ALL THE WORLD'S A STAGE

A Constructivist Approach to the Theory of Games and the Security Dilemma

LE MONDE ENTIER EST UNE SCÈNE

Une Approche Constructiviste de la Théorie des Jeux et du Dilemme de la Sécurité

A Thesis Submitted to the Division of Graduate Studies
of the Royal Military College of Canada
by

Elizabeth Leigh Anne Seeley, B.A. (Hon.)
Second Lieutenant

In Partial Fulfillment of the Requirements for the Degree of
Master of Arts in War Studies

June 2025

DEDICATION

To Platoons 2205 and 2206 (BMOQ-A, CFB Gagetown, New Brunswick, 2022)

Non Quam Cede

ACKNOWLEGDEMENTS

Thank you, Dr. Jane Boulden, for your indispensable guidance, encouragement, and unyielding patience throughout this project. Moreover, I am grateful for my friends; my friends and colleagues at Global Affairs Canada; and, of course, my parents, for being there when I needed their support the most.

ABSTRACT

This thesis proposes a game-theoretic approach inspired by Constructivism to examine how states can overcome the security dilemma. Using rationalist models, this thesis seeks to propose a *via media* between Rationalism and Constructivism. Substantiating the constructivist view that states can come to value cooperation through their interaction, this thesis explores how identity and norms affect strategic choice.

Using Constructivism and Game Theory, the thesis explores the *identity-preference relationship* in the context of the security dilemma. To this end, it uses three rationalist devices: hyper-games, games of conditional reciprocity, and costly signalling games. In game-theoretic terms, the security dilemma is a collective action problem in which individual rationality results in collective sub-optimality. Mitigating the security dilemma entails transforming it into a coordination game. With this in mind, the thesis proposes a two-step solution to the security dilemma: through *tit-for-tat*, states internalize norms of reciprocity and transform the collective action problem into a coordination game, and through *costly signalling,* states prevent coordination failures.

The thesis formalizes the identity-preference relationship, which it explores, using repeated games of conditional reciprocity and costly signaling games, in the context of preference change within the security dilemma. It explores how states can transform the self-help system into a rule-based international society based on mutual reciprocity where cooperation occurs organically. This thesis demonstrates how constructivists can use Game Theory to explain this concept systematically without having to forgo their inter-subjective approach.

RÉSUMÉ

Cette thèse propose une approche théorique des jeux inspirée du constructivisme pour examiner comment les États peuvent surmonter le dilemme de la sécurité. En utilisant des modèles rationalistes, cette thèse cherche à proposer une via media entre le rationalisme et le constructivisme. En étayant le point de vue constructiviste selon lequel les États peuvent en venir à apprécier la coopération grâce à leur interaction, cette thèse explore la manière dont l'identité et les normes affectent le choix stratégique.

En s'appuyant sur le constructivisme et la théorie des jeux, la thèse explore la relation identité-préférence dans le contexte du dilemme de sécurité. À cette fin, elle utilise trois dispositifs rationalistes : les hyper-jeux, les jeux de réciprocité conditionnelle et les jeux de signalisation coûteux. En termes de théorie des jeux, le dilemme de la sécurité est un problème d'action collective où la rationalité individuelle se traduit par une sous-optimalité collective. Pour atténuer le dilemme de sécurité, il faut le transformer en un jeu de coordination. Dans cette optique, la thèse propose une solution en deux étapes au dilemme de la sécurité : par le biais du tit-for-tat, les États intériorisent les normes de réciprocité et transforment les problèmes d'action collective en jeu de coordination, et par le biais d'une signalisation coûteuse, les États peuvent développer la confiance.

La thèse formalise la relation identité-préférence, qu'elle explore à l'aide de jeux répétés de réciprocité conditionnelle et de jeux de signalisation coûteux, dans le contexte du changement de préférence au sein du dilemme de la sécurité. L'explore la manière dont les États peuvent transformer le système d'entraide en une société internationale fondée sur des règles et basée sur la réciprocité mutuelle, où la coopération se produit de manière organique. Cette thèse démontre comment les constructivistes peuvent utiliser la théorie des jeux pour expliquer ce concept de manière systématique sans avoir à renoncer à leur approche intersubjective.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I
Our Doubts are Traitors: The Security Dilemma

Introduction

This thesis addresses an ontological debate within International Relations Theory (IR) between Rationalism and Constructivism over the issue of preference change by states—the ways in which state goals and values evolve over time. Rationalist theories of IR, such as Neorealism, attribute preference change to changes in the general distribution of power within the international system. However, this thesis argues that preferences can change *absent* changes in material capabilities. Although this thesis challenges Rationalism's materialist ontology, it does by using a rationalist device, namely, Game Theory. By using Game Theory to substantiate Constructivism, this thesis proposes a *hybrid* approach centered on preference change.

To this end, the thesis treats Rationalism and Constructivism less as distinct ontologies and more as methodological tools.[1] Hence, it focuses not so much on defining the *nature* of preference change as explaining *why* and *how* preference change occurs. Using Constructivism, the thesis *defines* preference change as a function of norms, identities, and preferences, and using Game Theory, a rationalist device, it *models* this relationship vis-à-vis the security dilemma. It explores how resolving the security dilemma involves states transforming their preferences through the adoption of cooperative norms and identities.

This thesis explores how states can transform their preferences to overcome *self-help*, the pattern of interaction underlying the security dilemma. Like Constructivism, but unlike rationalist approaches, this thesis assumes that self-help is just *one* of many possible patterns of interaction that can come to define the international system.[2] A pattern of interaction is a specific constellation of preferences centered on a unique normative culture. Rooted in a culture of *egoism,* the self-help system consists of *competitors* driven by relative power. The thesis proposes a solution to the security dilemma that involves changing the *culture* of the international system. By adopting a *shared* identity, states come to prefer cooperation over competition.

Although the thesis uses rationalist models to analyse preference change, in doing so, the thesis adopts a constructivist approach centered on the *identity-preference relationship*, which refers to the way in which *identity,* how states perceive themselves in relation to others, affect preferences, or what states *want* or *value.* Although constructivists seldom use formal models to analyze this relationship, this thesis explores how constructivists can use Game Theory to explain this concept systematically without having to forgo their inter-subjective approach. The thesis uses hypergames to formalize the identity-preference relationship, which it explores, using repeated games of conditional reciprocity and costly signaling games, in the context of preference change by states within the security dilemma.

The thesis uses *hypergames* to model the relationship between identity, preferences, perception, and behaviour; *repeated games of conditional reciprocity* to show how states internalize cooperative norms across time; and *costly signalling games* to show how states can update their beliefs about each other in ways that facilitate trust. Using these three rationalist devices, the thesis looks to substantiate the constructivist identity-preference relationship. In doing so, the thesis proposes a *via media* between two frameworks that IR scholars have long treated as mutually exclusive, thereby opening up a new avenue for IR theorizing in the process.

[1] Friedrick Kratochwil, "Constructivism as an Approach to Interdisciplinary Study," in *Constructing International Relations,* ed. Karin Fierke (Armonk, New York: M.E. Sharpe, 2001).
[2] Stanely Hoffman, *Contemporary Theory in International Relations* (Englewood Cliffs, NJ: Prentice-Hall, 1960), 90.

Constructivism and Rationalism

To establish the groundwork for a hybrid theory of preference change, the thesis first examines how Constructivism and Rationalism differ ontologically. Much of their disagreement over how state preferences emerge, evolve, and shape behaviour stems from how they define the *identity-preference relationship*—or how the way states perceive themselves in relation to others affect what those states want. This thesis explores these differences by comparing Structural Constructivism with two rationalist theories, Neorealism and Neoliberal Institutionalism. After exploring how these three schools of thought interpret the security dilemma, the thesis assesses the explanatory, descriptive, and predictive utility of their respective methodologies. This chapter closes with an outline of a hybrid solution to the security dilemma.

Differences in Ontology

Rationalism views states as rational egoists driven by *instrumental rationality defined by utility maximization.* They are rational insofar as their actions are goal-driven—with that goal being to maximize self-interest. Under the rationalist view, states share a common, *given* identity: as an egoistic, utility-maximizing agent.[3] Moreover, Rationalism adopts a materialist ontology that minimizes the causal significance non-material factors such as norms and identities have on behaviour.

Rationalism's deductive, utility-based approach especially complements Neorealism and Neoliberalism, which conceptualize states as goal-driven egoists driven by utility maximization. Rationalist IR theories argue that how states act is a function of their *relative position* within the international system. Their *material* capabilities relative to those of other states within the global system affect how they behave. Preferences, or what states *want*, cannot change absent shifts in the distribution of power within the system. Rationalist theories explore how states with *given* preferences behave subject to structural constraints imposed by the international system.

Neorealism and Neoliberalism are *systemic* theories whose analytical focus is the agent-system relationship: the relationship between states and the system-level factors most relevant to their behaviour. *Systems* are comprised of *structures* and *agents*. Agents, in turn, consist of *attributes*. In the language of IR, *system* refers to the international system; *agents,* the states comprising the system; *attributes*, the national, and sub-national properties of states; and *structure*, the constraints and limitations the system imposes on state behaviour.[4] Rationalist theories isolate the structure-agent relationship.[5] By holding preferences constant, they assume that agents respond to structure in similar and, thus, predictable ways, thereby facilitating a generalizable theory of behaviour.

Systemic theories explore how different elements within the international system interact. While reductionist, non-systemic theories reduce interaction to attributes, systemic theories focus less on the attributes of states and more on their arrangement (structures).[6] Neorealists Kenneth Waltz,[7] John

[3] Thomas Risse, "Let's Argue: Communicative Action in World Politics," *International Organization* 54, no. 1 (2000): 3.

[4] Philip Cerny, *The Changing Architecture of Politics, Structure, Agency, and the Future of the State* (London: Sage Publications, 1990), 4.

[5] Kenneth Waltz, *Theory of International Politics* (Boston, Mass.: McGraw-Hill, 1979), 42

[6] Waltz, 18.

[7] Waltz, 20.

Mearsheimer,[8] and Joseph Grieco,[9] and neoliberalists, such as Robert Axelrod,[10] Charles Lipson,[11] and Robert Keohane,[12] examine how factors on the systemic, structural, level affect unit-level properties.[13] For neorealists, *power*, defined as the relative distribution of material capabilities between states, that is, their *arrangement*, is *structure*.[14] Power has *unit-level* effects; it determines how states interact within the international system. Their arrangement within the system affects how they interact by "disposing force on [them]".[15] Alternatively, interaction refers to a *pattern of relations* among states—as opposed to their *arrangement*.[16] Hence, neorealists treat interaction as a *unit-level* property, giving it less analytical emphasis vis-à-vis structure. Consequently, they attribute differences in behaviour not to variations in state attributes but *structure*.[17]

Neorealists argue that structure emerges from interaction spontaneously. However, once structure emerges, "[a system] becomes a force in itself, and a force that the constitutive units… cannot control."[18] Unit-level phenomena such as interaction have system-level effects only if they alter material structure. Hence, patterns of behaviour that preserve the prevailing distribution of relative power within the system reinforce structure while those that change it transform structure.[19]

Falling under the rationalist tradition, Neorealism and Neoliberalism reduce behaviour to utility-maximization based on fixed preferences. The two theories differ, however, in how they define utility. Although both conflate rationality with utility maximization, neorealists define rationality in terms of *relative* gain maximization while neoliberalists define it in terms of *absolute* gain maximization.[20] Under the neorealist view, if in selecting some option X over an alternative option Y, state A improves, or at least does not weaken, its relative position within the system, then A, is said to be acting *rationally*. However, for neoliberalists, should Y yield a higher absolute payoff than X for A, then A is said to be acting *irrationally*.

Unlike Neorealism and Neoliberalism, Constructivism does not define behaviour in terms of instrumental rationality. Instead, it argues that states are driven by *normative, rule-based rationality*. For constructivists, preferences are a function of *identity* or *self-perception*. Identity affects what states want, thereby mediating behaviour. Constructivists argue that shared inter-subjective beliefs about

[8] John Mearsheimer, *The Tragedy of Great Power Politics* (New York, NY: WW Norton, 2014).

[9] Joseph Grieco, "Anarchy and the Limits of Cooperation: A Realist Critique of the Newest Liberal Institutionalism," in *Neorealism and Neoliberalism,* ed. David Baldwin (New York, NY: Columbia University Press, 1993).

[10] Robert Axelrod, *The Evolution of Cooperation* (New York, NY: Basic Books, 1984)

[11] Charles Lipson, "International Cooperation in Economic and Security Affairs," *World Politics* 37, no. 1 (1984): 1-23.

[12] Robert Keohane, *After Hegemony: Cooperation and Discord in the World Political Economy.* (Princeton, NJ: Princeton University Press, 1984).

[13] Joseph Grieco, "Anarchy and the Limits of Cooperation: A Realist Critique of the Newest Liberal Institutionalism," in *Neorealism and Neoliberalism,* ed. David Baldwin (New York, NY: Columbia University Press, 1993), 486.

[14] Kenneth Waltz, *Theory of International Politics* (Boston, Mass.: McGraw-Hill, 1979), 18.

[15] Waltz, 72

[16] Waltz, 95.

[17] David Dessler, "What's at Stake in the Agent-Structure Debate?" *International Organization* 43, no. 3 (1989): 449.

[18] Dessler, 32.

[19] Robert Keohane, "Theory of World Politics: Structural Realism and Beyond," in *Neorealism and its Critics,* ed. Robert Keohane (New York: Columbia University Press, 1986), 166.

[20] Joseph Grieco, "Anarchy and the Limits of Cooperation: A Realist Critique of the Newest Liberal Institutionalism," in *Neorealism and Neoliberalism,* ed. David Baldwin (New York, NY: Columbia University Press, 1993).

appropriateness and legitimacy affect what states want and how they identify themselves in relation to others.[21] Unlike rationalists, constructivists *endogenize* preferences, arguing that preferences *vary* across states and different socio-cultural milieux. Constructivists argue that preferences can change absent changes in the distribution of the material capabilities between states. *Through their interaction,* states can change the system without altering material structure.[22]

For constructivists, what renders material structure, the relative distribution of power within the international system, significant is the *social context* within which states interact.[23] Constructivists argue that material factors are meaningful only by what they signify within a broader socio-cultural context.[24] The social environment within which states interact determines what kinds of entities they *are* and how, through interaction, states *make* the international system.[25] Inter-subjective meanings—comprised of values, norms, and beliefs—which states use to understand, create, reproduce, and transform the world, emerge from interaction. By examining how social structure—the distribution of ideas, knowledge, and shared values within the international system—shape behaviour, constructivists challenge rationalism's materialist ontology. Constructivism argues that international politics is *socially constructed* through interaction.[26] It explores how social structure—norms, values, and inter-subjective beliefs—affect how states acquire, develop, and transform social identities and, by extension, their preferences.

Conceptualizing the Security Dilemma

Neorealism, Neoliberalism, and Constructivism conceptualize the security dilemma differently. This divergence stems from how they define *international anarchy* and their assumptions about the nature of norms, identities, and preferences. These differences, in turn, affect how the two frameworks explain preference change.

Central to IR theory is the notion that the international political system is anarchic. Marked by the absence of a supranational authority capable of enforcing rules of conduct and adjudicating disputes between states, the international system is governed by a principle of *self-help.*[27] Within a self-help system, states look to decouple their security from others, relying instead on their own capabilities to survive within the system. The uncertainty states have of what other states within the international system will do in the future makes states come to believe that they cannot rely on anyone except themselves for survival. This *future uncertainty* drives states to enter a power competition where each state seeks to increase their relative capabilities.

In a bilateral arms race, arms acquisition by one state inadvertently threatens the security of the other state, prompting the latter to arm in kind so to re-establish the balance-of-power. However, in doing so, it prompts the former to arm and *vice versa*. This reciprocity characterizes the so-called *security dilemma*—a self-reinforcing power competition where states inadvertently reduce their security by

[21] Martha Finnemore and Kathryn Sikkink, "International Norm Dynamics and Political Change," *International Organization* 52, no. 4 (1998): 391.

[22] Alexander Wendt, *Social Theory of International Politics* (Cambridge: Cambridge University Press, 1999).

[23] Nicholas Onuf, "Constructivism: A User's Manual," in *International Relations in a Constructed World,* eds. by Vendulka Kubalkova, Nicholas Onuf, and Paul Kowert (Armonk, New York: M.E. Sharpe, 1998).

[24] John Ruggie, "What Makes the World Hang Together? Neo-Utilitarianism and the Social Constructivist Challenge," *International Organization* 52 (1998).

[25] Alexander Wendt, *Social Theory of International Politics* (Cambridge: Cambridge University Press, 1999).

[26] Bear Braumoeller, "Nested Politics: A New Systematic Theory of IR," (Waterhead Center for International Affairs, 2004), 13.

[27] Kenneth Waltz, *Man, the State, and War: A Theoretical Analysis* (New York, NY: Columbia University Press, 2001), 237.

threatening the security of others. Security dilemmas illustrate how increases in power do not always translate to increases in security. John Herz writes:

> Groups and individuals who live [in anarchy] … must be… concerned about their security from being attacked, subjected, dominated, or annihilated by [others]. Striving to attain security from such attacks, they are driven to acquire more and more power in order to escape the effects of the power of others. This, in turn, renders the others more insecure and compels them to prepare for the worst.[28]

Rationalist theories disagree with Constructivism over the causes of the security dilemma and whether states can overcome it. For defensive realists, such as Kenneth Waltz, Charles Glaser,[29] Robert Jervis,[30] Paul Roe,[31] and Jeffrey Taliaferro,[32] security dilemmas can occur only between states with security-seeking motives. So-called *benign,* security-seeking states differ from their so-called *greedy*, power-maximizing counter-parts.[33] While benign states seek power out of insecurity, greedy states are motivated by non-security-related goals.[34] For defensive realists, states are fundamentally benign; they arm not so much out of a *desire to exploit others* as to lessen their vulnerability to *exploitation by others.* The uncertainty benign states have of the present and, especially *future*, intentions of other states within the system heighten this perception of insecurity.[35]

Absent mechanisms by which security-seekers can differentiate between greedy and benign states, they seek security out of fear—to protect themselves from exploitation. However, in doing so, they cause other security-seeking states to reciprocate, catalyzing a self-defeating cycle of mutual escalation and security reduction. By acquiring power, states alter the distribution of power to its relative advantage. However, since improvements in relative power for one state presuppose decreases in the relative power of other states, states end up always trying to balance their power vis-à-vis others within the international system. A security dilemma thereby ensues.

