimal payoff is $u_{UJS_1}=\max\{U_{D-ST},U_{D-SA},U_{D-A}\}=\max\{-1.5,2,6\}=6$, and the optimal decision is Aid (A).
- At UJS_2, the optimal payoff is $u_{UJS_2}=\max\{U_{ND-ST},U_{ND-SA},U_{ND-A}\}=\max\{-4.8,-1,-2\}=-1$, and the optimal decision is Sanction (SA).

Stage 1 Decision Analysis for North Korea

We now analyze the payoffs of decisions made by North Korea (NK) in stage 1, considering the optimal decision of UJS in stage 2, and the optimal decision of CR in stage 3. The following figure shows that there is only one decision node related to NK with two branches: denuclearize or not.

If NK denuclearizes, according to the analysis of decision nodes UJS_1 in stage 2 and CR_3 in stage 3, UJS would provide assistance to NK, and CR would also assist NK. The complete decision chain is D-A-A, which corresponds to scenario 7 in which NK's payoff is: $U_D=u_{NK,s_7}=4$.

If NK does not denuclearize, according to the analysis of decision nodes UJS_2 in stage 2 and CR_5 in stage 3, UJS would impose economic sanctions on NK, but CR would choose to assist NK. The complete decision chain is ND-SA-A, which corresponds to scenario 14 in which NK's payoff is: $U_{ND}=u_{NK,s_{14}}=2$.



NK	UJS	CR
-10	4	-1
-4	-7	7
-10	6	-1
-5	-7	5
-2	2	2
-1	2	3
4	6	6
3	7	5
-9	4	-5
0	-7	3
-9	7	-2
1	-7	1
0	0	0
2	-1	1
6	-2	-1
5	-0.5	-2

D, Denuclearize; ND, Not Denuclearize; ST, Strike; SA, Sanction; A, Aid; NA, Not Aid; F, Fight; NF, Not Fight; PS, Participate in Sanction.

Payoffs and Optimal Decisions in Stage 1 for North Korea

Comparing the two expected payoffs for North Korea, we get the optimal choice for *NK*, which is marked in bold green in the previous figure and as below:

- At decision node *NK*, the optimal payoff is
  $u_{NK} = \max\{U_D, U_{ND}\} = \max\{4, 2\} = 4$,
  and the optimal decision is to Denuclearize (D).

Equilibrium Path

Based on the above analysis, we conclude that the equilibrium path would be D-A-A; that is, North Korea chooses denuclearization and the United States, Japan, and South Korea provide economic assistance, while China and Russia also provide aid for North Korea.

Conclusion and Future Research Directions

This paper uses game theory and multi-attribute utility theory to provide a framework to study DPRK's denuclearization decisions. James Platte[6] states that "the United States can

[6] Platte, in Popp, "How the US Can Work With Its Partners," 13.

best contest North Korean operations by closely coordinating with partners and allies in the region to deter by denial and by conducting an information campaign to help empower North Korean people." This paper's results concur with that conclusion: the United States, Japan, South Korea, China, and Russia provide aid to North Korea when it denuclearizes, which leads to a win–win situation.

In future research, it may be worth studying the subgames between the United States, Japan, and South Korea, as well as subgames between Russia and China. We could also consider some continuous-level decisions, such as the timing, method, and level for potential military strikes and economic sanction and/or aid. Repeated interactions, risk preferences (e.g., risk seeking, risk neutral, or risk averse), incomplete information, credibility, sensitivity analysis, and simulation could also be studied.

References

Betts, Richard K. *Nuclear Blackmail and Nuclear Deterrence*. Washington, DC: Brookings Institution Press, 1987.

Camerer, Colin F. "Behavioral Game Theory: Experiments in Strategic Interaction." *Cuadernos De Economía* 23, no. 41 (2004): 229–236.

Haphuriwat, Naraphorn, Vicki M. Bier, and Henry H. Willis. "Deterring the Smuggling of Nuclear Weapons in Container Freight through Detection and Retaliation." *Decision Analysis* 8, no. 2 (2011): 88–102.

He, F., and J. Zhuang. "Modelling 'Contracts' between a Terrorist Group and a Government in a Sequential Game." *Journal of the Operational Research Society* 63, no. 6 (2012): 790–809.

Hymans, Jacques E. C. *The Psychology of Nuclear Proliferation: Identity, Emotions and Foreign Policy.* Cambridge: Cambridge University Press, 2006.

Levi, Nicolas. "Applying Game Theory to North Korea-China Relations." *Journal of Modern Science* 33, no. 2 (2017): 355–366.

Myerson, Roger B. *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press, 1997.

Popp, George. *How the US Can Work with Its Partners to Contest DPRK Operations: A Virtual Think Tank (ViTTa) Report*. Boston: NSI, 2018.