Although both flavours of Neorealism argue that states care about increasing their relative power, they disagree on *how much* relative power states pursue and for *what purpose*. Offensive realists, such as John Mearsheimer, argue that because states can never know how much power they would need to ensure their *long-term* security, states always seek to improve their position within the system. [36] Alternatively, defensive realists characterize states as security-seekers concerned only about acquiring *just enough* relative power to ensure their survival. For defensive realists, states are not so much driven by *power maximization* as by *loss aversion;* they argue that states do not seek out relative *gains* so much as try to avoid relative *losses*.[37] Common to both approaches, however, is the notion that for states, security is *sine qua non* to all other pursuits. As Raymond Aron writes:

[28] John Herz, *Political Realism and Political Idealism* (Chicago: University of Chicago Press, 1951), 157.

[29] Charles Glaser, "Realists as Optimists," *International Security* 19, no. 3 (1994): 67.

[30] Robert Jervis, "Realism, Neoliberalism, and Cooperation: Understanding the Debate," *International Security* 24, no. 1 (1999): 42-63.

[31] Paul Roe, "Actors' Responsibility in Tight, Regular, or Loose Security Dilemmas," *Security Dialogue* 32, no. 1 (2001).

[32] Jeffrey Taliaferro, "Security Seeking Under Anarchy," *International Security* 25, no. 3 (2000): 129.

[33] Shiping Tang, "The Security Dilemma: A Conceptual Analysis," *Security Studies* 18, no. 3 (2009): 613.

[34] Charles Glaser, "Political Consequences of Military Strategy," *World Politics* 44, no. 4 (1992): 499-508.

[35] Robert Jervis, *Perception and Misperception* (Princeton, NJ: Princeton University Press, 1976), 62.

[36] John Mearsheimer, *The Tragedy of Great Power Politics* (New York, NY: WW Norton, 2014), 35.

[37] Kenneth Waltz, *Theory of International Politics* (Boston, Mass.: McGraw-Hill, 1979), 126.

> Politics, insofar as it concerns relations among states, seems to signify—in both ideal and objective terms—simply the survival of states confronting the potential threat created by the [mere] existence of other states.[38]

For both realisms, structure imposes limits on cooperation. However, offensive realists disagree with defensive realists over the extent to which states can overcome these limitations and the security dilemma. Although states are driven by self-help, for defensive realists that does not mean that states do *not want* to or *cannot* cooperate. It does imply, however, that cooperation occurs only within parameters set by the structure of anarchy.[39] Defensive realists argue that states will cooperate so long as they trust each other.[40] *Contra* defensive realists, however, offensive realists argue that uncertainty precludes trust from developing and that, as a result, states cannot overcome security dilemmas.

Recall that, for neorealists, anarchy's self-help nature induces *balancing* rather than greedy power competition. Under this view, the pursuit of relative power within the system is driven not by greed but by insecurity. However, neoliberalists argue that states are indifferent to how *others* benefit or lose from interaction; they only care about how interaction impacts their *own* respective utilities. Under the neoliberalist view then, states are driven not so much by relative gains as by absolute gains; they are *greedy* by default.

Unlike neoliberalists, neorealists argue that states prioritize relative power over absolute power since shifts in absolute power do not always alter the overall balance of power within the system. Moreover, states pursue power not to maximize it *for its own sake* but to maintain their respective positions within the system or to improve them lest others strengthen theirs.[41] For neoliberalists, states only care about maximizing their absolute power. Attributing behaviour to a desire to maximize absolute utility, neoliberalism implies that the principal aim of states is not to maximize *security* but to maximize their *personal well-being*. Although they agree with neorealists that anarchy constrains state behaviour by making states less willing to cooperate, neoliberalists argue that neorealists effectively overestimate these constraining effects while underestimating the extent to which institutions can mitigate them.[42] For neoliberalists, institutions can help states identify opportunities for mutual gain and coordinate, thereby facilitating cooperation.

Conversely, for constructivists, anarchy is what states 'make' of it. Although it does not reject the idea that the security dilemma can only occur within a competitive, self-help system, Constructivism challenges the neorealist notion that self-help is *given* by the structure of anarchy. For neorealists, the structural constraints anarchy imposes on states within the international system make reciprocal balancing and power competition inevitable.[43] Alternatively, for constructivists, international anarchy is only *one* of many possible *cultural instantiations* that can define the system.[44] Underlying each 'instantiation' is a unique shared culture created through social practices. Wendt attributes self-help to a specific pattern of interaction or *shared culture* based on egoism, writing:

[38] Raymond Aron, *Peace and War.* Garden City, NY: Doubleday, 1966.

[39] Aron, 116.

[40] Charles Glaser, "The Security Dilemma Revisited," *World Politics* 50, no. 1 (1997): 183.

[41] E.H. Carr*, Twenty Years Crisis* (London, UK: Palgrave Macmillan, 2001), 111.

[42] Joseph Grieco, "Anarchy and the Limits of Cooperation: A Realist Critique of the Newest Liberal Institutionalism," in *Neorealism and Neoliberalism,* ed. David Baldwin (New York, NY: Columbia University Press, 1993), 486.

[43] Kenneth Waltz, *Man, the State, and War: A Theoretical Analysis* (New York, NY: Columbia University Press, 2001), 237.

[44] Alexander Wendt, "Anarchy is What States Make of It: The Social Construction of Power Politics," *International Organization* 46, no. 2 (1992).

> Self-help and power politics do not follow either logically or causally from anarchy… a self-help world [is due] to process [interaction], not structure. Self-help [is an] institution, not [an] essential featur[e] of anarchy.[45]

Rationalist theories argue that the *principal* determinant of state behaviour is the distribution of material power within the system; it emphasizes the constraining effect material, as opposed to social, structure has on behaviour. How states are *arranged* within the system affects how they behave and, by extension, their interaction. Rationalist theories argue that material structure demarcates cooperation under anarchy. Alternatively, constructivists reduce the limits of cooperation to *social* structure, such as *identities* and *values.* While rationalist theories argue that the self-help system can change *only* if there is a change in material structure, constructivists argue that system change can occur following a change in preferences—even absent a change in the distribution of power within the international system.

Differences in Methodology

Unlike Constructivism, rationalist theories focus less on preference *change* and more on how states act upon *fixed* preferences. By treating preferences exogenously, holding them constant, they treat *behaviour,* as opposed to the *nature of preferences*, as its principal analytical focus. Rationalist theories ask not so much *what states want* as *given what they want, how will utility-maximizing states behave*. Similarly, rationalist models seek to describe rational decision-making in terms of actors optimizing exogenous preferences. Like rationalist theories, rationalist models are, thus, *ahistorical;* they exogenize preferences, abstracting them from a broader social context. Although they can describe how actors with given interests behave, their ahistoricism limits their ability to explain preference formation and change.

Conversely, adopting a historical approach, Constructivism endeavours to show how state preferences and identities evolve with changes in social structure. However, its emphasis on the past—and retrospective use of historical case studies—limits Constructivism's ability to address how *future uncertainty* induces relative power competition.[46] This historicism contributes to the relative sparsity of constructivist literature on the issue of future uncertainty vis-à-vis the security dilemma. This thesis seeks to address this lacuna. Instead of focusing on how the international system *came to be defined* by certain patterns of interaction, this thesis adopts a more forward-looking approach. To this end, in Chapter III, this thesis uses two rationalist devices centered on *inter-temporal choice*, namely: games of conditional reciprocity to explore how states internalize cooperative norms across time and costly signalling games to show how states update their beliefs about each other over multiple interactions. These rationalist devices explore how beliefs about the *future* affect *current* behaviour.

Bridging the Constructivist-Rationalist Divide

Neorealists reduce what states *want* to their relative position within the international political system. Accordingly, under this view, states pursue goals that are determined by forces out of their control. Absent changes within their respective material capabilities, states cannot change their preferences. Drawing upon Constructivism, this thesis argues that *non-material* factors such as identity and norms mediate preferences and, by extension, preference creation and preference change. The thesis challenges the rationalist assumption that preferences cannot change absent alterations to the distribution of power within the international system. Instead, it argues that preferences are a function not of one's relative position within the system but of *identity.* This thesis explores this *identity-preference relationship* in the context of the security dilemma.

[45] Alexander Wendt, "Anarchy is What States Make of It: The Social Construction of Power Politics," *International Organization* 46, no. 2 (1992): 394-395.

[46] Dale Copeland, "The Constructivist Challenge to Structural Realism: A Review Essay," *Social Theory of International Politics* 25, no. 2 (2000).

Moreover, this thesis assumes that self-help is just one of the many possible patterns of behaviour that can come to define the international system. Patterns of interaction can change absent changes in material structure. This thesis argues that collective identity change can alter patterns of interaction by affecting how states develop preferences. It argues that states can come to value cooperation by adopting a non-egoistic identity defined by a *preference* for joint gains. By adopting an identity centered on *collective,* as opposed to *self,* interest, states come to prefer cooperation over competition. States can, thus, replace self-help with a *new* pattern of interaction based not on egoism but cooperation.

This thesis assumes that a power competition is a security dilemma only if it occurs within a self-help system comprised of security-seeking states driven by fear—as opposed to greed. Since states are driven into security dilemmas not by a desire to maximize power but by *loss aversion*, resolving the security dilemma involves reducing future uncertainty, mitigating egoism, and cultivating trust. This thesis attributes the security dilemma to loss aversion driven by these three factors. States are less likely to risk exploitation when confronted with a high degree of future uncertainty and mistrust. Moreover, within a self-help system, states have an egoistic identity, causing them to view interaction in zero-sum, all-or-nothing, terms, thereby preventing them from seeking out opportunities for potential mutual gain.

This thesis seeks to show how constructivists can use Game Theory to articulate a solution to the security dilemma without having to forgo their inter-subjective approach. With this in mind, the solution that this thesis proposes to the security dilemma involves a *two-step* process modelled by two game-theoretic devices: games of conditional reciprocity and costly signalling games. The first step involves transforming the security dilemma, a *collective action problem*, into a *coordination game* through games of conditional reciprocity. This step involves resolving future uncertainty and overcoming egoism. The former entails reducing the risk of exploitation and making exploitation more costly; the latter, transforming state identities through norms of reciprocity. The second step of the solution involves using costly signalling games to build trust and prevent coordination failures.

A common theme throughout this thesis is the identity-preference relationship. Chapter II formalizes this relationship using hypergames while Chapter III applies it to a solution to the security dilemma. After providing an overview of Conventional Game Theory to familiarize the reader with game-theoretic language, concepts, and methodology, Chapter II uses hypergame theory to model the security dilemma. Unlike Conventional Game Theory, hypergame theory enables an inter-subjective, rationalist analysis of the relationship between identity, preferences, perception, and behaviour. Chapter III then proposes a two-step solution to the security dilemma. The first step involves overcoming future uncertainty and egoism through identity-preference *change*. The second step involves overcoming mistrust through identity-preference *revelation.*

CHAPTER II
The Mind Shows Us What We Want to See: Perception and Behaviour

Introduction

The purpose of this chapter is to formalize the preference-identity relationship, a concept central to a Constructivism. It explores how preferences and identities are linked and how preferences are a function of state types. This chapter provides the groundwork for Chapter III, where a solution to the security dilemma centered on this preference-identity relationship is proposed. Chapter II opens with an overview of Conventional Game Theory before exploring the explanatory utility of nonconventional, game-theoretic models vis-à-vis the security dilemma.

Conventional Game Theory

Rationalism is closely related to Rational Choice Theory (RCT). A key assumption underlying RCT is that individuals, when confronted with competing courses of action, will always select the strategy that best aligns with their respective preferences. RCT can be further divided into Decision Theory, which pertains to *individual* decision-making, and Game Theory, which deals with *strategic* interaction. Two players, A and B, are said to be in a strategic interaction if A's payoff is dependent upon B's choices, and *vice versa.* Game Theory explores how actors, or in this case, states, base their decisions on the beliefs they have about what others *will likely do* under certain constraints.

Game Theory defines actors by their preferences over a set of outcomes and the set of possible options available to them. Conventional Game Theory, a sub-set of Game Theory based on the axiomatic theory of utility, is a theory of preferences over actions; it focuses not so much on how actors *develop* preferences over a set of outcomes as how units act on their *given* preferences over a set of outcomes.[47] Recall from Chapter I that rationalist theories of IR propose a systemic approach to explain how states with given preferences behave when subject to the structural constraints of international anarchy. They take preferences as *given* before attempting to explain how states choose among competing outcomes *in light* of their preferences. For that reason, much of the use of Game Theory in IR has been centered on answering system-level questions.

In *Game Theory and Economic Behaviour*, mathematicians John von Neumann and Oskar Morgenstern (N-M) present the axiomatic theory of utility that underlies Conventional Game Theory.[48] According to the axiomatic theory of utility, agents act in accordance with their respective subjective utility function(s), which determines how they establish preferences over a set of outcomes. Preferences are a function of how agents calculate utility. N-M thereby reduces rational choice to decision-making based on maximizing the expected value of some utility function. A decision-maker is *rational* only if they act in a way that maximizes their utility function. For neorealists, states define their utility function in terms of *relative gain* maximization while for neoliberalists, in terms of *absolute gain* maximization.

N-M formulized the *two-person, zero-sum game in normal-form with complete information.* Normal-form and extended-form games represent simultaneous games by way of payoff matrices and information sets, respectively. Game theorists tend to prefer normal-form games over extensive-form games to represent simultaneous-move or non-sequential game. In zero-sum games, one player's gain is equal to the other player's loss, precluding opportunities for mutual gain. Normal-form games can be represented by a payoff matrix comprised of four cells. Each cell corresponds to a unique strategy combination yielding a specific payoff for each player.

[47] Robert Powell, "Neorealism and its Critics," *International Organization* 48, no. 2 (1994): 318

[48] John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behaviour* (Princeton, NJ: Princeton University Press, 1944)

Neumann-Morgenstern games (NMGs) consist of the following basic elements:

		B	
		b_1	b_2
A	a_1	a_1, b_1	a_1, b_2
	a_2	a_2, b_1	a_2, b_2

$$G = \{V_A, V_B\}$$
$$N = \{A, B\}$$
$$S_A = \{a_1, a_2\}, S_B = \{b_1, b_2\}$$
$$U = S_A \times S_B = \{[a_1, b_1], [a_1, b_2], [a_2, b_1], [a_2, b_2]\}$$

Figure 1: Neumann-Morgenstern Game in Normal Form

1. Players: $N = \{1,2 \dots i \dots n\}$

2. Strategies: a strategy of player i is a set of actions that i can take. The set of strategies for player i is defined by: $s_i \in S_i \mid \forall i \in N$

3. Outcomes: An outcome represents a possible result from the interaction once all of the players have selected some strategy from their respective strategy set. An outcome is denoted by: $u = \{s_1, s_2, \dots s_i \dots s_n\}$, whereas the set of outcomes is given by $U \mid u \in$ where $U = \{S_1 \times S_2 \dots \times S_i \dots \times S_n\}$

4. Preference Vectors: A preference vector is an ordered set comprised of an outcome set U and a specified relation R_i, which is defined over U and ranks all elements in U from most-to-least preferable for each player. The preference vector for player i is defined as $V_i = \{U, R_i\}$.

5. Game: defined by the set of preference vectors for all players in N, such that: $G = \{V_1, V_2, \dots V_n\}$[49]

In NMGs, all players seek to maximize their respective payoffs, selecting the strategy that corresponds to their most preferred outcome. NMGs are games of *complete information* comprised of players with *complete, fixed* preferences. Players are said to have complete information when they know, and know that others know, everyone's strategies, preferences, and options, but are yet unaware of which strategy others will play. Player A is said to have *complete* preferences if—for options $a, b,$ and c—A prefers a over b and b over c and a over c. Moreover, A's preferences are *fixed* if the preference-ordering $a > c > b$ is always true. Also central to NMGs is the notion of *perfect rationality,* which assumes that all players will always select the strategy that maximizes their respective utility function. Under this view, a decision-maker is *rational* only if it they act in a way that maximizes utility.

In zero-sum games, the aggregate gains and losses between the players equal zero. In *pure coordination games*, players have identical preferences—individual rationality naturally results in a mutually optimal outcome.[50] However, in non-zero-sum, *mixed-motive games,* player preferences conflict only partially, giving players incentives to cooperate and compete. Players in mixed-motive games base

[49] Keith Hipel, Muhong Wang, and Niall Fraser, "Hypergame Analysis of the Falkland/Malvinas Conflict," *International Studies Quarterly* 32, no. 3 (1988): 335-358.

[50] Robert Keohane, *After Hegemony: Cooperation and Discord in the World Political Economy* (Princeton: Princeton University Press, 1984), 51.

their respective actions on *how they expect* others to act. Since cooperation presupposes the existence of opportunities for mutual gain, this thesis models the security dilemma as a *non-zero-sum, mixed-motive game*.[51]

Since it consists of opportunities for mutual gain, the security dilemma is, in game-theoretic terms, a *non-zero-sum game*. Non-zero-sum games such as Deadlock, where self-interest and mutual benefit align, and Chicken, where the main barrier to cooperation is fear of disrepute, have payoff structures that conflict with the security dilemma.[52] Thus, this thesis models the security dilemma using Prisoner's Dilemma (PD) and Stag Hunt (SH), or *collective action problems* and *coordination games*, respectively.[53]

The game-types differ in the number of pure-strategy Nash equilibria. An outcome is *stable* (an *equilibrium*) when no player has an incentive to alter their strategy given the strategies of other players. A Nash equilibrium (NE) is a specific type of equilibrium where players, after taking the strategies of their opponents into account, develop their 'best reply' to their opponent's selected strategy. An outcome is a NE if, given the strategies of others, neither player has anything to gain from changing their strategy. PDs consist of a singular NE (mutual defection) whereas SHs consist of two NEs (mutual defection and mutual cooperation).

Prisoner's Dilemma

		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	3,3	0,5
	a_2 Defect	5,0	**1, 1***

Stag Hunt

		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	**3, 3***	0,2
	a_2 Defect	2,0	**1, 1***

Figure 2: Mixed-motive Games - Prisoner's Dilemma (L) and Stag Hunt (R)

		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	3,3	0,5
	a_2 Defect	5,0	**1, 1***

Figure 2.1: Prisoner's Dilemma

[51] Keohane, 51.

[52] M. Shahrabi Farahami and Majid Sheikmohammady, "A Review on Symmetric Games: Theory, Comparison, and Applications," *International Journal of Applied Operational Research* 4, no. 3 (2014): 98-100, 102-104.

[53] Robert Jervis, "Cooperation Under the Security Dilemma," *World Politics* 30, no. 2 (1978): 171.

$N = \{A, B\}$
$S_A = \{a_1, a_2\}, S_B = \{b_1, b_2\}$
$U = S_A \times S_B = \{[3, 3], [5, 0], [0,5], [1,1]\}$
$R_A = U \rightarrow V_A, V_A = \{[a_2, b_1], [a_1, b_1], [a_2, b_2], [a_1, b_2]\}\ or\ DC > CC > \boldsymbol{DD} > CD \therefore \boldsymbol{D}$
$R_B = U \rightarrow V_B, V_B = \{[a_1, b_2], [a_1, b_1], [a_2, b_2], [a_2, b_1]\}\ or\ DC > CC > \boldsymbol{DD} > CD \therefore \boldsymbol{D}$
$R_A = U \rightarrow V_A, V_A = \{[5,0], [3,3], [1,1], [0,5]\}$
$R_B = U \rightarrow V_B, V_B = \{[0,5], [3,3], [1,1], [5,0]\}$

A Prisoner's Dilemma (PD) is a non-cooperative game where cooperation is *Pareto-optimal* albeit *unstable* while mutual defection is *stable* but *Pareto-suboptimal. Stable* outcomes correspond to a game's equilibria where neither player has an incentive to change strategies. An outcome is *Pareto-sub-optimal* if there is another outcome in the payoff matrix that would improve the payoffs of at least one player without reducing the payoffs of others. Conversely, an outcome is *Pareto-efficient* (or *Pareto-optimal*) if there is no alternative outcome that would benefit at least one player without making the other player worse off. In PD, cooperation is Pareto-optimal and is, thus, a *Pareto-improvement* over the game's singular Nash Equilibrium: mutual defection.

In PD, unilateral defection yields a higher payoff than mutual cooperation. Thus, players can always improve their payoffs by adopting a non-cooperative strategy. S*ince both players always have an incentive to defect*, mutual cooperation is unstable. For instance, note that for A, $[a_2, b_1] > [a_1, b_1]$, since $5 > 3$, and for B, $[a_1, b_2], > [a_1, b_1]$, since $5 > 3$. In PD, players are always better off, irrespective of what their opponents do, to defect. Players are simultaneously tempted to *exploit* and defect to *prevent exploitation*, which in PDs, is modelled by the *sucker payoff*, where one player cooperates while the other defects.[54] In **Figure 2.1**, $[a_1, b_2]$ and $[a_2, b_1]$ are the sucker payoffs for players A and B, respectively. At these outcomes, A and B would receive a payoff of 1. For both players, defection constitutes a *strictly dominant* strategy, which is the best move that they can make regardless of what the other does. In PD, mutual defection is not only the game's single equilibrium but is also a Nash Equilibrium.[55] Players in PD have a strictly *dominant strategy to defect.*

According to the *dominance principle* of Classical Game Theory*,* a rational player should never play a dominated strategy, which, in PDs, is cooperation. However, had the players cooperated instead of following their dominant strategy to defect, they both would have received a higher payoff than they would have had they acted rationally. In PDs, players prefer a Pareto-optimal outcome that they can only obtain by acting irrationally. *Individual rationality* in PD thereby results in a *collectively sub-optimal* outcome. Thus, PD is a type of *collective action problem* where players defect even though they would all have been better off cooperating. In collective action problems, individual rationality results in collective irrationality with individual interests conflicting with group interests: a result emblematic of the security dilemma.

[54] **A** [5,0] , **B**[0,5]

[55] John Nash, "Non-Cooperative Games," *The Annals of Mathematics* 54, no. 2 (1951): 286-295.

		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	$\mathbf{3,3*}$	0,2
	a_2 Defect	2,0	$\mathbf{1,1*}$

Figure 2.2: Stag Hunt

$N = \{A, B\}$
$S_A = \{a_1, a_2\}, S_B = \{b_1, b_2\}$
$U = S_A \times S_B = \{[3,3], [2,0], [0,2], [1,1]\}$
$R_A = U \rightarrow V_A, V_A = \{[a_1, b_1], [a_2, b_1], [a_2, b_2], [a_1, b_2]\}\ or\ \boldsymbol{CC} > DC > \boldsymbol{DD} > CD$
$R_B = U \rightarrow V_B, V_B = \{[a_1, b_1], [a_1, b_2], [a_2, b_2], [a_2, b_1]\}\ or\ \boldsymbol{CC} > DC > \boldsymbol{DD} > CD$
$R_A = U \rightarrow V_A, V_A = \{[3,3], [2,0], [1,1], [0,2]\}$
$R_B = U \rightarrow V_B, V_B = \{[3,3], [0,2], [1,1], [2,0]\}$

Unlike PD, Stag Hunt (SH) is a type of coordination game. In SH, each player's best strategy depends on *what they think the other will do.* Alternatively, in PDs, their best strategy is to defect *irrespective of what their opponent does*. So long as neither player thinks that the other will defect, neither player in SH has an incentive to defect. As the cost of defecting for any given player outweighs the benefit they expect to gain should they defect while their opponents cooperate, players will always prefer to cooperate if they believe that others will cooperate.[56] Thus, the strategy any given player prefers is predicated upon their beliefs regarding how their opponent will likely act. Unlike players in PDs, players in SH have a *contingent,* as opposed to a *dominant,* strategy. In SH, players will not select a strategy without first considering which strategy their opponents will likely play. Players will cooperate if they expect others to cooperate and defect if they expect others to defect.

Like in PDs, mutual cooperation in SH is Pareto-optimal. However, unlike PDs, SHs consist of two NE: mutual cooperation and mutual defection. Between the two NE, mutual cooperation, $[a_1, b_1]$, is the *payoff-dominant* outcome since it corresponds to the highest payoff for both players while mutual defection, $[\mathrm{a}_2, \mathrm{b}_2]$, is the game's *risk-dominant* outcome since it minimizes the sucker payoff for both players. In SH, mutual defection is more likely to occur when players are uncertain of the strategy others will play. The *more uncertain* players are of how others will act the *more likely* that each player will try to minimize risk, resulting in mutual defection, a stable but inefficient outcome.

Security Dilemmas as Collective Action Problems

This thesis argues that the security dilemma is a *collective action problem* best modelled by the Prisoner's Dilemma (PD). The principal reason why this thesis defines the security dilemma as a PD and not as an SH lies in how the former addresses *future uncertainty.* Future uncertainty makes defection a *dominant strategy*. Since states try to self-insure not only against *near-term* but also *future* exploitation, *it is always better,* regardless of what others do *now*, to arm. Although states are driven fundamentally by loss aversion, they capitalize on opportunities to increase their relative gains to protect themselves from potential exploitation *in the future*. Absent assurances against exploitation and ways to alleviate future

[56] Martin Osborne, *An Introduction to Game Theory* (New Dehli, India: Oxford University Press, 2004).

uncertainty, it is *always* in the best interest of states to arm regardless of what others do. Since it lacks dominant strategies, SHs cannot account for the effect future uncertainty has on behaviour.

However, if states can come to believe that other states within the system will disarm—or believe that it is not *always* better to arm—then the security dilemma can *turn into* a coordination game. The first step in resolving the security dilemma, then, which will be explored in Chapter III, involves eliminating the dominant strategy to defect, replacing it with a *contingent strategy,* which transforms the collective action problem into a coordination game. The second step involves preventing coordination failures by reducing the uncertainty states have of each other's intentions.

Alternative Game-theoretic Models

Since Neorealism and Neoliberalism fall under the rationalist tradition, it is unsurprising that much of the scholarship on Game Theory within IR, a sub-set of RCT, seeks to substantiate those two theories. Notable contributions to the realist game-theoretic literature include the utility-based, cost-benefit analyses of Thomas Schelling's *The Strategy of Conflict*,[57] Kenneth Boulding's *Conflict and Defense*,[58] and Anatol Rapoport's *Fights, Games, and Debates.*[59] Moreover, neoliberal institutionalists, such as Robert Axelrod,[60] Robert Keohane,[61] Arthur Stein,[62] and Charles Lipson,[63] have used repeated, mixed-motive games, especially Stag Hunts, to explore how institutions can facilitate coordination.

Rationalist theorists have since used Game Theory to examine a variety of issues ranging from military engagements,[64] debt restructuring,[65] and treaty negotiations.[66] Comparably lacking in the current IR literature is constructivist scholarship on Game Theory—not least on its applications to the security dilemma.[67] This lacuna stems from the rationalist framework of Game Theory centered on the principles of *complete information* and *pure rationality.* Together, these principles oversimplify state behaviour in ways incommensurate withy Constructivism's inter-subjective approach.

According to the complete information principle, players know, and know that others also know, the strategies, preferences, and options defining the game. Similarly, rationalist theories of IR assume that states 'see' the world as others view it—that states believe others see the world as they do. However, since states can only perceive situations from their own *subjective* point of view and are constrained by future uncertainty, games of complete information cannot explain why the security dilemma.

The principle of perfect rationality assumes that states are rational insofar as they act in ways that maximize their utility. However, in the security dilemma, individual rationality only results in collective irrationality. States that seek to maximize their utility independently of what others do only arrive at sub-

[57] Thomas Schelling, *The Strategy of Conflict* (Cambridge: Harvard University Press, 1981).
[58] Kenneth Boulding, *Conflict and Defense: A General Theory* (Literary Licensing, LLC, 2012).
[59]Anatol Rapaport, *Fights, Games, and Debates* (University of Michigan Press, 1974).
[60] Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984).
[61] Robert Keohane, *After Hegemony: Cooperation and Discord in the World Political Economy* (Princeton University Press, 2005).
[62] Arthur Stein, *Why Nations Cooperate: Circumstance and Choice in International Relations.* (Ithaca: Cornell University Press, 1990).
[63] Charles Lipson, *Reliable Partners: How Democracies Have Made a Separate Peace.* (Princeton University Press, 2013).
[64] Bruce Bueno de Mesquita, *The War Trap* (New Haven, CT: Yale University Press, 1981).
[65] Rohan Pitchford and Mark Wright, "On the Contribution of Game Theory to the Study of Sovereign Debt and Default," *Oxford Review of Economic Policy* 29, no. 4 (2013).
[66] Harold Kuhn, "Game Theory and Models of Negotiation," *The Journal of Conflict Resolution* 6, no. 1 (1962).
[67] Friedrich Kratochwil, *Rules, Norms, and Decisions* (New York, NY: Cambridge University Press, 1989).

optimal outcomes. In the security dilemma, states receive *less utility* than they would have had they acted irrationally. Separately, the rationality principle does not permit different definitions of utility; it presumes that states define utility identically and only in terms of *self-interest.* They ignore how extrinsic factors—such as norms and value systems—affect how states calculate and define utility.

By the 1960s, game theoretic methodology began to shift from single-play games of complete information to repeated games of incomplete information with bounded rationality. Game-theoretic models also began to differentiate actors by their preferences. For example, Duncan Luce and Howard Raiffa's formalization of the *subjective utility function* emphasized how preferences can vary between players.[68] Moreover, John Harsanyi's so-called Bayesian games explore how various *player types*—each defined by a unique set of preferences—adjust their behaviour over time.[69]

Like Luce and Raiffa, and Harsanyi, this thesis assumes that states define their preferences using different subjective utility functions—with each function corresponding to a specific *type* of player. With this mind, it uses three games of incomplete information and bounded rationality to model the security dilemma.[70] This chapter uses hypergames to model how identity affects strategic choice, formalizing the *identity-preference* relationship. Chapter III elaborates on this relationship using games of conditional reciprocity and costly signalling games.

This chapter uses hyper-games to examine how preferences vary between different player types and how these differences affect interaction. It also explores how higher-order beliefs about player types affect behaviour. This chapter then combines hypergames with Kelley and Thibault's *given-effective matrix model* to show how different player types perceive interactions differently. The purpose of this chapter is to explore how identity mediates strategic choice and formalize the preference-identity relationship—or how preferences and identities are *linked.* This chapter establishes the groundwork for Chapter III, which proposes a solution to the security dilemma.[71] States can overcome the security dilemma by changing their identities in ways that make them come to *prefer* cooperation over competition.

Hypergames

Threat perception in the context of security dilemmas is a function of the higher-order beliefs states have of the *type of game* being played, *who* the relevant states are, and *how* their actions affect how other states within the system behave. The hypergame model of strategic interaction decomposes a game of incomplete information into a set of *concurrent, interdependent sub-games*.[72] The sub-games are *concurrent* in that they occur simultaneously, and *interdependent* in that the outcome of any given sub-game affects the outcome of other sub-games. Each sub-game corresponds to a particular player's perception of the relevant actors, strategies, and outcomes of a given interaction.

Hypergames have been used to model strategic interaction in a variety of contexts ranging from sport fan behavior[73] to crisis decision-making and arms races.[74] However, the literature has yet to see a hypergame analysis of the security dilemma. This thesis fills this lacuna by using hypergames to explain how collective action and coordination problems stem from misperceptions regarding player types and the

[68] Duncan Luce and Howard Raiffa, *Games and Decisions* (John Wiley & Sons, Inc, 1957).
[69] John Harsanyi, "Game Theory and the Analysis of International Conflict." *The Australian Journal of Politics and History* 11 (1965): 292-304.
[70] Masao Takahashi, Nial Fraser, and Keith Hipel, "A Procedure for Analyzing Hypergames," *European Journal of Operational Research* 18 (1984): 113.
[71] John Kelley and Harold Thibault, *The Social Psychology of Groups* (New Jersey: Transaction Publishers, 1959).
[72] Peter Bennett, "Toward a Theory of Hypergames," *OMEGA* 5, 1977: 749–751.
[73] Peter Bennett, "Using Hypergames to Model Difficult Social Issues: An Approach to the Case of Soccer Hooliganism," *Journal of the Operational Research Society* 31, no. 7 (1980): 621-635.
[74] Peter Bennett, "The Arms Race as a Hypergame," *Futures* 14, no. 4 (1982): 293-306.

game writ large. It examines how identities mediate strategic choice by showing how *different player types* perceive the same situation differently and how the ways in which players perceive others affect how they think they will act.

Hypergame theory assumes that players conceptualize interaction differently, playing their own unique sub-game within the context of a broader, overarching game. The *hypergame* refers to the overall game defining the interaction while each *sub-game* refers to how a specific player perceives that game. In hypergames*,* players view the game differently, with at least one player having incomplete information about the overarching game. Breaking down interaction into discrete, subjective sub-games, hypergame theory models how perception mediates strategic behaviour since it shows how players understand the game only through their respective sub-games.[75] While in NMGs, all players view the game identically, thereby effectively playing the same game, in hypergames, players play different sub-games within the context of a broader game.

Recall from **Figure 1:** A game $G$is defined by a set of preference vectors, V_n, for all player in N. That is: $G = \{V_1, V_2, \dots V_n\}$

A hypergame between two players is A and B is defined as: $H_1 = \{G_\mathrm{A}, G_\mathrm{B}\}$

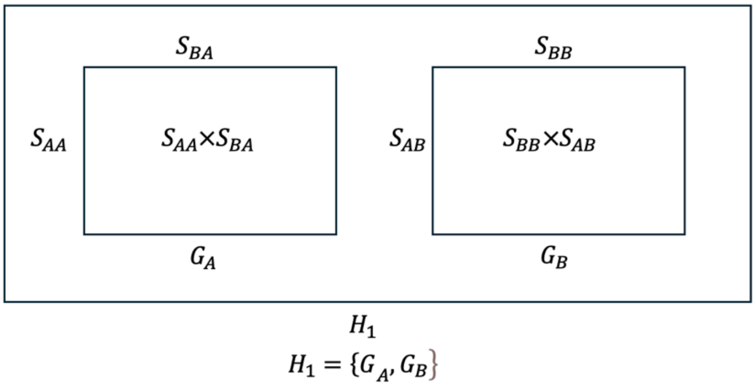


Figure 3: Basic Hypergame Model

In a two-player hypergame between states A and B, A's game is analyzed from A's interpretation of the situation while B's game is analyzed from B's point-of-view. The decisions A and B make depend not just on how they interpret their own payoffs but also on how they think the other is perceiving the game. Thus:

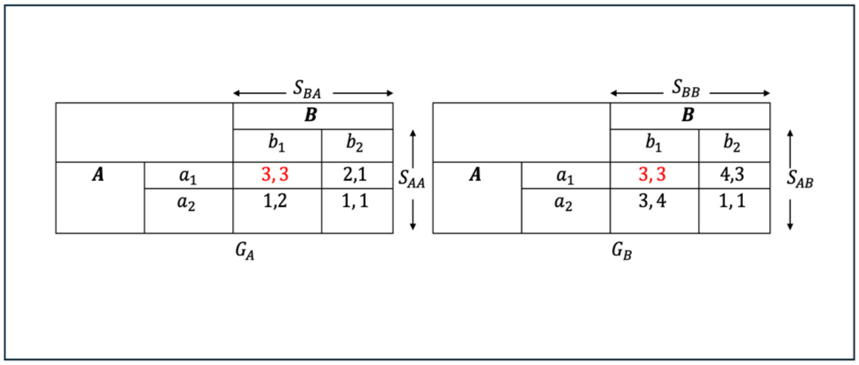
$G_A = \{V_A, V_B^A\}$: V_B^Arepresents $B's$ preference vector as perceived by A.
$G_B = \{V_B, V_A^B\}$: V_A^Brepresents $A's$ preference vector as perceived by B.

Note: A is said to have misperceived B if $V_B \neq V_B^A$

[75] Masao Takahashi, Nial Fraser, and Keith Hipel, "A Procedure for Analyzing Hypergames," *European Journal of Operational Research* 18 (1984): 113.