Powell, Robert. *Nuclear Deterrence Theory: The Search for Credibility.* Cambridge: Cambridge University Press, 1990.

Sagan, Scott D. "The Causes of Nuclear Proliferation." *Current History* 96, no. 609 (1997): 151–156.

———. "Letter to the Editor on "Proliferation Pessimism and Emerging Nuclear Powers.' " *International Security* 22, no. 2 (1997): 193–201.

Sagan, Scott D., and Kenneth N. Waltz. *The Spread of Nuclear Weapons: A Debate*. New York: W. W. Norton, 1995.

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

Shan, Xiaojun, and Jun Zhuang. "Subsidizing to Disrupt a Terrorism Supply Chain—A Four-Player Game." *Journal of the Operational Research Society* 65, no. 7 (2014): 1108–1119.

Singh, Sonali, and Christopher R. Way. "The Correlates of Nuclear Proliferation: A Quantitative Test." *Journal of Conflict Resolution* 48, no. 6 (2004): 859–885.

Solingen, Etel. *Nuclear Logics: Alternative Paths in East Asia and the Middle East*. Princeton, NJ: Princeton University Press, 2007.

Shan, Xiaojun (Gene), and Jun Zhuang. "Modeling Credible Retaliation Threats in Deterring the Smuggling of Nuclear Weapons Using Partial Inspection—A Three-Stage Game." *Decision Analysis* 11, no. 1 (2014): 43–62.

Zhuang, Jun, and Vicki M. Bier. "Balancing Terrorism and Natural Disasters—Defensive Strategy with Endogenous Attacker Effort." *Operations Research* 55, no. 5 (2007): 976–991.

Common Conjectures and International Norms and Law

James Morrow

Equilibrium in game theory requires two conditions: mutual best replies and a common conjecture. The common conjecture ensures that players' strategic expectations match the equilibrium strategies and allows the players to know that their strategies are optimal. Common conjectures provide a way to think about norms—shared understandings about appropriate behavior in a situation. International law codifies international norms, and so can be modeled as the common conjecture of the equilibrium being played. Not all patterns of behavior can be induced under international law because strategies still need to be best replies for the players. I conclude by applying these ideas to the security situation in Northeast Asia.

Introduction

International norms, often codified in international law, shape international politics by setting standards of appropriate conduct and proper responses to inappropriate acts.[1] These normative standards are often described as lying beyond rationality, but they can be modeled by a key but underemphasized concept in game theory, the common conjecture.[2]

Equilibrium in game theory requires two conditions. The first is *mutual best replies*—that each player's equilibrium strategy is a best reply against the other players' equilibrium strategies. This condition ensures that no player can unilaterally make herself better off by changing strategy. Finding sets of strategies that are mutual best replies is the focus of most game theory modeling and where the mathematical calculations are.

Many games, however, have multiple equilibria. For example, consider Chicken as shown in the following figure. It has two equilibria in pure strategies, (Escalate, De-escalate) and (De-escalate, Escalate), and another mixed strategy equilibrium. Each of these sets of strategies are mutual best replies, where the red stars mark the best replies for each player.

	De-escalate	Escalate
De-escalate	(6,3)	(–1,8) ★
Escalate	★ (10,0) ★	(–3,–6)

Chicken

[1] For example, Wendt, *Social Theory of International Politics*.

[2] Morrow, *Order within Anarchy*.

How do the players know which equilibrium they are playing? The second condition of equilibrium is that the players hold a *common conjecture*—a shared understanding that they will place their equilibrium strategy. Then each player knows that her strategy is optimal for her because she anticipates that others will play their equilibrium strategies and her equilibrium strategy is a best reply to them. Anticipations of what others will do are critical in strategic interactions, and the common conjecture allows players to know that their anticipations are correct. This does not require that they can predict the actions of others, as mixed strategies reflect uncertainty about exactly what a player will do.

Equilibria then are stable sets of behavior. No player wishes to change what she is doing, and no player is being fooled systematically. Equilibria are predictions in the sense that if the players are playing one, their behavior is predictable up to any uncertainty in the equilibrium strategies.

Common conjectures, in general, have to be common knowledge among the players; that is, something that all know, that all know that all know, and so on through all levels of knowledge.[3] For the special case of a two-player game, the common conjecture need be only mutual knowledge—something both know, even if they do not know that the other player knows they know it. With more than two players, the higher level of knowledge in common knowledge cuts the Gordian knot of recursions of logic such as "I know that you know that I will play my equilibrium, but how do I know that you know that?" With the common conjecture being common knowledge, all players understand how others will play, and so understand that their equilibrium strategies are best in that circumstance.

Common conjectures are shared understandings that operate as social facts—things that groups of people know that structure their social interactions. Because they are commonly known, individuals cannot change them on their own, unlike their personal beliefs. Further, each needs to consider them when they act because others will act according to them, even when those social facts limit their ability to achieve what they would like. Understanding these social facts active in a social setting helps us understand individual actions in that setting.