Hence, a two-person hypergame H, played by players A and B, is defined as:

$H_1 = \{(G_A, G_B^A), (G_B, V_A^B)\}$



		B	
		b_1	b_2
A	a_1	3,3	2,1
	a_2	1,2	1,1

		B	
		b_1	b_2
A	a_1	3,3	4,3
	a_2	3,4	1,1

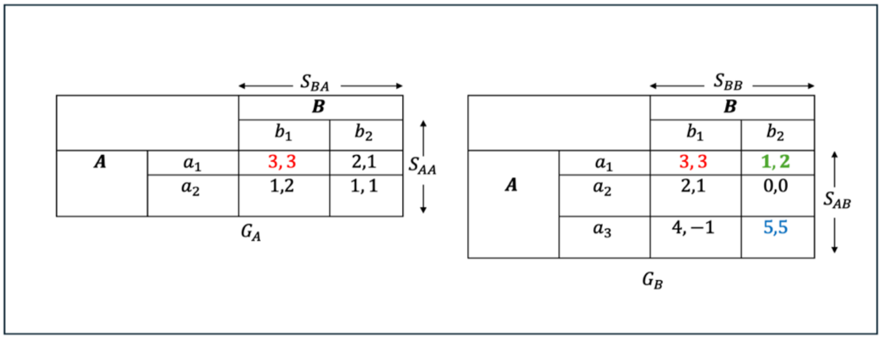
H_1

$H_1 = \{G_A, G_B\}$

Figure 4: Example of a Two-person Hypergame

The sub-games comprising the hypergame are *interdependent* in that the outcome of each sub-game affects the *outcome of other sub-games*.[76] The outcome of each sub-game affects the outcome of *other* sub-games, which, taken together, determine the outcome of the *overall* interaction.[77] Thus, if sub-games differ structurally from not only each other but also the hypergame itself, the equilibria players see will likely vary.[78] Moreover, although players may achieve Pareto-optimal outcomes in their *respective sub-games*, they may not acquire Pareto-optimality in the *overall* game since the game's overall payoff is a function of the outcomes of the sub-games of all the players. Players that act *irrationally* in the context of the overall game may be acting *rationally* in the game that they perceive. In this context, despite acting rationally within their own sub-games, players may end up with payoffs contrary to what they intended to obtain. This explains why players arrive at an individually *rational* but collectively *irrational* outcome in collective action problems such as the security dilemma.[79] To illustrate the nexus between sub-game and game equilibria, consider **Figure 5.**

[76] Masao Takahashi, Niall Fraser, and Keith Hipel, "A Procedure for Analyzing Hypergames," *European Journal of Operations Research* 18, no. 1 (1984): 111–122.
[77] Inohara, Takahashi, and Nakano, "Integration of Games and Hypergames Generated from a Class of Games," *Journal of the Operational Research Society* 48 (1997): 430.
[78] Yossi Feinberg, "Games with Awareness," *Stanford Graduate School of Business,* no. 2122. (2012): 26.
[79] Inohara, Takahashi, and Nakano, "Integration of Games and Hypergames Generated from a Class of Games," *Journal of the Operational Research Society* 48 (1997): 423-432. 430.



S_{BA}
S_{AA}
G_A
S_{BB}
S_{AB}
G_B

		B	
		b_1	b_2
A	a_1	3, 3	2,1
	a_2	1,2	1, 1

		B	
		b_1	b_2
A	a_1	3, 3	1, 2
	a_2	2,1	0,0
	a_3	4, −1	5,5

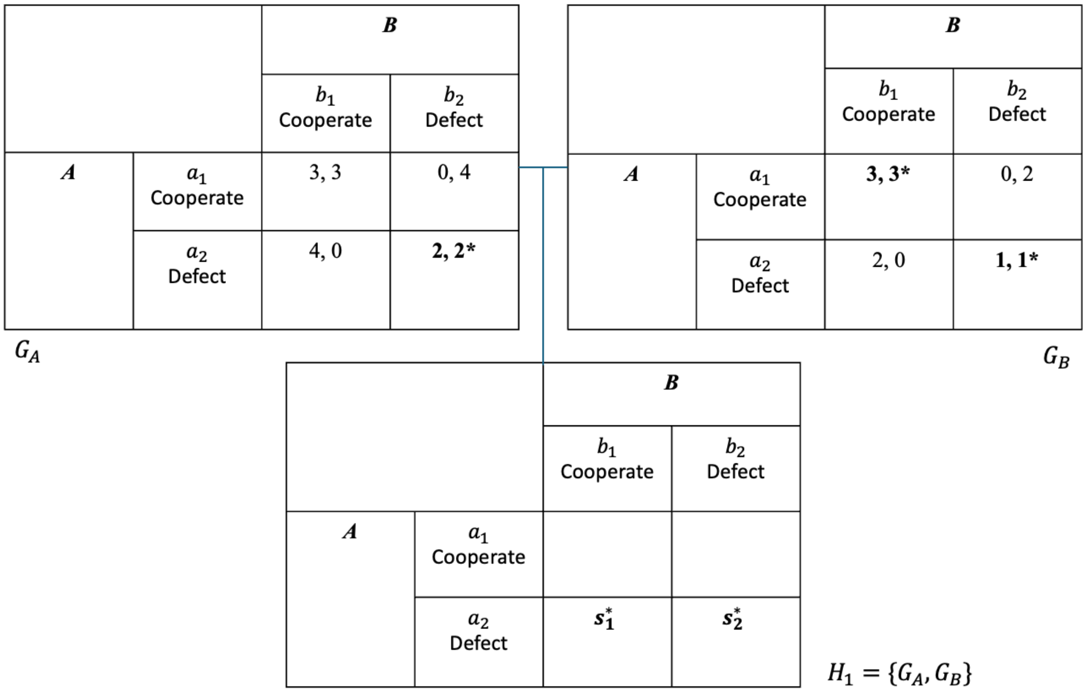
H_1

$H_1 = \{G_A, G_B\}$

Figure 5: Two-person Hypergame - Rational Individuality Results in Collective Pareto-Inefficiency

In this example, A expects B to play b_1, thus A plays a_1, obtaining a Nash Equilibrium at (a_1, b_1), which corresponds to the payoff (3,3). However, B perceives the game differently. B expects A to play a_3, its dominant strategy, so B plays b_2. If player B's perception of the game matches the game *proper* then what results from this interaction is (a_1, b_2), which is unstable and Pareto-inefficient. Although players may act rationally, and obtain *stable* outcomes within their respective sub-games, players may arrive at a *mutually sub-optimal, overall* outcome. Despite each player acting rationally within their respective sub-games, players can obtain a Pareto-inefficient outcome—much like in the security dilemma.

Security dilemmas can thus occur when players perceive the game differently. Consider **Figure 6**. Assume that A sees the situation as a PD while B sees it as a SH. A assumes that B's dominant strategy is to defect. However, say B believes that A will cooperate. A plays its dominant strategy and defects, not so much out of a desire to exploit B as to avoid exploitation by B. Expecting A to cooperate, B cooperates, resulting in outcome (D, C) for A and B, respectively. B's misplaced trust in A's willingness to cooperate results in B receiving the *sucker payoff,* its worst payoff. Alternatively, should B be uncertain of A's type, it will play the *risk-dominant strategy* and defect, resulting in mutual defection. Moreover, if B believes A is cooperator, then B will cooperate, resulting in B receiving the *sucker payoff* ($\boldsymbol{s}_1^*$). Alternatively, if B is uncertain of A's type, B will defect, resulting, much like in security dilemmas, in mutual defection $\boldsymbol{s}_2^*$**.**



		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	3, 3	0, 4
	a_2 Defect	4, 0	**2, 2***

G_A

		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	**3, 3***	0, 2
	a_2 Defect	2, 0	**1, 1***

G_B

		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate		
	a_2 Defect	s_1^*	s_2^*

$H_1 = \{G_A, G_B\}$

Figure 6: Two-person Hypergame - Perception Mismatch

The Nash Equilibrium in a hypergame (s_i^*, s_{-i}^*) is a Hyper Nash Equilibrium if, and only if, for any player i, s_i^* is a Nash Equilibrium of player i's subjective game. For instance, the hypergame depicted in Figure 6, consists of two Hyper Nash Equilibria: mutual defection and mutual cooperation.

Modelling Higher-order Perceptions

Classical game-theoretic models do not examine how higher-order beliefs mediate interaction. However, in international politics, states base their decisions on not only what they *think others will do* but also on what *they think others think they will* do. The strategy that a state selects depends not only on its perception of the world but also on its perception of how other states within the system perceive the world and what they believe *others* think about its own perceptions. Hyper-game models analyze strategic interaction in terms of the higher-order beliefs players have of each other and the game at play.

In *zero-order hypergames*—which are analogous to NMGs—each player's respective subgame aligns with the overall game. Alternatively, in *first-order hypergames*, which this thesis uses to model the security dilemma, players are unaware that they are perceiving the game differently (**Figure 6**). Moreover, in *second-order hypergames*, at least one player is aware that at least another player is perceiving the game differently, thereby effectively playing dissimilar sub-games.[80] Essentially, while in first-level hyper-games players unknowingly hold different perceptions of the same game, in second-level hyper-games, there is *at least one* player that knows that different games are at play and that misperceptions exist.

Player Types

A key limitation of systemic theories in IR lies in their inability to differentiate between states. Neorealists conceptualize states as *functionally identical* units; they differentiate units not by their

[80] Muhong Wang, Keith Hipel, and Niall Fraser, "Misperceptions and Hypergame Models of Conflict," *Behavioral Science* 33, no. 3 (1988): 207-223.

attributes but by their greater-or-lesser capabilities.[81] Although states are alike in that they seek self-preservation, they differ in their ability to achieve security.[82] States are, thus, differentiable only by their material position(s) within the international system. Rationalist theories ignore how states within the *same distribution of power* can possess different preferences. NMGs similarly treat player preferences as both fixed and universal.

Using hypergames and Kelley and Thibault's so-called *given-effective matrix dichotomy*, this thesis shows how different *types* of players with *type-specific* preferences interact, thereby showing how preferences are linked inextricably to identities, formalizing the identity-preference relationship.[83] It assumes that states may define their preferences using different utility functions—where each function corresponds to a specific *type* of player.[84] Like hypergames, Kelley and Thibault's given-effective matrix model shows how perceptions mediate strategic choice. The dichotomy decomposes games into an objective, *given* matrix and multiple subjective, *effective* matrices.[85] Each *effective* matrix represents a specific player's perception of the given matrix, which many not necessarily reflect the game's actual payoffs. Each effective matrix consists of a specific, subjective utility function.

Recall that a preference vector $V_i = \{U, R_i\}$ for some player i consists of a certain relation R_i, which ranks all the outcomes $z \in U$ from most-to-least preferable. [86] Utility functions represent a specific preference-ordering, or *preference vector*. They assign a certain value, $u(z)$, to each alternative z, such that, for any two alternatives z and z', if $u(z) \geq u(z')$, then $z \geq z'$, z is preferrable to z'. That is, if the utility of z is higher than the utility of z', then, assuming perfect rationality, the player will always prefer z over z'. Neorealists rank alternatives by their *relative* difference to the corresponding alternative in the *other* state's option set while neoliberalists rank them by their *absolute* values—that is, without reference to external option sets. The thesis broadens the set of feasible preference vectors to include one that ranks outcomes by their *joint value.* To that end, it explores three types of players: competitors, individualists, and cooperators. *Competitors* calculate utility using a rule of *difference maximization*; they seek to maximize their relative gain and minimize relative loss.[87] *Individualists* care only about maximizing their *absolute gains* while *cooperators* calculate utility based on *total joint gains.*

The thesis uses hypergames and the Kelley and Thibault's given-effective matrix dichotomy to model how perceptions of *player type* affect how players play the game and how different player types interpret a situation differently. Each player type bases their decisions on a type-specific effective matrix corresponding to a unique preference-ordering. Each type defines their preferences using a unique utility function.[88]

[81] Kenneth Waltz, *Theory of International Politics* (Boston, Mass.: McGraw-Hill, 1979), 97.
[82] Waltz, 96-97.
[83] John Kelley and Harold Thibault, *The Social Psychology of Groups* (New Jersey: Transaction Publishers, 1959).
[84] Charles McClintock and Wim Liebrand, "Role of Interdependence Structure, Individual Value Orientation, and Another's Strategy in Social Decision Making: A Transformational Analysis," *Journal of Personality and Social Psychology* 55, no. 3 (1988).
[85] John Kelley and Harold Thibault, *The Social Psychology of Groups* (New Jersey: Transaction Publishers, 1959).
[86] Michael Roloff, *Interpersonal Communication: The Social Exchange Approach* (California: Sage Publications,1981), 51.
[87] McClintock, Charles. "Motivational Bases of Choice in Three-Choice Decomposed Games." *Journal of Experimental Social Psychology* 9, no. 6 (1973): 575.
[88] Charles McClintock and Wim Liebrand, "Role of Interdependence Structure, Individual Value Orientation, and Another's Strategy in Social Decision Making: A Transformational Analysis," *Journal of Personality and Social Psychology* 55, no. 3 (1988).

Table 1: Typology of States

Assume that two states, A and B must choose between two outcomes: $(a_n, b_m)$and $(a_n, b_m)'$.

Order any two pairs $U_A(a_n, b_m)$ and $U_A(a_n, b_m)'$ by saying that $U_A(a_n, b_m)$ is preferred to $U_A(a_n, b_m)'$, that is, $(a_n, b_m) > U_A(a_n, b_m)'$.

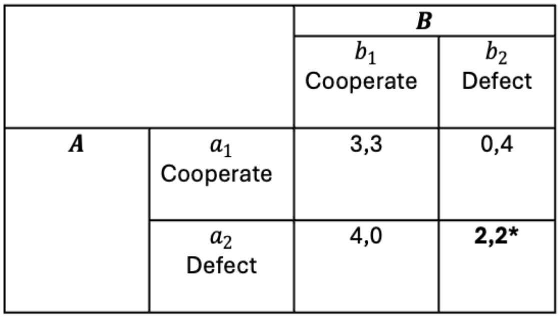
	Player Type	**Subjective Utility Function**
1	Competitors	Assume A is a competitor. A seeks to maximize the difference between its utility and B's. Thus: $U_A(a_n, b_m) > U_A(a_n, b_m)'$ if $[U_A(a_n, b_m) - U_B(a_n, b_m)] > [U_A((a_n, b_m)') - U_B((a_n, b_m)')]$ Difference maximization also corresponds with loss minimization.
2	Individualists	Assume A is an individualist. A only seeks to maximize its own utility. Thus: $U_A(a_n, b_m) > U_A(a_1, b_1)'$
3	Cooperators	Assume A is a cooperator. A seeks to the maximize the joint utility it has with B. Thus: $U_A(a_m, b_n) > U_A(a_m, b_n)'$ if $[U_A(a_m, b_n) + U_B(a_m, b_n)] > [U_A(a_m, b_n)' + U_B(a_m, b_n)]$

The type of player a player *is* not only affects how it calculates utility but also how it perceives the player types of others. Given two players A and B, the type of player A *is* determines A's preferences and what A believes about state B's type will depend on the type of state A is. For instance, competitors are less likely than cooperators and individualists to cooperate regardless of what others do while the opposite is true for cooperators.[89] Competitors are also more likely than cooperators to view others of being the same type.[90] Moreover, of the three types, only competitors have a *dominant strategy to defect*. If A is a competitor looking to maximize relative gain while also minimize relative loss and believes that B is also a competitor, A would deduce that B would defect as per its dominant strategy. A thus defects to avoid the sucker payoff.

Figure 8 illustrates why states arm in the security dilemma; it shows how the fears of exploitation pressure states to defect pre-emptively. Moreover, it shows that mutual cooperation cannot occur if there is at least one player that *believes* that the other is a competitor. The outcome of the hypergame depicted in **Figure 8**, modelled using the given-effective matrix, is a function of one state's misperception of another state's intentions and *vice versa.*

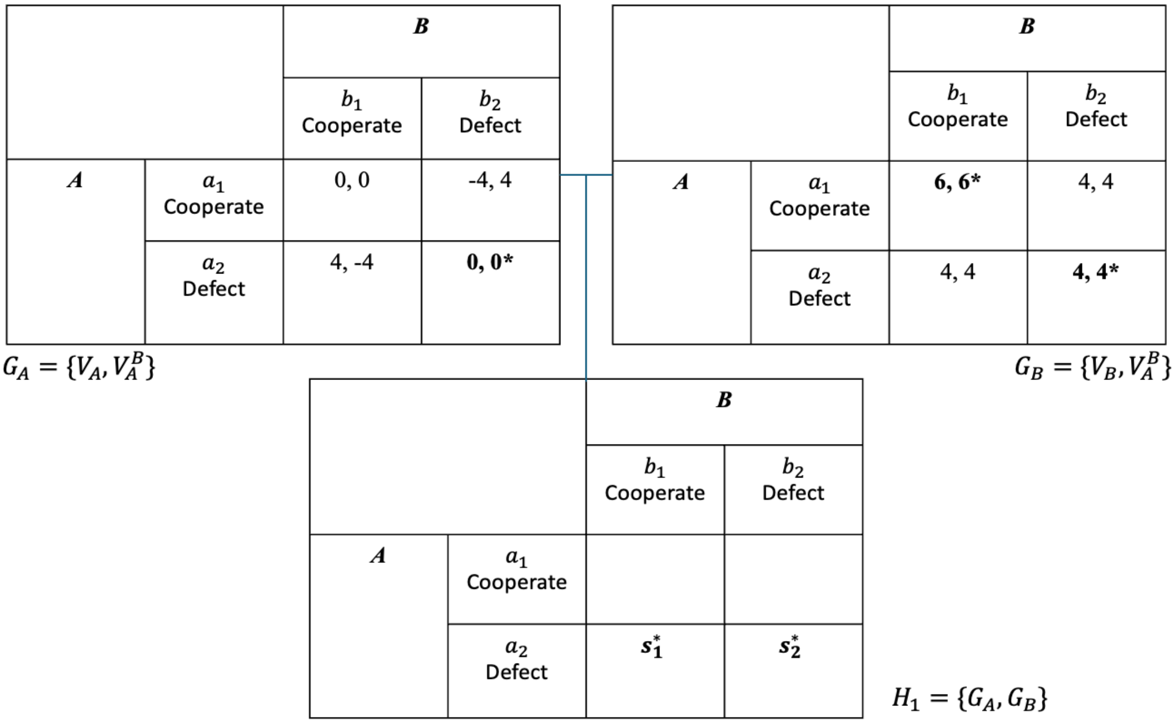
[89] Paul Van Lange, "The Pursuit of Joint Outcomes and Equality in Outcomes: An Integrative Model of Social Value Orientation," *Journal of Personality and Social Psychology* 77, no. 2 (1999): 338.