How can we understand how common conjectures shape behavior? We need to examine games with multiple equilibria because the common conjecture determines what the players will do. By comparing behavior under different equilibria, knowledge of which they are playing will shape how they act, allowing us to see what the common conjecture does. Iterated Prisoners' Dilemma is an excellent case because it has a wide range of equilibria where the common conjectures produce different behaviors and understandings of acts

[3] Aumann and Brandenburger, "Epistemic Conditions for Nash Equilibrium."

within a given equilibrium. The stage game is given in the following figure, and the players play the game repeatedly and indefinitely with the payoffs from future rounds discounted. When Prisoners' Dilemma is played only once, the only equilibrium is (Defect, defect), producing the dilemma that the players' self-interest leads them to a Pareto-suboptimal outcome—both receive less than if they had played (Cooperate, cooperate). When the game is played repeatedly, they may be able to enforce agreements that improve upon the Both Defect outcome with reciprocal threats where they respond to violations of their agreement by playing Defect for a period afterward. The well-known folk theorem demonstrates that such cooperative agreements can be enforced with reciprocal threats for some discount factor less than one.[4] Parties do not seize short-term gains because the long-term cost from reciprocal retaliation is too great.

	cooperate	defect
Cooperate	(1,1)	(–β,α)
Defect	(α,–β)	(0,0)

$\alpha > 1$ and $\beta > 0$

Prisoners' Dilemma

There is always the unhappy equilibrium where both players play Defect in every round. One obvious improvement on this unhappy state is an agreement to play Cooperate in every round backed up by the threat to play Defect for some period of time if either breaks the agreement by playing Defect when she is supposed to play Cooperate. Depending on the exact values of α and β, this punishment period could last longer than one round; the common conjecture needs to specify that so the players know when punishment ends and they should return to playing Cooperate. Further, the common conjecture also allows them to understand when plays of Defect are violations of their agreement and when they are part of reciprocal punishment, and so should not trigger any future reaction. But there are also asymmetric equilibria where the players are differentiated only by the common conjecture. For some values of α and β, an agreement where the players alternate between (Cooperate, cooperate) and (Defect, cooperate) can be supported in equilibrium. In this asymmetric equilibrium, the Row player gets to exploit the Column player every other round; the Column player allows this because the reward from cooperating in the other rounds is better for him than receiving the (Defect, defect) outcome in every round. The common conjecture that they are playing this asymmetric equilibrium differentiates

[4] Fudenberg and Maskin, "The Folk Theorem."

the players with distinct identities as exploiter and exploited along with the system of retaliation that enforces them.

Common conjectures allow players to form expectations about how others will act in many ways. They describe how players will cooperate and what acts are cooperative and which are violations of the standards. This ability enables them to understand how others will interpret the same actions as acceptable under some conditions and unacceptable under others. The common conjecture also spells out appropriate responses to unacceptable acts, even when those responses might be considered as violations if not taken as a response. Further, it makes all these standards common knowledge, which prevents misunderstandings. In games with communication, such as signaling games or cheap talk games, the common conjecture ensures that all know how signals should be interpreted, allowing those sending signals to anticipate how they will be understood and those receiving them to interpret them correctly. These interpretations are limited by the strategic dynamics of those signals. Finally, the common conjecture could differentiate players with the same structure of interests as in the asymmetric equilibrium above.

Common conjectures could arise from a number of sources.[5] The players could negotiate about which equilibrium they play before the game begins. They presumably will choose a Pareto-optimal equilibrium, but there can be multiple Pareto-optimal equilibria, as is the case in Chicken. A particular equilibrium might be distinguished from others by characteristics that are independent from payoffs, making it a focal point.[6] Equilibria that produce an equal outcome among the players could be focal. Common experience playing the game could lead the players to learn that they should play one equilibrium from that shared history of play. More broadly, a common culture could create a common conjecture by setting how the players should interact. Cultural explanations could account for the common conjecture underlying asymmetric equilibria.

In international politics, international law can be thought of as the codification of the common conjecture. It publicly sets standards of appropriate conduct and in some cases limits what responses can be made to inappropriate conduct. When complete standards of conduct cannot be set because all possible cases cannot be anticipated, international law sets principles that the parties can use to judge how novel cases should be handled. Of course, they may not agree on the application of those principles to specific cases, as was the case for the invasion of Iraq in 2003. International law is often created by the adoption of treaties through ratification of a single document (sometimes with versions in multiple languages) where a state is not bound by the law if it does not accept it through

[5] Kreps, *Game Theory and Economic Modelling*.