[90] Harold Kelly and Anthony Stahelski, "Social Interaction Basis of Cooperators' and Competitors' Beliefs About Others," *Journal of Personality and Social Psychology* 16, no. 1 (1970): 66-91.



		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	3,3	0,4
	a_2 Defect	4,0	**2,2***

Figure 7: Given Matrix - Prisoner's Dilemma



		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	0, 0	-4, 4
	a_2 Defect	4, -4	**0, 0***

$G_A = \{V_A, V_A^B\}$

		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	**6, 6***	4, 4
	a_2 Defect	4, 4	**4, 4***

$G_B = \{V_B, V_A^B\}$

		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate		
	a_2 Defect	$\boldsymbol{s_1^*}$	$\boldsymbol{s_2^*}$

$H_1 = \{G_A, G_B\}$

Figure 8: Effective Matrix - Prisoner's Dilemma – Mismatched Perceptions

A believes B is a competitor, but B is a cooperator who believes that A is also a cooperator, resulting in two different effective matrices.

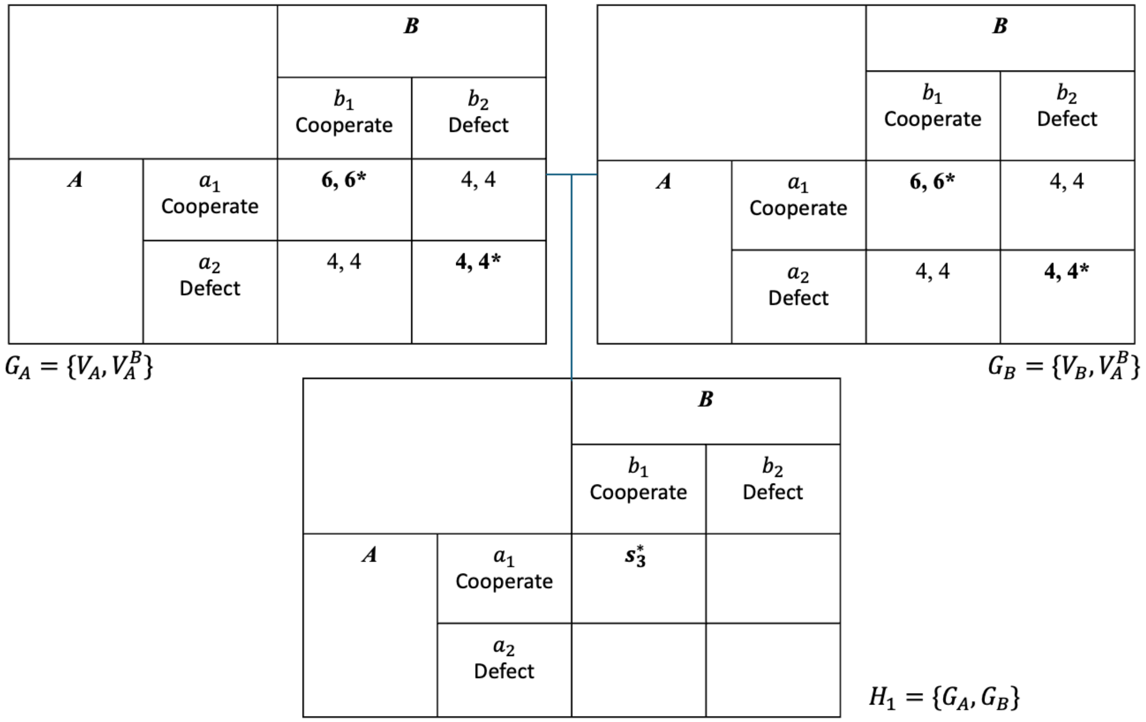
$R_A = U \rightarrow V_A, V_A = \{[a_1,b_2],[a_1,b_1],[a_2,b_2],[a_1,b_2]\}\ or\ DC > CC > \boldsymbol{DD} > CD$
$R_A = U \rightarrow V_A, V_A = \{[4,-4],[0,0],[0,0],[-4,4]\}$
$V_B^A = \{[a_1,b_2],[a_2,b_2],[a_1,b_1],[a_2,b_1]\ or\ DC > \boldsymbol{DD} \geq CC > CD \therefore \boldsymbol{D}$

$R_B = U \rightarrow V_b, V_b = \{[a_1,b_1],[a_1,b_2],[a_2,b_2],[a_2,b_1]\}\ or\ \boldsymbol{CC} > DC \geq DD \geq CD \therefore \boldsymbol{C}$
$R_B = U \rightarrow V_b, V_b = \left\{[6,6],[4,4]]\right\}$
$V_A^B = \{[a_1,b_1],[a_1,b_2],[a_2,b_2],[a_2,b_1]\}\ or\ \boldsymbol{CC} > DC \geq DD \geq CD \therefore \boldsymbol{C}$

B's game consists of two NE: mutual defection and mutual cooperation. Thinking that A is a cooperator, B cooperates. A's game consists of only one Nash equilibrium: mutual defection. Thinking that B will compete, A deduces that B's dominant strategy is to defect. Hence, A defects. The Hyper Nash equilibrium of (a_2, b_1) or $\boldsymbol{s_1^*},$ where B receives the sucker payoff, obtains. Alternatively, if B believes that A is a competitor, then B will defect, resulting in the Hyper Nash equilibrium (a_2, b_2) or $\boldsymbol{s_2^*}.$ To hedge against the risk of exploitation, it would be in B's best interest to *assume* that A is a competitor and to defect pre-emptively.

Moreover, **Figure 9** suggests that, within the security dilemma, cooperation can only occur *between cooperators* and that the existence of cooperators is a *necessary condition* for overcoming security dilemmas. Competitors could relinquish their dominant strategy to defect if they turned into cooperators. This transformation in player type, which will be explored, in Chapter III, using conditional games of reciprocity, entails substituting the dominant strategy to defect with a contingent strategy, thereby transforming the *game itself*—from a collective action problem into a coordination game.

Although necessary, the existence of cooperators by itself does not guarantee that cooperation will occur in coordination games. The players must also *trust* that others will cooperate, presupposing a mechanism by which states can reduce uncertainty of each other's intentions. In coordination games, cooperators will cooperate only if they believe others will also cooperate. For both competitors and cooperators, their perceptions of *who* their opponents are impact how they behave.



		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	**6, 6***	4, 4
	a_2 Defect	4, 4	**4, 4***

$G_A = \{V_A, V_A^B\}$

		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	**6, 6***	4, 4
	a_2 Defect	4, 4	**4, 4***

$G_B = \{V_B, V_A^B\}$

		B	
		b_1 Cooperate	b_2 Defect
A	a_1 Cooperate	$\boldsymbol{s_3^*}$	
	a_2 Defect		

$H_1 = \{G_A, G_B\}$

Figure 9: Effective Matrix - Prisoner's Dilemma – Aligned Perceptions

In this scenario, A and B are both cooperators and see each other as such.

$R_A = \ U \ \rightarrow V_A, V_A = \{[a_1, b_1], [a_1, b_2], [a_2, b_2], [a_2, b_1]\}\ or\ \boldsymbol{CC} > DC \geq DD \geq CD\ \therefore \boldsymbol{C}$
$R_A = \ U \ \rightarrow V_A, V_A = \{[6{,}6], [4{,}4]]\}$
$V_B^A = \{[a_1, b_2], [a_2, b_2], [a_1, b_1], [a_2, b_1]\ or\ \boldsymbol{CC} > DC \geq DD \geq CD\ \therefore \boldsymbol{C}$

$R_B = \ U \ \rightarrow V_b, V_b = \{[a_1, b_1], [a_1, b_2], [a_2, b_2], [a_2, b_1]\}\ or\ \ \boldsymbol{CC} > DC \geq DD \geq CD\ \therefore \boldsymbol{C}$
$R_B = \ U \ \rightarrow V_b, V_b = \{[6{,}6], [4{,}4]]\}$
$V_A^B = \{[a_1, b_1], [a_1, b_2], [a_2, b_2], [a_2, b_1]\}\ \ or\ \ \boldsymbol{CC} > DC \geq DD \geq CD\ \therefore \boldsymbol{C}$

Although in this situation, mutual defection is also stable NE, the payoff each player receives at this outcome is *identical* to the payoff they receive from defecting unilaterally or from being suckered. Hence, so long as they are rational—that is, motivated to choose the option yielding the highest utility—players will cooperate. By reducing the sucker payoff and the gains associated with defecting, cooperation becomes more likely to occur. Eliminating the dominant strategy to defect involves transforming the incentive structure of the payoff matrix. In Chapter III, the thesis explores this transformation in the context of *norm internalization* via games of conditional reciprocity. This scenario also suggests that cooperation is stable only among cooperators and when both states believe that the other is also a cooperator. Chapter III will examine in this dynamic in the context of costly signalling games.

By affecting how players calculate utility and develop preferences, identity has a transformative effect on interaction. Since behaviour is driven by preferences, by mediating preferences, *identity affects interaction.* Should two competitors come to value joint gain maximization, becoming cooperators, then, so long as they trust that the other is also a cooperator, they can overcome the security dilemma. In the next chapter, the thesis explores how competitors transform into cooperators through games of conditional reciprocity and how cooperators can facilitate trust through costly signaling games.

The security dilemma *cannot* consist of individualists as they are indifferent to relative gains, which is central to the problems of future uncertainty and egoism. This thesis argues that the security dilemma can be modelled as only one of *two* types of games*:* a *collective action problem comprised only of* competitors and a *coordination game involving cooperators* burdened by a high level of uncertainty and mistrust. Mutual defection occurs in collective action problem between competitors since they both have a dominant strategy to defect. This dominant strategy stems from their utility function, which is defined by relative difference maximization and difference minimization. In coordination games where cooperators are uncertain about the likelihood of others cooperating, mutual defection occurs because players are unwilling to risk exploitation.

Conclusion

Chapter II explored how higher-order beliefs shape behaviour; it examined how perceptions of player type affect how players play the game and how *different player types* calculate utilities differently. They demonstrate how perception, preferences, and motivations vary across different identities. Chapter II formalizes the identity-preference relationship, proving that states can infer what others *want* from *who* they think others are. Chapter III elaborates on this relationship by examining how preferences and identities *can change.* This thesis argues that states can forgo one identity for another by accepting certain norms and they can—through repeated interactions—update their beliefs about who others *are,* and by extension, what they *want.* These dynamics will be explored in Chapter III using games of conditional reciprocity and costly signalling games, respectively.

CHAPTER III

All the World's a Stage: Overcoming the Security Dilemma

Introduction

In Chapter I, the thesis characterized the security dilemma as a problem of risk aversion. Three principal factors drive this aversion: future uncertainty, egoism, and mistrust. Since states are driven into security dilemmas not by a desire to maximize power but by *loss aversion*, resolving the security dilemma involves reducing these three factors. With this in mind, this thesis proposes a two-step solution to the dilemma. The first step involves transforming the security dilemma, a collective action problem involving *competitors*, into a coordination game comprised of *cooperators*. The second step involves overcoming coordination failures. The thesis uses two game-theoretic models, repeated games of conditional reciprocity, or tit-for-tat, and costly signalling games, to model each step, respectively.

This chapter builds on Chapter II by exploring further—and in the context of the security dilemma—the identity-preference relationship and its effect on interaction.[91] Chapter III consists of two parts. Drawing upon Regime Theory, the first part of this Chapter defines terms relevant to the study of preference change. The second part of this Chapter consists of two sections. Together they explore how resolving security dilemmas entails overcoming loss aversion driven by future uncertainty, egoism, and mistrust.

In the first section, the thesis explores, using games of conditional reciprocity, how states can transform collective action problems into coordination games. Recall that in collective action problems, mutual cooperation is *Pareto-optimal* but *unstable.* As such, states always have an incentive to defect regardless of what others do since unilateral defection yields a higher payoff than cooperation. Loss aversion driven by future uncertainty and egoism reinforces the dominant strategy to defect. Resolving collective action problem involves eliminating this dominant strategy by reducing the sucker payoff and increasing the costs of early defection.

In the second section, the thesis uses costly signalling games to show how states can build trust and achieve cooperation in coordination games. In coordination games, mutual cooperation is *stable* but so too is mutual defection. Players will cooperate only if they believe that others will cooperate. The main barrier to cooperation in coordination games is loss aversion driven by mistrust. Hence, preventing coordination failures involves developing trust.

Towards a Solution to the Security Dilemma

Much of the scholarship on resolving collective action problems falls under the rationalist tradition. Defensive realists, and neoliberal institutionalists, such as Robert Axelrod,[92] Charles Lipson,[93] and Arthur Stein,[94] argue that *institutions* can reduce future uncertainty to a level conducive to long-term cooperation. Alternatively, constructivists propose an *inter-subjective* approach: a solution based on reducing insecurity through the collective internalization of cooperative norms and rules. In this way,

[91] Michael Barnett and Martha Finnemore, *Rules for the World* (Ithaca, NY: Cornell University Press, 2004), 33.

[92] Charles Lipson, "International Cooperation in Economic and Security Affairs," *World Politics* 37, no. 1 (1984): 1-23.

[93] Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984).

[94] Arthur Stein, "Coordination and Collaboration: Regimes in an Anarchic World," *International Organization* 36, no. 2 (1982): 299-324.

Constructivism mirrors Regime Theory*,* which, despite its neoliberalist origins, have since expanded into constructivist circles.[95]

For realists, *material* incentives condition actor behaviour in pursuit of their *a priori* interests. Ignoring the independent causal power of norms, realists seldom explore how norms can *transform* or *create* interests.[96] Unlike Realism, Constructivism and Regime Theory explore how inter-subjective rules, norms, and principles signify, enable, and transform behaviour. Regime theorists such as Stephen Krasner, Robert Keohane, Joseph Nye, and Hedley Bull argue that international regimes can facilitate cooperation by delineating legitimacy and routinizing behaviour.[97] For Bull, regimes are rules that delimit behaviour in specific, prescribed ways.[98] Similarly, for Keohane and Nye, regimes are "governing arrangements [comprised of] rules, norms, and procedures that regularize behaviour and control its effects".[99]

For constructivists and regime theorists, rules give actions meaning; they help agents ascertain, *or make sense of*, of what they, and others "can do [and] cannot do".[100] Rules thereby constrain behaviour in ways that facilitate predictability.[101] By expressing notions of legitimacy, they impose inter-subjective limits on action. *Regulative* rules establish the parameters of some interaction; they are either prescriptive or proscriptive.[102] They take the form of: *in some context C, do action x but not action y*. Alternatively, *constitutive* rules *define* the interaction, delineating what certain actions *mean* and in *what contexts*. They take the form of: *in some context C, x counts as y*. For instance, an arms treaty may contain a rule that defines the act of arming for non-defensive purposes as a violation of the treaty. Simultaneously regulative and constitutive, such a rule prohibits arming for non-defensive purposes while also establishing *what* constitutes a violation and specifying within what *contexts* arming constitutes a violation.

Moreover, constitutive and regulative rules clarify notions of permissibility, effectively defining the parameters of interaction within a specific issue-area. For instance, the 1974 ABM Treaty consisted of both regulative and constitutive rules regarding the development, use, and disposal of nuclear weapons technology. Article V of the Treaty contained a regulative rule prohibiting the production, testing, and deployment of non-static weapon-systems.[103] On the other hand, Article II *defined* an ABM system as a missile system whose express purpose is to counter, intercept, or destroy 'strategic ballistic missiles or their elements in flight trajectory'.[104]

Although Article V of the ABM Treaty proscribed the deployment of space-based ABM systems, in 1983, the Reagan Administration announced that it would build a space-based, missile defence system called the Strategic Defence Initiative (SDI). Citing Agreed Statement D of the Treaty, the United States argued that Article V did not apply to ABM systems based on technology invented *after* 1974. According to Agreed Statement D:

[95] Benjamin Meiches and Raymond Hopkins, "Regime Theory," *Oxford Research Encyclopedia of International Studies.* 2018.

[96] Sara Hellenmuller, Jamie Pring, and Oliver Richmond, "How Norms Matter in Mediation: An Introduction," *Swiss Political Science Review* 26, no. 4 (2020): 349.

[97] Steven Krasner, *International Regimes* (Ithaca, NY: Cornell University Press, 1983).

[98] Hedley Bull, *The Anarchial Society: A Study of Order in World Politics* (New York: Columbia University Press, 1977), 54.

[99] Robert Keohane and Joseph Nye, *Power and Interdependence* (London, UK: Longman, 2012), 19.

[100] David Dessler, "What's at Stake in the Agent-Structure Debate?" *International Organization* 43, no. 3 (1989): 448.

[101] Anthony Giddens, *The Constitution of Society* (Berkeley: University of California Press, 1984), 21.

[102] Nicholas Onuf, "Constructivism: A User's Manual," In *International Relations in a Constructed World,* eds. Vendulka Kubalkova, Nicholas Onuf, and Paul Kowert (New York, NY: Routledge, 1998), 68.

[103] "Treaty Between the United States of America and the Union of Soviet Socialist Republics on the Limitation of Anti-Ballistic Missile Systems (ABM Treaty), 1972.

[104] Ibid.

> The Parties agree that in the event ABM systems based on other physical principles and including components capable of substituting for ABM interceptor missiles, ABM launchers, or ABM radars *are created in the future*, specific limitations on such systems and their components would be *subject to discussion* in accordance with Article XIII and agreement in accordance with Article XIV of the Treaty.[105]

Limiting the scope of the Treaty to technologies built at the time of its ratification and thereby clarifying *what is* subject to the Treaty and *what is not,* Agreed Statement D served as an important constitutive rule defining the parameters of the Treaty.

Like rules*, norms* contain notions of legitimacy, permissibility, and appropriateness. Although like rules norms distinguish legitimate behaviour from illegitimate behaviour, *norms do so implicitly*. Regardless of its source, concepts of legitimacy affect the *kind* of goals states pursue and *how* they pursue them. Moreover, norms often constitute the general principles and beliefs underlying formal, rule-based agreements. For example, the Non-Proliferation Norm (NPN), which undergirds the Treaty on the Non-Proliferation of Nuclear Weapons (NPT), delegitimizes the weaponization of nuclear technology.[106] NPT differentiates between two types of states, which internalize different aspects of the NPN. Under Article I, *nuclear-weapon states* pledge to refrain from actions that could enable non-nuclear-weapon states to obtain nuclear weapons. Under Article II and III, respectively, *non-nuclear-weapon states* pledge to not seek out nuclear technologies and to comply with verification checks. The NPN is a vehicle of beliefs and expectations about how states *should* employ nuclear technology; it underlies the specific obligations and formal commitments made in the NPT, which differentiates between the two classes. For nuclear-weapon states, it delegitimizes *the transfer, facilitation, and exchange* of nuclear weapon technologies, while, for non-nuclear-weapon states, it delegitimizes their *acquisition* and *production*.[107] The NPT translates the definitions of legitimacy, permissibility, and appropriateness provided by the NPN into a set of *rules* that define and clarify the roles nuclear and non-nuclear-weapon states play in the nuclear non-proliferation agenda.

Although centered on the NPT Treaty, the nuclear non-proliferation regime also consists of the International Atomic Energy Agency (IAEA) and Nuclear-Weapon-Free Zones.[108] Serving as a body of rules, IAEA verification programs delineate what states *can* and *cannot* do in terms of the development, employment, storage, and transportation of nuclear assets.[109] Moreover, NWFZs delineate how states within certain geographical areas can employ nuclear technology. For instance, NWFZs prohibit states from *weaponizing* nuclear assets, limiting the use of nuclear technology to peaceful purposes only. Each state within a NWFZ may operate nuclear energy facilities so long as they are used for civilian purposes and in compliance with IAEA safeguards.[110]

Recall that for constructivists, preferences are a function of identity. *Contra* neorealism, constructivists argue that preferences evolve *independently* of changes in the global distribution of (material) power. Under the constructivist view, inter-subjective understandings mediated by norms affect

[105] Joshua O'Donnell, "The Anti-Ballistic Missile Treaty Debate: Time for Some Clarification of the President's Authority to Terminate a Treaty," *Vanderbilt Law Review* 35, no. 5 (2002): 1609.
[106] Mario Caranza, "The Stability of the Nuclear Nonproliferation Norm: A Critique of Non-Contestation Theory," *Nonproliferation Review* 26, no. 1 (2019): 20.
[107] Caranza, 20.
[108] Joseph Siracusa and Aiden Warren, "The Nuclear Non-Proliferation Regime: An Historical Perspective," *Diplomacy and Statecraft* 29, no. 1 (2018): 5.
[109] Karl Pieragostini, "Arms Control Verification: Cooperating to Reduce Uncertainty." *The Journal of Conflict Resolution* 30, no. 3 (1986): 424.
[110] Joseph Siracusa and Aiden Warren, "The Nuclear Non-Proliferation Regime: An Historical Perspective," 14.

identity and, *by extension*, preferences.[111] For instance, constructivist Bruce Cronin defines systems in terms of collective identities. He argues that not only do concerts differ from a balance-of-power system in its distribution of power, but they also differ in the *type* of states they consist of.[112] He defines a concert as a system comprised of states with *common, transnational identity*. Under the neorealist view, however, states in a bipolar balance-of-power system would not differ in *identity* from states in other types of systems.

Recall from Chapter II how the kinds of goals actors pursue depends largely on *who* they are. By adopting a cooperative identity, states view themselves, and others, as belonging to an in-group: an *international society*. An international society differs from the international system in that a society involves "the institutionalization of shared interest and identity [amongst states]".[113] States can develop an international society or adopt a shared identity by associating *their interest* with the *collective interest*—by valuing *joint utility*. States that view themselves as belonging to a society are less worried about relative gains *within that group* than they are about the balance of power between *their in-group* and *other out-groups*. Thus, states can come to accept future uncertainty about the intentions of other states within their respective in-groups.

In anarchy, states are, by default, *competitors* driven by the norm of egoism: the view that one should maximize their own self-interest. However, by moving "from a definition of self as unique and distinct to one that perceives the self as [part of] a conceptual social group", states overcome egoism.[114] Competitors can change their identity by replacing egoism with an *alternative* norm—such as *conditional reciprocity*. In doing so, they adopt a *new* identity based on maximizing *mutual*, as opposed to, self, interests. In effect, they become *cooperators.*

Neoliberal institutionalists argue that states can overcome collective problems by creating institutions. Institutions are stable, recognizable patterns of rules and related shared practices rooted in shared values specific to a certain issue area.[115] While regimes and institutions typically center on a specific issue area, an *organization* can span various issue areas and encompass multiple regimes. For instance, the IAEA promotes the nuclear non-proliferation regime, which, in turn, is part of a broader collective security regime.[116] Under the neoliberalist view, states can achieve mutual gains *absent* a transformation of the system itself. Alternatively, constructivists argue that states can *transform* the system in ways that facilitate cooperation. Using games of conditional reciprocity, this thesis models this transformation in the context of the transition between a collective action problem to a coordination game. The thesis explores how, by engaging in tit-for-tat, competitors internalize the norm of conditional reciprocity and come to associate *their interest* with the *collective interest*. In doing so, states become cooperators, which, recall from Chapter II, are defined by their preference for maximizing joint utility.

Transforming Collective Action Problems into Coordination Games

Adopting a constructivist approach, while also using repeated games of conditional reciprocity, or *tit-for-tat*, this section explores how states can adopt a new identity based on collective interest. Through this process, states can overcome future uncertainty and egoism. Instead of using other rationalist devices

[111] Shiping Tang, "Fear in International Politics: Two Positions," *International Studies Review* 10 (2008): 464.
[112] Bruce Cronin, Community Under Anarchy: Transnational Identity and the Evolution of Cooperation (New York: Columbia University Press, 1999), 4.
[113] Cronin, 7.
[114] Cronin, 31.
[115] Nicholas Onuf, "Constructivism: A User's Manual,"in *International Relations in a Constructed World,* eds. Vendulka Kubalkova, Nicholas Onuf, and Paul Kowert (New York, NY: Routledge, 1998), 61.
[116] Anu Bradford, "Regime Theory," in *Max Plank Encyclopedia of Public International Law* (Social Science Research Network: 2007), 1-10.

such as Aumann's games of infinite length[117] and games of social survival, [118]the thesis uses tit-for-tat due to its *rule-based* nature.

Overcoming future uncertainty entails altering the payoff structure of collective action problems in a way that eliminates the dominant strategy to defect. Resolving the problem of future uncertainty in security dilemmas entails transforming the dilemma into a coordination game. This transition involves substituting the *dominant* strategy to defect with a *contingent* strategy based on conditional reciprocity. Eliminating the dominant strategy to defect involves increasing the costs of defection and decreasing the costs of cooperation. Using repeated games of conditional reciprocity, or tit-for-tat, this thesis models this transition. Changes in *time horizons* and *identities* mediate this process.

States can alter the payoff structure of collective action problems—and eliminate the dominant strategy to defect—by lengthening time horizons.[119] States judge the costliness of cooperation by comparing its near-term utility with its long-term utility.[120] Time horizons refer to the value players ascribe to long-term cooperation vis-à-vis early defection.[121] Two states with identical substantive preferences but different time horizons evaluate costs differently. *Myopic* states, who have *short* time horizons, behave opportunistically and often at the expense of long-term cooperation. Myopic states are more likely than *non-myopic* states, who have longer time horizons, to defect early in collective action problems. Players guided by short time horizons engage in *hyperbolic discounting*, valuing short-term gain enabled by early defection *more* than the long-term gain associated cooperation.[122] Hence, although they would benefit more in the *future* from cooperating, myopic states defect.

Recall that in security dilemmas, future uncertainty causes states to have the dominant strategy to defect. Driven by loss aversion, states often forgo potential long-term gain to prevent short-term loss. Longer time horizons *reverse* this loss aversion by making defection more costly. The dominant strategy to defect weakens as defection becomes costlier and longer time horizons make early defection *more costly relative* to long-term cooperation. As time horizons lengthen, the incentive to defect decreases, reducing the risk of exploitation attendant with cooperating.

In one-shot Prisoner's Dilemma (PD), players will always employ their respective dominant strategies, resulting in mutual defection. Because each player expects others to defect everyone defects. However, Axelrod discovered that, in *repeated* PD, where players expect to interact over multiple, or an indefinite, number of rounds, they will refrain from defecting on the first round. [123] In one-shot PD, there are no future payoffs to consider. Thus, the *discount rate*, which corresponds to how much states value *future payoffs relative to current payoffs*, is zero, resulting in short time horizons. [124] Thus, players defect. A state with low discount values value earlier payoffs more than later ones and are, thus, more liable than states with higher discount values to defect early.

[117] Robert Aumann, "Acceptable Points in General Cooperative *n*-Person Games," in *Contributions to the Theory of Games IV, Annals of Mathematics Study*, eds. Albert Tucker and Robert Luce (Princeton, NJ: Princeton University Press, 1959).

[118] Martin Shubik, "Game Theory, Behaviour, and the Paradox of the Prisoner's Dilemma," *The Journal of Conflict Resolution* 14, no. 2 (1970): 190.

[119] Kyle Haynes, "A Question of Costliness: Time Horizons and Interstate Signalling," *The Journal of Conflict Resolution* 63, no. 8 (2019).

[120] Robert Axelrod, *The Evolution of Cooperation* (New York, NY: Basic Books, 1984).

[121] Martin Shubik, "Game Theory, Behaviour, and the Paradox of the Prisoner's Dilemma," *The Journal of Conflict Resolution* 14, no. 2 (1970): 190.

[122] Maureen Cropper and David Laibson, "The Implications of Hyperbolic Discounting for Project Evaluation" (The World Bank Working Paper, Development Research Group, Washington, D.C., 1998), 1.

[123] Ibid.

[124] Kenneth Oye, "Explaining Cooperation under Anarchy: Hypotheses and Strategies," *World Politics* 38, no. 1 (1985): 1-24.

If the discount parameter, δ, is sufficiently large, however, then there is no one best strategy, or dominant strategy, the player can employ independently of the other player's strategy. Hence, the higher the discount value, the *weaker* the dominant strategy to defect and the more likely it is that states will risk exploitation *now* in hopes of achieving greater gains in the *future*. For any given time, the *more states value future payoffs relative to current payoffs* the *less likely they are to defect*.[125] By increasing their time horizons, and the discount value, states overcome future uncertainty and the attendant pressure to defect early.

Samuelson's Discounted Utility Model illustrates how players with high discount values are more likely to cooperate than those with low discount values. This model represents preferences as an intertemporal utility function: u_τ.[126] The utility at time *t* of some set of goods (e.g. security), accruing at times $t+1$, $t+2$, $t+3$,… T is given by:

$$V_t \equiv \sum_{t=i}^{\infty} D(t) \times u_{t+i}$$

Where: $D(t)$ or $D(t) = \delta^t$

Figure 10: Samuelson's Discounted Utility Model (Condensed)

The present value at time t of a stream of payoffs u_t, u_{t+1}, u_{t+2}, u_{t+3}…u_{t+i}, …, with a discount factor $0 < \delta < 1$ is:

$$V_t \equiv \sum_{t=0}^{\infty} \delta^i \cdot u_{t+i} = u_t + \delta \cdot u_{t+1} + \delta^2 \cdot u_{t+2} + \delta^3 \cdot u_{t+3} + \cdots \delta^i \cdot u_{t+i} + \cdots$$

Figure 11: Samuelson's Discounted Utility Model

Assume that:

- Cost associated with initiating cooperation during first round: u_t
- Instantaneous benefit iniator acquires from joint cooperation: u_{t+1}
- Perceived utility at a specific moment t in period T: u_t, u_{t+1}, u_{t+2}, u_{t+3}
- Discount factor (measures how an actor discounts utility in *later* periods relative to earlier periods): δ

For Examples I, II, and III:

- Assume two players: Player 1, P_1, and Player 2, P_2
- Let the instantaneous cost of cooperating for P_1, equal -1
- Let the instantaneous benefit, ω, P_1 gains should P_2 reciprocate at t equal 1.5
- Assume that the two players engage in tit-for-tat indefinitely $t + i$

Given $u_{t+\tau} = -1 + \delta^{t+i}(1.5)$

[125] Robert Axelrod, *The Evolution of Cooperation* (New York, NY: Basic Books, 1984).

[126] Angelina Lazaro, Ramon Barberan, and Encarnacion Rubio, "The Discounted Utility Model and Social Preferences: Some Alternative Formulations to Conventional Discounting," *Journal of Economic Psychology* 23, no. 3 (2002): 317-337.

- $If\ u_{t+i} > 0$, then P_1 is willing to initiate cooperation at t
- $If\ u_{t+i} < 0$, then P_1 is unwilling to initiate cooperation at t

Example I: High Discount Value

Let the discount factor δ equal 0.95

At t $u_t = -1 + 0.95^0(1.5) = 0.5$
At $t + 1$ $u_{t+1} = -1 + 0.95^1(1.5) = 0.425$
At $t + 2$ $u_{t+2} = -1 + 0.95^2(1.5) = 0.35$
At $t + 3$ $u_{t+3} = -1 + 0.95^3(1.5) = 0.286$

Example II: Moderate Discount Value

Let the discount factor δ equal 0.8

At t $u_{t+1} = -1 + 0.8^0(1.5) = 0.5$
At $t + 1$ $u_{t+1} = -1 + 0.8^1(1.5) = 0.2$
At $t + 2$ $u_{t+1} = -1 + 0.8^2(1.5) = -0.04$
At $t + 3$ $u_{t+1} = -1 + 0.8^3(1.5) = -0.232$

Example III: Low Discount Value

Let the discount factor δ equal 0.2

At t $u_{t+1} = -1 + 0.2^0(1.5) = 0.5$
At $t + 1$ $u_{t+1} = -1 + 0.2^1(1.5) = -0.7$
At $t + 2$ $u_{t+1} = -1 + 0.2^2(1.5) = -0.94$
At $t + 3$ $u_{t+1} = -1 + 0.2^3(1.5) = -0.988$

Corroborating Axelrod's theory, Example I, II and III, suggest that players with high discount values are more likely to initiate cooperation than players with low discount values.

Example IV: Low Discount Value and High Instantaneous Benefit

For Example IV:

- Maintain the instantaneous cost of cooperating for P_1 at -1
- *Double* the instantaneous benefit P_1 gains from jointly cooperating with P_2
- Let the discount factor δ equal 0.2

Thus:

At t $u_{t+1} = -1 + 0.2^0(3) = 2$
At $t + 1$ $u_{t+1} = -1 + 0.2^1(3) = -0.4$
At $t + 2$ $u_{t+1} = -1 + 0.2^2(3) = -0.88$
At $t + 3$ $u_{t+1} = -1 + 0.2^3(3) = -0.976$

Example IV emphasizes the effect discount values have on the likelihood of states initiating cooperation. Increasing the instantaneous benefit of joint cooperation at t does not offset the effects the low discount factor has on the player's willingness to risk exploitation.

The Samuelson's Discounted Utility Model can illustrate the dynamics of *tit-for-tat,* where the initiator incurs an initial cost upon cooperating first but sees long-term value in cooperation. Assume that players P_1, and P_2 are playing Prisoner's Dilemma *ad infinitum* with the one-shot payoff matrix.

	P_2		
		b_1	b_2
P_1	a_1	3, 3	0, 4
	a_2	4, 0	2, 2

Figure 12: Prisoner's Dilemma – One-shot Payoff Matrix – Tit-for-Tat

In a two-person tit-for-tat, a player, or, in this example, P_1, chooses to cooperate in the first period, and thereafter selects the same move the other player, P_2, choses in the preceding round. Tit-for-tat is a Nash Equilibrium strategy profile. For instance, consider P_1's incentives to deviate in period 0, assuming P_2 plays according to tit-for-tat, and P_1reverting to tit-for-tat thereafter.

If P_1cooperates in each period, the present value of the stream of payoffs is:

$$3 + \delta \cdot 3 + \delta^2 \cdot 3 + \cdots + \delta^i \cdot 3 + \cdots = 3 \cdot \left(1 + \delta + \delta^2 + \cdots + \delta^i + \cdots\right) = 3 \cdot \frac{1}{1-\delta}$$

If P_1 deviates in period 1, while P_2 plays according to tit-for-tat and chooses to cooperate, P_1 receives a payoff of 4 in period 1. Moreover, P_2will defect in period 2 (mirroring P_1's action in the previous period), resulting in P_1 receiving a payoff of 0 in period 2, 4, in period 3, and so on, and so forth, with payoffs alternating between 4 and 0. The present value of this stream of payoffs is:

$$4 + \delta \cdot 0 + \delta^2 \cdot 4 + \delta^3 \cdot 0 + \delta^4 \cdot 4 + \cdots = 4 \cdot (1 + \delta^2 + \delta^4 \ldots) = 4 \cdot \frac{1}{1-\delta^2}$$

P_1will cooperate if, and only if:

$$3 \cdot \frac{1}{1-\delta} \geq 4 \cdot \frac{1}{1-\delta^2} \Rightarrow 3 \cdot \frac{1+\delta}{1-\delta^2} \geq 4 \cdot \frac{1}{1-\delta^2} \Rightarrow 3 \cdot (1+\delta) \geq 4 \Rightarrow 3 \cdot \delta \ \geq 1 \Rightarrow \delta \geq \frac{1}{3}$$

Thus, if P_1's discount factor is high enough ($\delta > \frac{1}{3}$), then P_1will cooperate. The same argument applies to P_2.