[6] Schelling, *Strategy of Conflict*.

ratification. The process of treaty ratification helps to create common knowledge of what standards are encoded in the law and which states accept those standards. Ratification does not compel a state to comply with the law, but states that refuse to ratify a treaty send the signal that they will not be bound by that law. This public rejection allows other states to adjust their anticipations of the future conduct of such states. International law does not force states to abide by the norms codified in law, but it can trigger different strategic dynamics among the parties by changing their anticipations of how others will act, clarify when parties breach those norms, and allow them to coordinate responses to those in breach. For instance, the arms control treaties between the United States and the Soviet Union during the Cold War did not end their strategic nuclear competition; it redirected away from anti-ballistic missiles and toward multiple independently targetable reentry vehicles. The interaction of legal principles and state strategies is critical to the process. Treaties and deterrence are complements, not opposites. Treaties clarify what actions are subject to deterrence and thereby avoid mistaken responses. In turn, deterrence undergirds treaties by making parties less likely to engage in inappropriate acts.

The laws of war, more properly known as international humanitarian law, exemplify the interaction of law and strategy.[7] The treaties that lay out these laws, most notably the four Geneva Conventions, advance general humanitarian principles, state some specific standards for particular issues, and create a framework for warring parties to ascertain that one another are following those rules. Joint ratification of a treaty before war breaks out increases compliance with the standards of that treaty by strengthening reciprocity between states at war with one another. The worst behavior comes when one side has not ratified the relevant treaty, and has thereby triggered strategic dynamics that produce violations by both militaries. But establishing common principles of humanitarianism is not enough to ensure that all parties comply with the law. Patterns of compliance vary across issues because some issues, such as the treatment of civilians, confront an agency issue, namely that soldiers can commit violations on their own in violation of orders from command authorities, that complicate compliance and confound verification that the other side is complying. Law and strategy are entwined in practice.

What does this argument about norms, laws, and common knowledge say about the security situation in Northeast Asia? International law could help set standards of conduct for all parties that could reduce the suspicion among the parties. Six countries—the United States, North Korea, South Korea, China, Japan, and Russia—can play a key role in Northeast Asia, and so must be included in any system to limit security competition in the region. Such a system could focus on a wide range of issues, including trade as well

[7] Morrow, *Order within Anarchy*.

as military competition. It could set standards for North Korean behavior that seek to limit the most provocative behavior while not demanding solutions to all the outstanding issues between North Korea and the United States. That agreement would seek to push strategic competition in the region toward less dangerous forms, rather than to end it. The other parties are important because they can support and reward better behavior as the nature of the competition changes. Finally, the viability of the system must also be considered; why will the parties follow through and pursue their interests within this legal regime instead of breaking out of it? International law can ameliorate international conflicts, but it cannot end them.

References

Aumann, Robert, and Adam Brandenburger. "Epistemic Conditions for Nash Equilibrium." *Econometrica* 63, no. 5 (1995): 1161–1180.

Fudenberg, Drew, and Eric Maskin. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54, no. 3 (1986): 533–554.

Kreps, David M. *Game Theory and Economic Modelling*. New York: Oxford University Press, 1990.

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

Morrow, James D. *Order within Anarchy: The Laws of War as an International Institution*. New York: Cambridge University Press, 2014.

Wendt, Alexander. *Social Theory of International Politics*. New York: Cambridge University Press, 1999.

Strategic Causes of Proliferation: Northeast Asia in Comparative Perspective

Alexandre Debs

This presentation introduces a strategic framework to understand the causes of proliferation, and the effectiveness of counterproliferation tools, placing Northeast Asia in comparative perspective. Acquiring nuclear weapons takes time and effort. Before a nuclear weapons program comes to fruition, adversaries and allies may offer threats and assurances to prevent proliferation. The stronger a potential proliferator is, the more likely it is to succeed in its attempt to acquire nuclear weapons. Threats are most effective against weak potential proliferators, and assurances are most expedient when offered to strong potential proliferators. In Northeast Asia, threats of preventive war have been ineffective in preventing North Korea from acquiring nuclear weapons, given its ability to inflict severe damage on Seoul. Assurances have been crucial in getting South Korea and Japan to forgo nuclear weapons. Looking ahead, a coercive approach toward North Korea is unlikely to be effective, and questions about assurances offered to South Korea and Japan risk spurring their proliferation.

Summary

Nuclear weapons have had a serious impact on world politics since 1945, ushering in what some have called a revolution in international relations. Washington has been concerned to various degrees about the risks of nuclear proliferation, and this concern seems to have been heightened since the end of the Cold War. Yet the rate of nuclear proliferation has been relatively slow—slower than many administrations and academics have predicted—and it has gone down since the end of the Cold War. Over the last two and a half decades, North Korea is one of the only countries to have acquired nuclear weapons.[1] Under what conditions does nuclear proliferation occur? What counterproliferation tools can effectively slow down nuclear proliferation? What implications should we draw for US foreign policy in Northeast Asia?