The principle of Mutually Assured Destruction (MAD) illustrates one way of increasing discount values. MAD increases this value by making defection prohibitively costly, that is, by increasing the cost of defection. During the 1960s, the Soviets, perceiving nuclear parity as a necessary condition for effective deterrence, accelerated its Intercontinental Ballistic Missile production. In 1967, the Soviets enhanced their first-strike capabilities by constructing an Anti-Ballistic Missile (ABM) defense-system around Moscow. The Soviet arms build-up in missile defence systems weakened American retaliatory capability, tipping the strategic balance to the Soviet Union. By the 1970s, however, both the United

States and the Soviet Union had secured second-strike capabilities, thereby entering a gridlock predicated upon mutual deterrence and MAD.[127]

Under MAD, the two countries refrained from launching pre-emptive first strikes against each other lest the other state retaliates with their second-strike capabilities. Effective deterrence under MAD presupposed the existence of adequate retaliatory capabilities; it was also predicated upon the *belief* that countries will make use of such capabilities. Increasing the credibility of threats of retaliatory attacks, this belief rendered first-strikes highly risky, thereby increasing the discount value.[128] Unlike MAD, tit-for-tat increases the costs of defection *normatively;* it is a vehicle for the norm of reciprocity, which *delegitimize* defection.

Recall that for realists, interaction is a function *not of identity* but of *structure*. Alternatively, constructivists advance a: "Cognitive, inter-subjective conception of process [interaction] in which identities and interests are endogenous to interaction [as opposed to exogenous]".[129] Norms not only constrain or regulate the behavior of states, but they also create, define, and transform their identities and, by extension, interests. Inter-subjective understandings mediated by norms affect how states define themselves and others.[130]

Overcoming egoism involves moving away from "a definition of self as unique and distinct to one that perceives the self as [constituting] a conceptual social group.[131]. States can come to view themselves as constituting a social group where they willingly forgo short-term interests for the sake of securing common interests in the long-term. The norm of reciprocity makes *competitors* come to value the *joint gains* associated with long-term cooperation over *relative gain maximization*, thereby transforming them into *cooperators*. By embracing joint utility, competitors come to identify with a social group, thereby eschewing egoism. States can, thus, come to view each other as part of a unique *social group of cooperators*, or an international society, framing each other not as *adversaries* but as *collaborators* unified by a shared identity and interest.[132]

Players can sustain cooperation by adopting a strategy of conditional reciprocity and engaging in tit-for-tat. Since it is a rule-based strategy, tit-for-tat effectively replaces the dominant strategy to defect within collective action problems with a *contingent* strategy based on reciprocity. The rule of reciprocity enables states to coordinate their actions, thereby reducing future uncertainty and rendering cooperation stable. By eliminating the dominant strategy to defect, tit-for-tat transforms the collective action problem, which, recall from Chapter II, consists of a singular equilibrium at mutual defection, into a coordination game comprised of two equilibria, one at mutual cooperation, and one at mutual defection.[133] Moreover, by adopting the norm of reciprocity, states cultivate a cycle of cooperative behaviour where competitors increasingly value cooperation over defection. Competitors thereby *become* cooperators, overcoming egoism. States thereby adopt a system within which cooperation occurs organically: a rule-based international society based on reciprocity.

[127] Joshua O'Donnell, "The Anti-Ballistic Missile Treaty Debate: Time for Some Clarification of the President's Authority to Terminate a Treaty," *Vanderbilt Law Review* 35, no. 5 (2002): 1603.
[128] O'Donnell, 1603.
[129] Alexander Wendt. "Anarchy is What States Make of It: The Social Construction of Power Politics," *International Organization* 46, no. 2 (1992): 392.
[130] Shiping Tang, "Fear in International Politics: Two Positions," *International Studies Review* 10 (2008): 464.
[131] Tang, 31.
[132] Tang, 7.
[133] Dean Pruitt, "Twenty Years of Experimental Gaming: Critique, Synthesis, and Suggestions for the Future," *Annual Review of Psychology* 28, no. 1 (2003): 370.

Achieving Cooperation in Coordination Games

To overcome the security dilemma, states must understand *who others are* and *what they value*. In coordination games, cooperators arm *only because they do not know the type of player* others are. Cooperators arm if they believe that their opponent will arm; however, if cooperators knew that others were also cooperators, then they would refrain from arming since they would no longer fear exploitation by potential competitors. [134]

In coordination games, a key barrier to cooperation is insufficient trust.[135] The literature on trust theory as it relates to IR centers on two main traditions: the rationalist approach and the binding approach.[136] The rationalist approach emphasizes the effect interests have on trust-building. Proponents of the rationalist approach argue that trusting relationships depend on *interest-based* calculations contingent upon the pay-off structure underlying the interaction.[137] Under this view, State A trusts State B to the extent that A believes *that it is in B's interest* to respect A's interests.[138] Alternatively, according to the binding approach, A trusts B because it thinks that B values its relationship with A *for its own sake*.[139] Proponents of the binding approach, whose arguments complement constructivism, believe that states value trusting relationships independently of the underlying pay-off structure.[140]

This thesis articulates a hybrid approach. The payoff structure of the security dilemma is constructed in such a way that incentivizes the *risk-maximizing outcome*: mutual defection. Although cooperators may value cooperation for its own sake, they will defect if confronted with a *low* level of trust. Hence, as proponents of the rationalist approach suggests, the incentive structure affects behaviour *even* among players who value certain outcomes for their own sake. It is the lack of trust that incentivizes defection. Hence, for states to come to prefer the payoff-maximizing outcome over the risk-maximizing one in coordination games, they must trust each other.

The uncertainty states have regarding *who others are* and, by extension, *what they want*, impedes trust. Through costly signalling, states can discern the possible *types* of other states and increase the credibility of their commitments towards cooperation. Research on costly signalling spans various domains within IR. Avidit Acharya and Krisopher Ramsay,[141] Brian Blankenship,[142] and Brandon Haynes and Kyle Yoder,[143] apply costly signalling to reassurance; Anne Sartori,[144] Robert Trager,[145] and Allan

[134] Kristopher Ramsay, "Information, Uncertainty, and War," *Annual Review of Political Science* 20 (2017): 520.

[135] Vincent Keating and Jan Ruzicka, "Trusting Relationships in International Politics: No Need to Hedge," *Review of International Studies* 40, no. 4 (2014): 759.

[136]Jan Ruzicka and Nicholas Wheeler, "The Puzzle of Trusting Relationships in the Nuclear Non-Proliferation Treaty," *International Affairs* 86, no. 1 (2010): 70.

[137] Ruzicka and Wheeler, 71.

[138] Russell Hardin, *Trust and Trustworthiness* (New York, NY: Russell Sage Foundation, 2002), 3.

[139] Jan Ruzicka and Nicholas Wheeler, "The Puzzle of Trusting Relationships in the Nuclear Non-Proliferation Treaty," *International Affairs* 86, no. 1 (2010): 73.

[140] Ruzicka and Wheeler, 74.

[141] Avidit Acharya and Kristopher Ramsay, "The Calculus of the Security Dilemma," *Quarterly Journal of Political Science* 8, no. 2 (2013): 183-203.

[142] Brian Blankenship and Erik Lin-Greenberg, "Trivial Tripwires? Military Capabilities and Alliance Reassurance," *Security Studies* 31, no. 1 (2022): 92-117.

[143] Brandon Yoder and Kyle Haynes, "Signalling under the Security Dilemma: An Experimental Analysis," *Journal of Conflict Resolution* 65, no. 4 (2020).

[144] Anne Sartori, *Deterrence by Diplomacy* (Princeton, NJ: Princeton University Press, 2005).

[145] Robert Trager, "The Diplomacy of War and Peace," *Annual Review of Political Science* 19 (2016): 205-228.

Dafoe, Johnathan Renshon, and Paul Huth,[146] to deterrence. This thesis uses Andrew Kydd's costly signalling game model—a type of Bayesian game where states take turns making increasingly costly moves over multiple interactions—to model how states can develop trust in coordination games.[147]

In coordination games, cooperators can reveal their respective types through costly signalling, thereby mitigating the uncertainty that prevents trust from developing. Costly signalling shows how through interaction states obtain information with which they can update their beliefs regarding *who others are* and *what they value.* States deduce *who* other players are from how they act, associating specific actions with certain types. Through such games, states can develop trust by signalling their willingness to forgo the short-term gains associated with early defection.

Costly signalling games evoke Wendt's concept of *reciprocal typications* that reinforce concepts of identity.[148] Showing how over multiple interactions states develop, reinforce, and create identities, costly signalling consists of what Wendt defines as *social acts.* Wendt writes:

> The first social act creates [tentative] expectations on both sides about each other's future behaviour… From this initial signal by A, B responds to A's initial signal. Based on this knowledge, *ego* makes a new gesture, again signifying the basis on which it will respond to *alter*, and again alter responds, adding to the pool of knowledge each has about the other, and so on over time.[149]

In Kydd's costly signalling games, each actor knows their *own* type but is uncertain about the type of player the *other* is. This uncertainty complicates cooperation since neither state wants to risk the sucker payoff.

To illustrate, recall the typology of states explored in Chapter II, and consider a coordination game comprised of two cooperators, A and B. If A believes B is a competitor, then A expects B to defect as per its dominant strategy. A would thus defect to avoid its least preferred outcome: unilateral defection by B. Similarly, if A does *not know* what type of player B is, or is *highly uncertain* of B's identity, then A would still *prefer* the risk-dominant payoff and defect. However, if A thinks that B is a cooperator, and *vice versa*, then both players come to *prefer* the payoff-maximizing outcome and cooperate. Recall from Chapter II that coordination failures occur when there is at least one player that *believes* that the other is a competitor (**Figure 8**). Costly signalling can reverse this belief between cooperators. If A *believes* that B *is a cooperator*, then A would risk cooperation.

The Soviet Union's shift towards détente in the mid-1980s illustrates how by increasing the costliness of cooperative signals states can facilitate reciprocation. Vincent Keating and Jan Ruzicka argue that the relationship between the two superpowers was based on *mutual confidence*.[150] Confidence occurs when an actor expects another actor to comply with a social norm or reciprocate costly signals.

In 1985, the Soviet Union, under Mikhail Gorbachev, signalled a desire for rapprochement with the United States. Significantly underestimating the costs the Soviets risked, or incurred, in their initial attempts at reconciliation, the United States initially dismissed the Soviet attempts at rapprochement as insincere.[151] Recall that effective signalling is a function of how costly the signal is not so much to the

[146] Allan Dafoe, Jonathan Renshon, and Paul Huth, "Reputation and Status as Motives for War," *SSRN* (2014): 1-23.

[147] Andrew Kydd, "Trust, Reassurance, and Cooperation," *International Organization* 54, no. 2 (2000).

[148] Alexander Wendt. "Anarchy is What States Make of It: The Social Construction of Power Politics," *International Organization* 46, no. 2 (1992): 405.

[149] Wendt, 405.

[150] Vincent Keating and Jan Ruzicka, "Trusting Relationships in International Politics: No Need to Hedge," *Review of International Studies* 40, no. 4 (2014): 753-770.

[151] Todd Hall and Keren Yarhi-Milo, "The Personal Touch: Leaders' Impressions, Costly Signaling, and Assessments of Sincerity in International Affairs," *International Studies Quarterly* 56, no. 3 (2012): 567.

sender as to the *receiver.* Accordingly, in 1987, the Soviets agreed to the Intermediate-range Nuclear Forces (INF) Treaty, accepting more intrusive monitoring arrangements, including on-site inspections. The Soviet decision to sign the INF, and consent to a more comprehensive verification regime, marked the first Soviet action that the United States deemed *costly enough* to warrant a reversal in American opinion of the Soviet Union as an expansionist power.[152] The United States regarded the Soviet decision to sign the INF as an act of *bona fide* costly reassurance. Moreover, by the late 1980s, the Soviet Union underwent comprehensive domestic reform. The United States regarded such costly actions as signalling of a Soviet willingness to improve Soviet-U.S. relations.

Strategic interaction in Kydd's costly signalling games spans two rounds. In each round, players choose, simultaneously, whether to cooperate or defect.[153] Each player enters the first round with identical levels of *prior trust* (t). Upon observing the other's move at the end of the first round, each player updates their belief about the other, forming a *posterior level of trust* (t'). The players then conduct their respective second-round moves. Cooperators are more likely to cooperate in the second round if they began the game with a relatively *high level of prior trust* than *low-to-moderate* prior trust.[154] However, in security dilemmas, cooperators are confronted with a *low* initial level of trust. Although under intermediary stakes, cooperators with low levels of trust may risk exploitation by cooperating, security dilemmas are characterized by *high* stakes. In *low trust, high stakes* situations, such as the security dilemma, the risk of exploitation is high, and hence cooperators defect, resulting in coordination failures.[155] By lowering the stakes just enough—but not too low to the point where cooperative signals lose credibility—trust can develop.[156]

Cooperation entails vulnerability. Initiating tit-for-tat involves relatively high upfront costs and risks. For a tit-for-tat to occur, the initiator must risk exploitation by their opponent. Costly signalling games can help states achieve a baseline of risk that is low enough to facilitate the initiation of tit-for-tat but *high* enough to promote trust.[157] In the context of the security dilemma, the extent to which a state prefers to arm depends on the perceived cost of arming. If, by arming, a state incurs little cost, then that state's dominant strategy is to arm regardless of what their opponents do. So-called high-cost cooperators prefer to arm only if they believe with *near certainty* that others will arm.[158] Low-cost cooperators will arm if they believe that their opponent will arm even with the *slightest* probability—as in the case of security dilemmas.

Central to costly signalling games is communication. Instead of *updating* their beliefs in light of new information, states tend to *assimilate* new information into their prior—and potentially erroneous—beliefs.[159] Moreover, when states *do* alter their beliefs, they often do so through a process of 'asymmetric updating' whereby they evaluate new information against prior beliefs. Nyhan and Reifler write:

> Humans are goal-directed information processors who tend to evaluate information with a directional bias toward reinforcing their pre-existing views… [people] tend to evaluate

[152] Hall and Yarhi-Milo, 567.

[153] Andrew Kydd, "Trust, Reassurance, and Cooperation," *International Organization* 54, no. 2 (2000): 328.

[154] Brandon Yoder and Kyle Haynes, "Signalling under the Security Dilemma: An Experimental Analysis," *Journal of Conflict Resolution* 65, no. 4 (2020): 690.

[155] Yoder and Haynes, 691.

[156] Yoder and Haynes, 678

[157] Charles Lipson, "International Cooperation in Economic and Security Affairs," *World Politics* 37, no. 1 (1984): 17.

[158] Kristopher Ramsay, "Information, Uncertainty, and War," *Annual Review of Political Science* 20 (2017): 520

[159] Joshua Kertzer, Brian Rahbun, and Nina Rathbun. "The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations," *International Organization* 74, no. 1 (2020): 96.

> information with a directional bias toward reinforcing their pre-existing views… [and they tend to disparage information] that contradict their views.[160]

Effective costly signalling presupposes a mechanism by which states can overcome asymmetric updating, such as the Standing Consultative Commission (SCC), a joint US-USSR institution created by Article XIII of the 1974 ABM Treaty. Through the SCC, the superpowers discussed compliance-related issues in ways that often challenged the pre-existing beliefs they had of each other. Although it did not formalize a costly signalling mechanism, the SCC provided a forum for negotiation, information-sharing, and dispute resolution, through which the superpowers could clarify misperceptions and reinforce their commitments towards compliance.[161] Speaking more broadly, states can signal their commitment by assenting to international regimes and institutions; undertaking costly actions such as reducing military expenditures; agreeing to arms control protocols, and investing in defensive weapon-systems.

The first part of this section showed how states can adopt the norm of conditional reciprocity to overcome future uncertainty and egoism. A change in collective identity from *competitors* to *cooperators* transforms the payoff structure of the game in a way that eliminates the dominant strategy to defect, transforming the collective action problem into a coordination game. The second part of this section examined how states can remedy low levels of initial trust through costly signalling games.

The thesis articulated these processes using rationalist models centered on the identity-preference relationship. It uses repeated games of conditional reciprocity to explore how changes in identity coincide with preference change, and costly signalling games to examine how the beliefs states have of *who others are* affect their beliefs on what others *want.* To that end, the chapter proposed a potential *via media* between Rationalism and Constructivism.

[160] Brendan Nyhan and Jason Reifler, "When Corrections Fail: The Persistence of Political Misperceptions," *Political Behaviour* 32, no. 2 (2010): 307.

[161] Mathew Bunn, *Foundation for the Future: The ABM Treaty and National Security* (Washington, D.C.: Arms Control Association, 1990), 22.

CONCLUSION

Rationalist theories, such as Neorealism and Neoliberalism, argue that interaction within the international system is a function of *material* structure. They generalize state behaviour by analyzing how the distribution of power within the international system predisposes states to act in certain, predictable ways. Lacking in rationalist approaches is attention to how preferences and identities vary across states—and how they shape interaction. Rationalist theories discount the effect *inter-subjective* factors have on behaviour and, by extension, the pattern of interaction that can come to define the international system. In doing so, they centre their solutions to the security dilemma on changes in *material* structure or the creation of institutions—rather than changes *to the system itself.* For rationalist theories, it is material structure, or the distribution of power within the international system that determines the pattern of interaction defining it. Moreover, they assume that interaction cannot change the structure of the system.

However, like Constructivism, but unlike rationalist approaches, this thesis assumes that self-help is just *one* of many possible patterns of interaction that can come to define the international system. Each pattern of interaction is centred on a unique normative culture. Within a self-help system, that culture is based on egoism. Accordingly, this thesis proposes a solution to the security dilemma that involves changing the *culture* of the international system. Since security dilemmas can only occur within a self-help system, it follows that states can overcome security dilemmas if they can transform the international system from a self-help system to an international society.

Rationalist theories argue that within the international system states can only achieve *ad hoc* cooperation. Realists associate cooperation with alliance-building while neoliberalists focus on how institutions can facilitate coordination. For both theories, the only alternative to the self-help system—excluding a war-against-all—would be a system of World Government, which would entail a move from anarchy to hierarchy. This transition would involve not a *transformation* of the international anarchic system but its *elimination.* The World Government thesis implies that eliminating security dilemmas entails eliminating anarchy. However, this thesis argued that states can overcome security dilemmas by *transforming* anarchy rather than eliminating it. States can transform the self-help system into a rule-based international society based on mutual reciprocity where cooperation occurs organically, thereby nullifying the need for a central authority.

In Chapter I, the thesis reduced the security dilemma to loss aversion driven by three principal factors: future uncertainty, egoism, and mistrust. Recall that rationalist theories ignore how values and preferences vary across states*,* consequently downplaying the causal effect inter-subjective factors have on interaction. In Chapter II, the thesis explored how the beliefs states have about what other states *want* and *what they know* feed into subsidiary beliefs about *who* those states are and how they will *act.* It then examined how different state types calculate utility, establishing that what players *want* is linked inextricably to *who* they are. In doing so, the thesis formalized the identity-preference relationship. that underlies the two-step solution to the security dilemma proposed in Chapter III.

Chapter III explored how resolving the security dilemma entails transforming state preferences through the adoption of cooperative norms and identities. The thesis modelled preference change in the context of transforming collective action problems into coordination games. It then examined how in coordination games states can overcome risk aversion. Using games of conditional reciprocity, or tit-for-tat, the thesis showed how states can overcome future uncertainty and egoism through identity change. A change in identity from competitors to cooperators transforms the payoff structure of the security dilemma in a way that eliminates the dominant strategy to defect, turning the collective action problem into a coordination game. The thesis then explored, using costly signalling games, how cooperators can reveal their respective types, facilitating trust.

Drawing upon Constructivism, the thesis *defined* preference change as a function of norms, identities, and preferences, and using Game Theory, it *modeled* this relationship in the context of the security dilemma. This thesis articulated thusly how constructivists can use Game Theory to explain the identity-preference relationship systematically without having to forgo their inter-subjective approach, thereby presenting a *via media* between Rationalism and Constructivism. In doing so, the thesis addressed Constructivism's inability to address future uncertainty and Rationalism's ahistorical approach vis-à-vis the security dilemma.

The purpose of the hybrid approach proposed in this thesis is not so much to rival Rationalism or Constructivism as strengthen their explanatory, descriptive, and predictive potential. This approach could provide a departure point for a generalizable model of preference change with applications not limited to the security dilemma. It may also facilitate future research into the transformative effect the identity-preference relationship and has on interaction. For instance, hypergame analysis could facilitate the analysis of how other subjective factors such as risk propensity, ideology, and language affect perception and, by extension, decision-making. Moreover, one important aspect of norm theory that this thesis did not address is *norm compliance.* Compliance with regime norms depends partly on the extent to which those norms align with state interests or preferences. To correct misalignment, states may adopt new interests—or reframe pre-existing ones. As shown in this thesis, states could change their preferences by adopting new identities. The modelling approach presented in this thesis could thereby guide future research into how alignment could occur—research that could have applications to policymaking in fields such as diplomacy and international security.

In addressing the ontological debate between Rationalism and Constructivism over the nature of preference change by states, the thesis broadened the explanatory, predictive, and methodological utility of formal theories to IR. The thesis proposed a hybrid approach towards resolving the security dilemma, and to show that states can, indeed, 'make' the international system.

BIBLIOGRAPHY

Acharya, Avidit and Kristopher Ramsay. “The Calculus of the Security Dilemma.” *Quarterly Journal of Political Science* 8, no. 2 (2013): 183-203.

Aron, Raymond. *Peace and War*. Garden City, NY: Doubleday. 1966.

Aumann, Robert. “Acceptable Points in General Cooperative n-Person Games.” In *Contributions to the Theory of Games IV, Annals of Mathematics Study,* edited by Albert Tucker and Robert Luce. Princeton: Princeton University Press, 1959.

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

Barnett, Michael and Martha Finnemore. *Rules for the World*. Ithaca, NY: Cornell University Press, 2004.

Bennett, Peter. “The Arms Race as a Hypergame.” *Futures* 14, no. 4 (1982): 293-306.

Bennett, Peter. “Toward a Theory of Hypergames.” *OMEGA* 5 (1977): 749–751.

Bennett, Peter. “Using Hypergames to Model Difficult Social Issues: An Approach to the Case of Soccer Hooliganism.” *Journal of the Operational Research Society* 31, no. 7 (1980): 621-635.

Blankenship, Brian and Erik Lin-Greenberg. “Trivial Tripwires? Military Capabilities and Alliance Reassurance.” *Security Studies* 31, no. 1 (2022): 92-117.

Boulding, Kenneth. *Conflict and Defense*. New York, NY: Harper and Brothers. 1962.

Braumoeller, Bear. “Nested Politics: A New Systematic Theory of IR.” *Waterhead Center for International Affairs.* 2004.

Bradford, Anu. “Regime Theory.” In *Max Plank Encyclopedia of Public International Law,* 1-10. Social Science Research Network: 2007.

Bueno de Mesquita, Bruce. *The War Trap*. New Haven, CT: Yale University Press, 1981.

Bull, Hedley. *The Anarchial Society: A Study of Order in World Politics*. New York: Columbia University Press. 1977.

Bunn, Mathew. *Foundation for the Future: The ABM Treaty and National Security*. Washington, D.C.: Arms Control Association. 1990.

Caranza, Mario. “The Stability of the Nuclear Non-proliferation Norm: A Critique of Non-Contestation Theory.” *Non-proliferation Review* 26, no. 1 (2019): 7-22.

Carr, E.H. *Twenty Years Crisis.* London, UK: Palgrave Macmillan. 2001.

Cerny, Philip. *The Changing Architecture of Politics, Structure, Agency, and the Future of the State*. London: Sage Publications. 1990.

Copeland, Dale. “The Constructivist Challenge to Structural Realism: A Review Essay.” *Social Theory of International Politics* 25, no. 2 (2000): 187-212.

Cronin, Bruce. *Community Under Anarchy: Transnational Identity and the Evolution of Cooperation.* New York: Columbia University Press, 1999.

Cropper, Maureen and David Laibson. "The Implications of Hyperbolic Discounting for Project Evaluation." The World Bank Working Paper, Development Research Group, Washington, D.C., 1998.

Dafoe, Allan and Jonathan Renshon and Paul Huth. "Reputation and Status as Motives for War." *SSRN* (2014): 1-23.

Dessler, David. "What's at Stake in the Agent-Structure Debate?" International Organization 43, no. 3 (1989): 441-473.

Feinberg, Yossi. "Games with Awareness." *Stanford Graduate School of Business*, no. 2122. (2012): 1-52.

Finnemore, Martha and Kathryn Sikkink. "International Norm Dynamics and Political Change." *International Organization* 52, no. 4 (1998): 887-917.

Giddens, Anthony. *The Constitution of Society*. Berkeley: University of California Press. 1984.

Glaser, Charles. "Political Consequences of Military Strategy." *World Politics* 44, no. 4 (1992): 497-538.

Glaser, Charles. "Realists as Optimists." *International Security* 19, no. 3 (1994): 50-90.

Glaser, Charles. "The Security Dilemma Revisited." World Politics 50, no. 1 (1997): 171-201.

Grieco, Joseph. "Anarchy and the Limits of Cooperation: A Realist Critique of the Newest Liberal Institutionalism." *International Organization* 42, no. 3 (1988): 485-507.

Grieco, Joseph. "Understanding the Problem of International Cooperation: The Limits of Neoliberal Institutionalism and the Future of Realist Theory." In *Neorealism and Neoliberalism*, edited by David Baldwin, 301-331. New York, NY: Columbia University Press. 1993.

Hall, Todd and Keren Yarhi-Milo. "The Personal Touch: Leaders' Impressions, Costly Signaling, and Assessments of Sincerity in International Affairs." *International Studies Quarterly* 56, no. 3 (2012): 560-573.

Hardin, Russell. *Trust and Trustworthiness*. New York, NY: Russell Sage Foundation. 2002.

Harsanyi, John. "Game Theory and the Analysis of International Conflict." *The Australian Journal of Politics and History* 11 (1965): 292-304.

Haynes, Kyle. "A Question of Costliness: Time Horizons and Interstate Signalling." *The Journal of Conflict Resolution* 63, no. 8 (2019): 1939-1964.

Haynes, Kyle. "Trust, Cooperation, and the Trade-offs of Reciprocity." *Conflict Management and Peace Science* 41, no. 1 (2023): 26-46.

Hellenmuller, Sara and Jamie Pring, and Oliver Richmond. "How Norms Matter in Mediation: An Introduction." *Swiss Political Science Review* 26, no. 4 (2020): 345-363.

Herz, John. *Political Realism and Political Idealism*. Chicago: University of Chicago Press. 1951.

Hipel, Keith and Muhong Wang, and Niall Fraser, "Hypergame Analysis of the Falkland/Malvinas Conflict," *International Studies Quarterly* 32, no. 3 (1988): 335-358.

Hoffman, Stanely. *Contemporary Theory in International Relations*. Englewood Cliffs, NJ: Prentice-Hall. 1960.

Inohara, Takahashi, and Nakano, "Integration of Games and Hypergames Generated from a Class of Games," *Journal of the Operational Research Society* 48 (1997): 423-432.

Jervis, Robert. "Cooperation Under the Security Dilemma." *World Politics* 30, no. 2 (1978): 167-214.

Jervis, Robert. *Perception and Misperception*. Princeton, NJ: Princeton University Press. 1976.

Jervis, Robert. "Realism, Neoliberalism, and Cooperation: Understanding the Debate." International Security 24, no. 1 (1999): 42-63.

Keating, Vincent and Jan Ruzicka. "Trusting Relationships in International Politics: No Need to Hedge." *Review of International Studies* 40, no. 4 (2014): 753-770.

Kelley, Harold and Anthony Stahelski. "Social Interaction Basis of Cooperators' and Competitors' Beliefs About Others." *Journal of Personality and Social Psychology* 16, no. 1 (1970): 66-91.

Kelley, John and Harold Thibault, *The Social Psychology of Groups.* New Jersey: Transaction Publishers. 1959.

Keohane, Robert and Joseph Nye. *Power and Interdependence*. London, UK: Longman. 2012.

Keohane, Robert. "Theory of World Politics: Structural Realism and Beyond." In *Neorealism and its Critics*, edited by Robert Keohane, 158-204. New York: Columbia University Press, 1986.

Kertzer, Joshua. "The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations," *International Organization* 74, no. 1 (2020): 95-118.

Krasner, Steven. *International Regimes*. Ithaca, NY: Cornell University Press, 1983.

Kratochwil, Friedrick. "Constructivism as an Approach to Interdisciplinary Study." In *Constructing International Relations: The Next Generation*, 13-35. Armonk, New York: M.E. Sharpe. 2001.

Kratochwil, Friedrich. *Rules, Norms, and Decisions*. New York, NY: Cambridge University Press. 1989.

Kuhn, Harold. "Game Theory and Models of Negotiation." *The Journal of Conflict Resolution* 6, no. 1 (1962): 1-4.

Kydd, Andrew. "Trust, Reassurance, and Cooperation." *International Organization* 54, no. 2 (2000): 325-357.

Lazaro, Angelina, Ramon Barberan, and Encarnacion Rubio. "The Discounted Utility Model and Social Preferences: Some Alternative Formulations to Conventional Discounting." *Journal of Economic Psychology* 23, no. 3 (2002): 317-337.

Lipson, Charles. “International Cooperation in Economic and Security Affairs.” *World Politics* 37, no. 1 (1984): 1-23.

Lipson, Charles. *Reliable Partners: How Democracies Have Made a Separate Peace.* Princeton University Press. 2013.

Luce, Duncan and Howard Raiffa, *Games and Decisions,* John Wiley & Sons, Inc. 1957.

McClintock, Charles. “Motivational Bases of Choice in Three-Choice Decomposed Games.” *Journal of Experimental Social Psychology* 9, no. 6 (1973): 572-590.

McClintock, Charles and Wim Liebrand. “Role of Interdependence Structure, Individual Value Orientation, and Another’s Strategy in Social Decision Making: A Transformational Analysis.” *Journal of Personality and Social Psychology* 55, no. 3 (1988): 396-409.

Mearsheimer, John. *The Tragedy of Great Power Politics*. New York, NY: WW Norton, 2014.

Meiches, Benjamin and Raymond Hopkins. "Regime Theory." *Oxford Research Encyclopedia of International Studies.* 2018.

Morgenthau, Hans. *Politics Among Nations*. New York: Alfred Kopf. 1956.

Nash, John. “Non-Cooperative Games.” *The Annals of Mathematics* 54, no. 2 (1951): 286-295.

Nyhan, Brendan and Jason Reifler. “When Corrections Fail: The Persistence of Political Misperceptions.” *Political Behaviour* 32, no.2 (2010): 303-330.

O’Donnell, Joshua. “The Anti-Ballistic Missile Treaty Debate: Time for Some Clarification of the President’s Authority to Terminate a Treaty.” *Vanderbilt Law Review* 35, no. 5 (2002): 1601-1636.

Osborne, Martin. *An Introduction to Game Theory.* New Dehli, India: Oxford University Press. 2004.

Onuf, Nicholas. “Constructivism: A User’s Manual.” In *International Relations in a Constructed World*, edited by Vendulka Kubalkova, Nicholas Onuf, and Paul Kowert, 58-78. New York, NY: Routledge. 1998.

Oye, Kenneth. “Explaining Cooperation under Anarchy: Hypotheses and Strategies.” *World Politics* 38, no. 1 (1985): 1-24.

Parkhe, Arvind. “Strategic Alliance Structuring: A Game Theoretic and Transaction Cost Examination of Interfirm Cooperation.” *The Academy of Management Journal* 36, no. 4 (1993): 794-829.

Pieragostini, Karl. “Arms Control Verification: Cooperating to Reduce Uncertainty.” *The Journal of Conflict Resolution* 30, no. 3 (1986): 420-444.

Pitchford, Rohan and Mark Wright. “On the Contribution of Game Theory to the Study of Sovereign Debt and Default.” *Oxford Review of Economic Policy* 29, no. 4 (2013): 649–667.

Powell, Robert. “Neorealism and its Critics.” *International Organization* 48, no. 2 (1994): 313-344.

Pruitt, Dean. “Twenty Years of Experimental Gaming: Critique, Synthesis, and Suggestions for the Future.” *Annual Review of Psychology* 28, no. 1 (2003): 363-392.

Quek, Kai. “Four Costly Signaling Mechanisms.” *American Political Science Review* 115, no. 2 (2021): 537-549.

Ramsay, Kristopher. “Information, Uncertainty, and War.” *Annual Review of Political Science* 20 (2017): 505-527.

Rapoport, Anatol. *Fights, Games, and Debates*. Ann Arbour, MI: University of Michigan Press. 1974.

Risse, Thomas. “Let’s Argue: Communicative Action in World Politics.” *International Organization* 54, no. 1 (2000): 1-39.

Roe, Paul. “Actors’ Responsibility in Tight, Regular, or Loose Security Dilemmas.” *Security Dialogue* 32, no. 1 (2001): 103-116.

Roloff, Michael. *Interpersonal Communication: The Social Exchange Approach*. California: Sage Publications. 1981.

Ruggie, John. “What Makes the World Hang Together? Neo-Utilitarianism and the Social Constructivist Challenge.” *International Organization* 52 (1998): 855-885.

Ruzicka, Jan and Nicholas Wheeler, “The Puzzle of Trusting Relationships in the Nuclear Non-Proliferation Treaty.” *International Affairs* 86, no. 1 (2010): 69-85.

Sartori, Anne. *Deterrence by Diplomacy*. Princeton, NJ: Princeton University Press. 2005.

Schelling, Thomas. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press. 1980.

Shahrabi Farahami, M. and Majid Sheikmohammady, “A Review on Symmetric Games: Theory, Comparison, and Applications.” *International Journal of Applied Operational Research* 4, no. 3 (2014): 91-106.

Shubik, Martin. “Game Theory, Behaviour, and the Paradox of the Prisoner’s Dilemma.” *The Journal of Conflict Resolution* 14, no. 2 (1970): 181-193.

Siracusa, Joseph and Aiden Warren. “The Nuclear Non-Proliferation Regime: An Historical Perspective.” *Diplomacy and Statecraft* 29, no. 1 (2018): 1-26.

Stein, Arthur. “Coordination and Collaboration: Regimes in an Anarchic World.” *International Organization* 36, no. 2 (1982): 299-324.

Stein, Arthur. *Why Nations Cooperate: Circumstance and Choice in International Relations.* Ithaca: Cornell University Press. 1990.

Takahashi, Masao, Nial Fraser, and Keith Hipel. “A Procedure for Analyzing Hypergames.” *European Journal of Operational Research* 18 (1984): 111-122.

Taliaferro, Jeffrey. “Security Seeking Under Anarchy.” *International Security* 25, no. 3 (2000): 128-161.

Tang, Shiping. “Fear in International Politics: Two Positions.” *International Studies Review* 10 (2008): 451-471.

Tang, Shiping. “The Security Dilemma: A Conceptual Analysis,” *Security Studies* 18, no. 3 (2009): 587-623.

Trager, Robert. “The Diplomacy of War and Peace.” *Annual Review of Political Science* 19 (2016): 205-228.

“Treaty Between the United States of America and the Union of Soviet Socialist Republics on the Limitation of Anti-Ballistic Missile Systems (ABM Treaty).” 1972.

Van Lange, Paul. “The Pursuit of Joint Outcomes and Equality in Outcomes: An Integrative Model of Social Value Orientation.” *Journal of Personality and Social Psychology* 77, no. 2 (1999): 337-349.

von Neumann, John and Oskar Morgenstern. *Theory of Games and Economic* Behaviour. Princeton, NJ: Princeton University Press. 1944.

Waltz, Kenneth. *Man, the State, and War: A Theoretical Analysis*. New York, NY: Columbia University Press. 2001.

Waltz, Kenneth. *Theory of International Politics*. Boston, Mass.: McGraw-Hill. 1979.

Wang, Muhong and Keith Hipel, and Niall Fraser. “Misperceptions and Hypergame Models of Conflict.” *Behavioral Science* 33, no. 3 (1988): 207-223.

Wendt, Alexander. “Anarchy is What States Make of It: The Social Construction of Power Politics.” *International Organization* 46, no. 2 (1992): 391-425.

Yoder, Brandon and Kyle Haynes. “Signalling under the Security Dilemma: An Experimental Analysis.” *Journal of Conflict Resolution* 65, no. 4 (2020): 672-700.