To answer these questions, this summary first introduces a strategic framework to analyze the causes of nuclear proliferation, and then it reflects on its implications for US foreign policy in Northeast Asia, building on prior work with my colleague Nuno Monteiro.[2] The analysis begins with the idea that nuclear proliferation is a costly investment with delayed

[1] By some accounts, it is the only country to have done so, as most experts believe that Pakistan crossed the nuclear threshold in the late 1980s.

[2] Debs and Monteiro, *Nuclear Politics*; Debs and Monteiro, "Known Unknowns"; Debs and Monteiro, "Cascading Chaos"; and Monteiro and Debs, "Strategic Logic of Nuclear Proliferation."

returns. If a potential proliferator is tempted to acquire nuclear weapons, its adversaries and allies may worry about the consequences of proliferation. Between the moment a country decides to pursue nuclear weapons and the moment it acquires them, adversaries and allies may offer threats and assurances to prevent proliferation. Thus, to understand nuclear proliferation, we need to understand the strategic interaction among a potential proliferator, its adversaries, and its allies, leading up to the acquisition of nuclear weapons or to the end of a nuclear program.

Consider first a simple interaction between two states, the potential proliferator and its adversary. The potential proliferator decides whether to invest in nuclear weapons. The adversary receives some information about the potential proliferator's decision and chooses whether to issue nonproliferation threats or actually strike preventively. Assume that the conventional capabilities of the potential proliferator have two effects on the parameters of the model. First, they increase the cost of preventive war. The stronger the potential proliferator is, the costlier it would be for the adversary to destroy its program. Second, greater conventional capabilities reduce the effect of proliferation. This assumption follows simply from the assertion that nuclear weapons are the "weapons of the weak"—in other words, that states that would otherwise be defeated in a conventional contest may greatly benefit from acquiring such weapons—for example taking regime change off the table. From these assumptions follow the prediction that stronger potential proliferators are more likely to acquire nuclear weapons. The stronger the potential proliferator is, the greater the cost of preventive war and the smaller the effect of proliferation, so that the threat of preventive war is less credible. Stronger potential proliferators can best shield themselves against preventive pressure.

Now consider a richer model with a third actor, an ally of the potential proliferator, that decides whether to support the potential proliferator in any conflict with the adversary, current or future. This model predicts that the presence of the ally weakens the effect of conventional capabilities on nuclearization. Strong potential proliferators may forgo the option of acquiring nuclear weapons, content with the support of the ally. Weak potential proliferators may now have the opportunity to acquire nuclear weapons.

The model also suggests that the effectiveness of nonproliferation tools depends on the balance of power between the potential proliferator and its adversary. With strong potential proliferators, coercive threats are unlikely to be successful, and assurances are likely to be sufficient to prevent proliferation. Threats are ineffective because such states would be able to acquire nuclear weapons on their own. Assurances are expedient because these states' desire to acquire nuclear weapons is relatively small, as the country could do relatively well with conventional weapons alone. With weak potential proliferators, the

reverse holds. Threats can be effective at preventing proliferation, and assurances may be prohibitively costly. Threats can be effective because such states would not be able to acquire nuclear weapons on their own. Assurances could be too costly because these states' desire to acquire nuclear weapons is relatively large.

Turning to the empirical record, we see that these claims are largely borne out in the data.[3] Among states that do not have a nuclear ally, strong potential proliferators have been more likely to acquire nuclear weapons. If the Soviet Union was relatively impervious to counterproliferation threats, states like Iraq, Syria, and Iran either saw their nuclear programs destroyed or were coerced to remain nonnuclear. The presence of a nuclear ally weakens the relationship between conventional capabilities and nuclearization. For example, South Korea, a state with strong conventional capabilities, was eventually convinced that it didn't need its own nuclear weapons, while Pakistan, a state with weak conventional capabilities, benefited from US support in crossing the nuclear threshold. Qualitative analysis also supports the theoretical claims on the effectiveness of nonproliferation tools. South Korea remained nonnuclear because of US assurances, while Taiwan remained nonnuclear because of US coercive pressure.

In broad terms, these empirical patterns suggest that there have been some great successes in US counterproliferation measures. US security assurances have probably been sufficient to convince many allies that they do not need their own nuclear weapons. US threats of preventive war, and actual preventive strikes, have kept some adversaries from acquiring nuclear weapons. If the United States was really concerned about proliferation since the end of the Cold War, it went to great length to prevent the spread of nuclear weapons.

We can now turn our attention to Northeast Asia to gain some perspective on its nuclear history. North Korea began its nuclear program in the 1960s and performed a first test in 2006, followed by subsequent tests in 2009, 2013, and 2016. North Korea had good reasons to doubt the reliability of the security assurances offered by the Soviet Union and China. After the end of the Cold War, Pyongyang accelerated its nuclear weapons program. Ultimately, US threats of preventive action lacked credibility, because of Seoul's proximity to the border with North Korea. As General Gary Luck, commander of US forces, put it in the fall of 1993, the issue was not whether the United States would win a conflict; the issue was that such operations would be too costly. Even the Republican administration of George W. Bush, despite its aggressive rhetoric, was deterred from taking preventive measures. North Korea crossed the nuclear threshold in the early 2000s.

[3] See, e.g., Figure 3.1 in Debs and Monteiro, "Known Unknowns."

South Korea, for its part, began its nuclear weapons program in the wake of the Nixon Doctrine, when the United States called on its East Asian allies to do more for their defense. After Washington learned of Seoul's efforts to acquire a nuclear capability, it first applied coercive measures to end its nuclear program, but South Korea's interest in nuclear weapons lingered. It was only when the administration of Ronald Reagan reaffirmed US security assurances that South Korea terminated its nuclear weapons program.

Japan, finally, never pursued the option of acquiring nuclear weapons. Under the Yoshida Doctrine, Tokyo was content to rely on US assurances for its security needs. In the late 1960s, the Japanese government commissioned a study on the costs and benefits of nuclear weapons. Ultimately, it concluded that the country could acquire nuclear weapons, but that it was not in its interest to do so. An autonomous nuclear deterrent would raise concerns in the region, eventually undermining Japan's security. Since then, Japan has remained committed to nonproliferation.

Looking ahead, this analysis sheds some light on the effectiveness of counterproliferation tools. Coercive measures applied against North Korea are unlikely to obtain its denuclearization. North Korea perceives that nuclear weapons can increase its security. Threats of preventive action lack credibility, as they are too costly, and they are thus unlikely to be effective. At the same time, questions about the reliability of US assurances toward South Korea and Japan may ultimately spur their acquisition of nuclear weapons. So far, both countries have been content to forgo the nuclear option, instead relying on US security guarantees. But they could conceivably complete a nuclear weapons program if they chose to do so. Washington should consider whether it is in its best interest to face future crises in Northeast Asia where its two allies control their own nuclear weapons, or whether it would prefer to control the escalation ladder.

References

Debs, Alexandre, and Nuno P. Monteiro. "Cascading Chaos in Nuclear Northeast Asia." *The Washington Quarterly* 41, no. 1 (2018): 97–113.

———. "Known Unknowns: Power Shifts, Uncertainty, and War." *International Organization* 68, no. 1 (2014): 1–31.

———. *Nuclear Politics: The Strategic Causes of Proliferation*. New York: Cambridge University Press, 2017.

Monteiro, Nuno P., and Alexandre Debs. "The Strategic Logic of Nuclear Proliferation." *International Security* 39, no. 2 (2014): 7–51.

4

DISCUSSION

Previous Johns Hopkins University Applied Physics Laboratory (JHU/APL) work concluded that game theory was a potentially useful analytic framework for assessing multilateral nuclear stability. Stemming from this work was the premise that for the full potential of game theory to be realized, game theorists need to work closely with the national security and nuclear policy communities to produce policy-relevant insights. The primary aim of this workshop was to bring together prominent figures from these two communities to explore the validity of this premise, study potential disconnects and strife between the communities, and identify beneficial avenues for future research and collaboration.

Multilateral Nuclear Stability and Game Theory

A primary finding from the workshop is that game theory provides a useful and rigorous framework to facilitate logical reasoning concerning a complex issue, such as multilateral nuclear stability. Game theory values simplicity and requires the analyst to break down an elaborate problem into its rudimentary elements and to prune all negligible variables. This forced simplification is in itself useful for reaching the core of an issue. For the resulting simplified problem, game theoretic methods have the potential to provide insights on conflicts by building causal connections between the players' motivations, objectives, and actions and the outcomes of their strategic interactions. Deriving causal connections is a key advantage of game theoretic techniques over statistical analysis alone.

Uniqueness

While game theory provides a framework for understanding which outcomes of a conflict are the logical result of the assumed motivations, objectives, and strategies available to the players, game theory is not a tool that derives unique insights about conflicts that cannot be determined through other means. Because of the logical soundness of game theory, when utilized correctly, many of a game's insights might seem intuitive and obvious. However, game theory verifies that those insights are logically substantiated based on assumptions about the game's setup.

Game Theory and Player Objectives

Prior JHU/APL work on game theory concluded that unsupported or incorrect assumptions about the players or the conflict significantly diminish the policy relevance of a game. In response to this issue, one of the goals of the workshop was to help understand how game theorists can utilize the work of international relations and policy professionals. In doing so, game theorists would be able to produce more policy-relevant games based on an accurate and nuanced understanding of the strategies available to, and outcome preferences of, the players. For many games related to the North Korean nuclear crisis, the players are the various nations with the most influence and/or the most at stake in the crisis. These are the Democratic People's Republic of Korea (DPRK), the Republic of Korea (ROK), China, the United States, Japan, and Russia. The myriad of poorly understood, highly uncertain, and often conflicting objectives of these nations is perhaps one of the most daunting challenges in accurately addressing this crisis.

To stimulate thinking for the workshop, JHU/APL sent workshop invitees a draft list of objectives in the North Korean nuclear crisis, from the perspective of the United States,

the DPRK, the ROK, China, Russia, and Japan. We hoped that such a list would serve as a focal point for discussion about prioritizing game outcomes. However, the JHU/APL list of objectives instead sparked an important discussion on when and how these types of lists would be useful for developing and analyzing games. It became clear during the workshop discussions that such a list of objectives cannot serve as a sufficient starting point for game development, nor can it provide the information necessary to fully mitigate the presence of incorrect assumptions folded into a game theory analysis. Additionally, a direct one-to-one mapping between a nation's objectives and its preferred outcomes does not always exist, or is at least elusive. Finally, because game theory often requires simplification of a conflict scenario, an objectives list is often too complex to lend itself easily to inclusion in games.

Papers describing game theory methods tend to focus on types of conflicts, rather than specific situations, which is evident from the scarcity of published game theory analyses specifically about the North Korean nuclear crisis, as discussed in "Game Theory and the North Korean Nuclear Crisis." Therefore, game theorists do not begin developing games by thinking through the objectives or motivations of any particular nation. Rather, they focus on general trade-offs or strategies in a particular conflict or scenario. When game theory is applied to a specific scenario, however, understanding the likely actions a player would take based on assumptions of that player's objectives is an important step. This provides another opportunity for a list of the player's objectives to make substantial contributions to the validity of the resulting game. Unfortunately, there does not exist a mapping between a list of objectives of a nation, broadly developed for a complex and multifaceted scenario like the North Korean nuclear crisis, and the likely actions a player might take or the player's preferred outcomes given the actions and preferences of an opponent. Values assigned to the preferences or payoffs in a game are contextually dependent on the specific situation, meaning the hierarchies of the objectives can change drastically as a situation develops. An example given during the workshop was the game of chess: while the ultimate objective is to take the king, the objective at the beginning of the game is instead to seize the center of the board to create more opportunities to gain an advantage. Additionally, any one outcome might achieve several objectives while failing others. The objective list, as presented at the workshop, did not provide sufficient information on what a player might be willing to risk or what price a player might be willing to pay to achieve a particular objective.

Last, game theorists were skeptical about the utility of such a list because of the complexity and seemingly ambiguous nature of the objectives. Some game theorists argued that the list went against the mathematical desire for simplicity in game theoretic models. Others argued that the list included both intrinsic and derived objectives, and the relations

between the objectives were ignored, which can often be important in translating objectives to possible outcomes. Additionally, some game theorists argued that the objectives list fell prey to the exact issue they were trying to prevent: much of the list was factually unsupported, disagreed upon by the regional and policy experts present, and in general prone to misinterpretations and miscalculations.

However, despite the skepticism of the utility of such a list to directly influence game development, there was a general consensus on the value in making such a list to facilitate methodological thought. Creating and debating a list of objectives can help game theorists improve their understanding of a situation, think through interrelationships of objectives and motivations, and understand a game's players to develop a more realistic and accurate game. As one workshop participant summarized, "making a list is valuable; reading it afterward is not."

Conflicts, Misunderstandings, and Communication Issues

Throughout the workshop, it became clear that there was a disconnect between the game theory and policy communities, not necessarily in the strategic scenarios discussed but in the expectations for a game's results. This disconnect was much deeper than simple vocabulary differences between the communities or an overstatement of confidence in the numerical values presented in a game, although these two factors certainly increased the strife. One example of a simple vocabulary difference was each group's respective use of the word *solution*. While the policy community tended to interpret *solution* as the course of action that will resolve a conflict, the game theory community used the term to represent the mathematical outcome that will result from the game play, which is usually expressed as an equilibrium.

The deeper disconnect between the two communities was more complex. When policy makers turn to game theory, they are looking for possible solutions to a complex strategic problem with many degrees of freedom and large uncertainties. It became clear that some of the policy community experts believed that game theory claims to be a deductive theory of human behavior, from which solutions to complex strategic problems can be determined. This misconception led the policy experts to be skeptical about and distrustful of game theoretic methods. However, while game theory can model certain aspects of human behavior well, it never claims to be an all-encompassing theory of human behavior. Indeed, because game theory excels at showing logical "best" strategies given some game setup, incorrect assumptions about players, their possible actions, and/or their underlying motivations can invalidate any analysis and conclusions resulting from a game.

The Benefits of Collaboration

Despite the divergent expectations, conflicting goals, and skepticism, interaction between the game theorists and the policy community has the potential to produce a deeper understanding of nuclear stability issues. Collaboration between the two communities holds great potential to augment the decision-making process and to organize the thinking of both sides.

The primary benefit from a close collaboration would be increased communication between the communities. The policy community could communicate with game theorists on the most critical conflicts, interactions, and research questions to study, and they could draw attention to important potential applications of game theory methods. Another benefit of this collaboration would be in challenging the assumptions of game theoretic analyses. For example, a key assumption for a game theoretic model is that the players act in their own best self-interests, which is not necessarily true when considering governments and decision-making bodies of nations. What is best for a nation's leader is not always best for that leader's nation; it can be important to know which self-interest will have priority for any specific country of interest (like North Korea) and how and why decisions get made there. The policy community and regional experts can provide such important information for the United States and other nations to ensure that all necessary assumptions are met.

The policy community would gain the benefits from breaking a problem down into its core issues and looking at each of these in a comparative context using game theoretic methods. Working with game theorists who have studied types of conflict, instead of specific scenarios, also has the potential of bringing to light previously ignored variables.

While closer collaboration between the game theory and policy communities would benefit both communities, several obstacles stand in the way. First, the policy community is still largely skeptical about the utility of game theory. Although many of the participants understood the game theory presentations and saw value in the methodologies they described, few saw a direct benefit in game theory helping them to make difficult decisions. Additionally, the academic game theorists have academic responsibilities and motivations, which can be orthogonal to those of the policy community. The academic game theorists are working to obtain tenure, to publish in prestigious journals, and/or to ensure a steady stream of funding for more research. The result of these academic pressures is that there is little motivation for the academic game theorists to develop close collaborations with the policy community.

Future Work

The workshop was specifically planned to address a large number of conflicts and interactions in relation to the North Korean nuclear crisis. However, this broad approach hindered the workshop's ability to produce in-depth insights on the North Korean nuclear crisis itself. To derive insights on the scenario, a smaller workshop between game theorists and the policy community, focused on a specific piece of the conflict, might have had more potential to produce a greater understanding of the conflict and to provide more insight into how game theory could help. During the workshop discussions, participants raised several key questions that game theoretic analyses have the potential to address. These key questions could potentially be the subject of a future workshop focused on a specific piece of the topic and include:

1. Could game theory model the Trump/Kim war of words and help us understand whether it was beneficial for either leader? Did either leader come out ahead? How could this war of words backfire? What is the role of credibility in this situation?
2. What is the "breaking point" for North Korea? Assuming North Korea will not give in to pressure of sanctions, what would cause it to be desperate enough to resort to selling its nuclear weapons/material?
3. How can the United States beat the "commitment game"? If the United States truly wants to strike a deal with North Korea, the countries will have to figure out a way to trust each other.
4. How does an imbalance in a conventional military force affect nuclear deterrence? Does having a worse conventional military force shorten the escalation ladder, and does this lead to increased instability?
5. What leads states to opt for conventional, peripheral war over nuclear war? What is the trade-off for conventional war when a state has nuclear weapons? Would a conventional, peripheral war be less likely or less intense, given that a state does or does not possess nuclear weapons?
6. Do lower-yield nuclear weapons make conventional war more or less likely? How does this change when only one side (compared to both) has such low-yield nuclear weapons?
7. Having patience can matter greatly in nuclear crises. How does cheating on a deal (or worrying about the other party cheating on a deal) affect such a scenario?

8. How does the North Korean nuclear crisis change once conventional preemptive strikes become a legitimate possibility? When does a conventional preemptive strike become the most rational strategy for the United States?
9. Because Schelling's work seems widely appreciated among policy makers, how can his games/insights apply to the North Korean nuclear crisis? Are his games still valid? How do Schelling's assumptions differ between the North Korean nuclear crisis and Cold War scenarios?

5

CONCLUSIONS

The workshop confirmed that game theory is a unique framework for deductive reasoning that can be applied to conflicts of interest, including those involving multilateral nuclear stability and the North Korean nuclear crisis. The presentations by the game theorists helped to bring greater understanding to aspects of the crisis, including proliferation, denuclearization, negotiations, and extended deterrence. However, because game theory cannot directly provide policy makers the optimum strategy to achieve their goals, it is unclear how the insights produced from game theory analyses can directly support policy decisions.

The utility of game theory could be improved if the game theory and policy communities work more closely together. A close collaboration between these communities has the potential to ensure that game theory is correctly applied to nuclear stability issues and that the results of the analyses are understood, including their limitations. Additionally, this collaboration would help to identify policy-relevant issues that are potentially addressable with game theoretic methods, challenge assumptions made in game theory analyses, and further develop the field of game theory.

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY