Theories of Mind in International Relations

Why Individuals' Inferences about States' Motives Vary



Paola Magdalena Solimena
University College
University of Oxford

Thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in International Relations

Hilary Term 2023

Word Count: 65,817

Abstract

In international politics, communication about motives can make the difference between bargaining and war, policies of mutual reassurance and arms races, a state of perpetual conflict and the pursuit of a shared vision for international society. However, despite its centrality, empirical research on the communication of motives has long lagged behind theoretical scholarship, and fundamental questions about the microfoundations of motives remain: how do individuals infer a state's motives from its actions, and why do individuals' inferences dynamically vary, sometimes converging, other times diverging, in response to a state's signal? This thesis draws on psychology and the philosophy of mind to explore these questions. First, I examine the theoretical foundations of a recent wave of psychological research in International Relations and show that this research shares with rationalism a conception of the mind as computer. In order to explain variation in belief formation and updating about motives, I argue, we must move beyond this conception and integrate rationalist and psychological approaches to information processing with constructivist insights into social facts and systems of meaning. Second, I develop a domain-specific theory of attribution to explain how individuals infer motives in international politics. Combining a cross-national survey experiment with an in-depth analysis of evolving perceptions of Soviet motives during the Carter Administration, I find support for the costly signaling hypothesis, yet also show that the way individuals infer motives is shaped by their ontological assumptions about the target. Importantly, differences in these assumptions do not always result in divergent responses to costly signals: whether or not they do depends on the signal in question. Offering novel evidence on the attributional processes that underpin costly signaling, this thesis has important implications for the risk of escalatory dynamics between states and the study of the mind in International Relations.

To my mom

Acknowledgements

I thank University College and the UK Economic and Social Research Council for funding my DPhil. Without their support, I could not have pursued a doctorate at Oxford University.

This thesis would not have been possible without the encouragement, patience, and support of many people. I am especially indepted to Dominic Johnson and Jonathan Leader Maynard, who supervised my MPhil and DPhil theses. Dominic nurtured my fascination with interdisciplinary research at the intersection of International Relations and cognitive science, encouraged me to use experimental methods at a time when they were not yet part of the graduate curriculum in International Relations at Oxford, and supported me in taking risks to pursue my ideas. Jonathan nurtured my interest in the theoretical foundations of International Relations, and his questions, analytical rigor, and meticulous proof-reading of chapter drafts challenged me to balance nuance with clarity in developing my ideas.

Intellectually, I am also indepted to Eddie Keene whose MPhil course on 'The making of modern international society' instilled me with an appreciation for, and lasting interest in, intellectual and disciplinary history. Eddie's probing questions continue to shape the way I think, learn, and develop ideas about international politics.

Throughout the process of researching and writing this thesis, I have benefitted greatly from the constructive and sharp feedback I received from Janina Dill and Neil MacFarlane, who were my confirmation examiners and diligently commented on early drafts of chapters 1, 3, and 5, as well as from Todd Hall, Emanuel Adler, Duncan Snidal, and Stefano Guzzini, who generously engaged with my work at various stages of the DPhil.

The Centre for Experimental Social Sciences (CESS) provided invaluable support in conducting my first experimental study. I especially thank Noah Bacine for generously sharing his time and expertise on experimental methods and discussing drafts of the experimental design, Ray Duch for offering invaluable comments on the first and final drafts of the design, and Melanie Sawers for her patience and administrative support in implementing the experiment. I thank the Department of Politics and International Relations for funding the study.

I also received valuable comments and encouragement from participants at the IR Colloquium, CESS Colloquium, and American Politics and Graduate Seminar at Oxford, as well as Annual Meetings of the International Studies Association and International Society of Political Psychology.

Importantly, completing this DPhil was a personal journey as much as it was an academic journey, and my greatest debt is to the staff, mentors, friends, and family who supported me. In particular, I thank Alice and Violet for bringing cheerfulness and warmth to long days at the department, and Andrew for

helping resolve many an administrative issue, even – and especially – during the pandemic.

I thank University College for bringing me into its welcoming and lively community, and its faculty and staff for creating a safe and stimulating environment in which to explore ideas and make new experiences. My particular thanks go to Andrew Gregory for taking the time to listen.

Looking back onto the DPhil journey, I consider myself enormously lucky that the lows of that journey are most memorable, not for the challenges they posed, but for the friendships they gave rise to. I especially thank Jackie, whose wit, humor, and empathy have brightened many a moment on the DPhil, Arthur for exploring the magical lands that lie beyond the stairwell with me, and Keno, Lili, Hendrik, Dan, David, and Vanessa for their wonderful and enriching companionship.

I am also fortunate to have had the loyal support of friends who have known me long before I embarked on the DPhil journey. In particular, I thank Lisa and Johanna – I could not have wished for two more inspiring, dedicated, and self-determined women to weather my late teens and twenties and share many moments of the DPhil with.

I would not have kept faith in my project or my ability to complete it without the unwavering support of my college advisor, Emily Jones. Emily went above and beyond to mentor me throughout the DPhil: she took the time to talk through important decisions with me, she cheered me on through the most challenging moments of the DPhil, and she celebrated the highs of the DPhil with me. The deep sense of purpose she brings to her work inspires me to one day become a great mentor to my own students, too.

Finally, I would not have had the courage and determination to pursue and complete a DPhil at Oxford without the unconditional love, support, and great loyalty of my grandparents, aunt, and sister, and the deep trust of my mom, who has modeled the courage it takes to lead a self-determined life. I dedicate this thesis to her.

Contents

List of Figures xi

List of Tables xiii

1 Introduction 1
- 1.1 Argument . . . 3
- 1.2 Theoretical Foundations . . . 5
- 1.3 Theoretical Contributions . . . 8
- 1.4 Methods . . . 11
- 1.5 Plan of the Thesis . . . 13

2 Theories of Mind in International Relations 17
- 2.1 Motives in International Relations: Theories and Methods . . . 17
- 2.2 Conceptualizing Motives . . . 22
- 2.3 Motives and the Puzzle of Attribution . . . 24
- 2.4 A Theory of Attribution for International Relations . . . 31

3 Rationalism and Psychology in International Relations 35
- 3.1 Introduction . . . 36
- 3.2 A Tale of Two Minds . . . 38
- 3.3 Mind Metaphors Matter . . . 44
  - 3.3.1 Rational Choice Theory: The Mind as a Black Box? . . . 44
  - 3.3.2 Psychology: The Mind as Computer . . . 47
  - 3.3.3 Of One Mind: Rational Choice and Psychology . . . 52
- 3.4 Implications for International Relations . . . 56
  - 3.4.1 Reform, Not Revolution . . . 56
  - 3.4.2 Variation in Mind . . . 62
  - 3.4.3 A Meeting of Minds . . . 66
- 3.5 Conclusion . . . 67
- References . . . 69

4 Implicit Theories and Signaling in International Politics 73
- 4.1 Introduction . . . 74
- 4.2 Costly Signaling and Attribution . . . 77
  - 4.2.1 Attribution and Heterogeneity in Costly Signaling . . . 79
- 4.3 Research Design . . . 83
  - 4.3.1 Hypotheses . . . 88
- 4.4 Results . . . 89
  - 4.4.1 Sample . . . 89
  - 4.4.2 Randomization and Balance Checks . . . 91
  - 4.4.3 Analysis . . . 92
- 4.5 Discussion . . . 115
- 4.6 Conclusion . . . 117
- 4.7 Appendix . . . 121
  - 4.7.1 Ethical Approval . . . 121
  - 4.7.2 Experimental Instructions . . . 122
  - 4.7.3 CONSORT Participant Flow Diagram . . . 124
  - 4.7.4 Supplementary Data Analysis . . . 126

4.7.5 UK Sample: Experimental Findings . . . 135
4.7.6 Note on Timing of Data Collection . . . 138
References . . . 139

5 Implicit Theories, Attribution, and the End of Detente 145
5.1 Introduction . . . 146
5.2 Costly Signaling and the Puzzle of Attribution . . . 149
5.3 Attribution in International Politics . . . 153
5.4 Research Design . . . 161
5.5 Attribution and the End of Detente . . . 164
5.5.1 Implicit Theories of the Soviet Union . . . 164
5.5.2 Intervention in the Horn . . . 168
5.5.3 Soviet Brigade Crisis . . . 176
5.5.4 The Invasion of Afghanistan . . . 185
5.6 Conclusion . . . 195
References . . . 198

6 Conclusion 207
6.1 Summary of Arguments and Findings . . . 207
6.2 Avenues for Future Research . . . 211
6.2.1 The Causal Role of Implicit Theories . . . 211
6.2.2 Other Types of Signals . . . 212
6.2.3 Attribution and the Levels of Analysis Problem . . . 213

Bibliography 215

List of Figures

1.1 Implicit theories in international politics: a typology 4
1.2 Explaining attributional patterns: a theoretical framework 5

2.1 Distinguishing motives from related concepts: a schematic depiction . . . 23

4.1 Design and survey flow 84
4.2 The effect of cost of action and strategic context on perceived expansionism and behavioral predictions 99
4.3 The marginal effect of cost of action and strategic context on perceived expansionism and behavioral consistency 100
4.4 The effect of strategic context on perceived defensiveness and resolve . . . 101
4.5 Perceived expansionism across values of entity theory and entitativity . . 104
4.6 Perceived expansionism across values of interdependence and independence 106
4.7 Behavioral predictions across values of entity theory and entitativity . . . 108
4.8 Behavioral predictions across values of interdependence and independence 110
4.9 Perceived defensiveness across values of entity theory, entitativity, interdependence, and independence 113
4.10 CONSORT flow chart for the U.S. sample 124
4.11 CONSORT flow chart for the UK sample 125
4.12 Attrition in the U.S. sample: randomization inference 126
4.13 Attrition in the UK sample: randomization inference 127
4.14 Covariate balance for U.S. sample 128
4.15 Covariate balance for UK sample 129
4.16 Randomization check for the U.S. sample 131
4.17 Randomization check for the UK sample 132
4.18 Heat map: correlations of pre-treatment covariates for the U.S. data . . . 133
4.19 Distributions of key covariates in the U.S. sample 134
4.20 The marginal effect of cost of action and strategic context on perceived expansionism and behavioral predictions for the UK sample 135

List of Tables

2.1 Theories of motives and the puzzle of attribution. 30

4.1 U.S. sample size and randomization 91
4.2 UK sample size and randomization 91
4.3 Treatment effects with and without covariate adjustment 98
4.4 Implicit theories and self-construal as moderators of signaling 105
4.5 Implicit theories and self-construal as moderators of signaling 107
4.6 Implicit theories and self-construal as moderators 111
4.7 Treatment effects with and without covariate adjustment 136
4.8 Treatment effects with and without covariate adjustment 137

5.1 Implicit theories and attribution in international politics: theoretical predictions. 161
5.2 Implicit theories in the Carter Administration. 164
5.3 Belief updating about Soviet motives during the Carter Administration: key findings 194

1 Introduction

How do actors in international politics infer states' unobservable motives? Why do states sometimes succeed at peacefully negotiating mutually acceptable agreements and other times fail? These questions are central to scholarship on signaling and bargaining in international relations. While some scholars maintain that another state's motives are all but unknowable and, therefore, remain marginal to states' strategic calculations,[1] many scholars point to the ubiquity of assessments of states' motives in foreign policy and strategic decision making and agree that understanding these assessments is essential to explaining international outcomes.[2] Hardly any event since the end of the Cold War has done more to highlight this than the Ukraine war and the distinct possibility of a military confrontation between NATO and Russia: as decision makers walk a fine line between deterrence and escalation, their ability to assess Russian motives accurately and convey NATO's motives credibly is vital to avoiding the widening of war.

In an effort to understand how they do, scholars in international relations have argued that a state's domestic characteristics or costly behavior serve as readily observable and credible indicators of its motives.[3] Alternatively, they have emphasized the role of prior beliefs in the inference of states' motives, theorizing that individuals – decision makers or members of the public alike – attend to information that concurs with their prior beliefs about a state, or the nature of international politics generally, but disregard it otherwise.[4]

However, international politics abounds with instances where individuals' inferences about a state's motives neither consistently converge in line with a state's domestic characteristics and costly behavior nor consistently diverge in accordance with individuals' prior beliefs. Instead, their inferences dynamically vary, sometimes diverging and other times converging, in response to a state's behavior. For example, when the Soviet Union intervened in the Ogaden War on behalf of Ethiopia in 1977, top-level decision makers within the Carter Administration drew divergent conclusions about Soviet motives. By contrast, the same decision makers temporarily converged in their concern over Soviet motives in 1979, when U.S. intelligence reports identified a Soviet brigade with an unknown mission

[1] Mearsheimer 2014; Rosato 2015; Waltz 1979.
[2] For example, Glaser et al. 2016; Trager 2017; Yarhi-Milo 2014.
[3] For example, Fearon 1997; Glaser 2010; Kydd 2005.
[4] For example, Holsti 1962; Jervis 1976; Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Yarhi-Milo 2014.

in Cuba. Subsequently, as additional information about the brigade became available, they diverged again. And, when the Soviet Union invaded Afghanistan in 1979, the same decision makers made similar inferences about Soviet motives: although their substantive reasoning and policy recommendations partly differed, they all agreed that Soviet goals were more expansive than the Carter Administration had previously assumed.

In this example, superpower rivalry and opposing political systems predisposed decision makers on both sides of the Iron Curtain toward mistrust. Moreover, in all three cases, Soviet behavior – or what appeared to be Soviet behavior – incurred a significant cost to the Soviet Union. Yet, top-level officials within the Carter Administration did not consistently or gradually[5] converge in their interpretations of that behavior. Nor, however, did they consistently diverge in keeping with their prior beliefs about Soviet motives. Rather, the inferences individual decision makers drew from Soviet actions sometimes converged, other times diverged, varying both across decision makers and within decision makers over time. What explains this variation?

A key claim I make in this thesis is that attribution – that is, the ascription of behavior to underlying causes – is key to explaining this variation. In psychology, the important role that attribution, and our capacity as human beings to infer another's motives, play in navigating social interactions, and forming and maintaining relationships, is well-recognized.[6] I argue that attribution plays a similarly vital role in interactions between states, with important consequences for states' ability to communicate their motives and avoid costly military disputes.

The idea that attribution extends from interpersonal to interstate interactions and can shed light on the microfoundations of bargaining between states is not itself new. International relations scholars have suggested that attributional biases found in interpersonal and intergroup settings in psychology extend to interstate interactions, predisposing actors to underestimate their own role in driving another state's malign policies and increasing the risk of arms races.[7] However, scholarship to date has not addressed the question of how individuals attribute the behavior of states as distinct from the behavior of individu-

[5] Depending on individuals' priors and the event in question, Bayes' rule implies that beliefs will converge gradually in response to several successive events rather than immediately in response to a single event (e.g. Jackman 2009).

[6] A seminal example is Heider 1958.

[7] Jervis 1976; Larson 1997a; Rosenberg and Wolfsfeld 1977.

als or individual group members.[8] When do individuals attribute a state's behavior to its underlying motives, and when to its security environment? When do they attribute the behavior of states to individual state actors, and when to states as a whole?

This thesis begins to address these questions. It develops a domain-specific theory of attribution to understand how individuals – decision makers and members of the public alike – ascribe the behavior, and infer the motives, of states. By systematically investigating the attributional processes that underpin signaling between states, the theory provides international relations scholars with an explanation of why individuals' beliefs about a state's motives dynamically vary, sometimes converging, other times diverging, in response to costly signals.

1.1 Argument

A central argument I advance is that individuals' assumptions about the target of attribution in international politics – states and individual state actors – are essential to developing a theory of attribution for international relations and to explaining variation in individuals' inferences about states' motives. Accordingly, I take 'implicit theories', a concept that originates in social psychology, as a starting point for building such a theory.[9] An implicit theory is a set of assumptions that individuals tacitly hold about the agents that populate the social world, their fundamental properties, and the relations between them.[10] As I will argue, information processing in general, and attribution in particular, presuppose representational content, including implicit theories, and implicit theories shape the way individuals attribute a state's behavior and, thus, form beliefs about its motives. Put simply, differences in individuals' implicit theories result in differences in their attributional processes.

I focus on those components of an individual's implicit theory that most immediately bear on the attribution of states' and state actors' behavior: individuals' assumptions

[8] A notable exception to this is Rosenberg and Wolfsfeld 1977, which focuses on the attribution of group behavior.

[9] 'Implicit theories' and 'lay theories' are used interchangeably in psychology. In this thesis, I adopt the former in order to avoid an emphasis on the contrast between lay and scientific theories (see Levy, Chiu, and Hong 2006, pp. 6-7), which raises fundamental questions at the intersection of epistemology and psychology, and instead emphasize the tacit nature of these theories.

[10] Levy, Chiu, and Hong 2006; Plaks, Levy, and Dweck 2009; Molden and Dweck 2006; Haslam, Rothschild, and Ernst 2004.

about individuals, states, and their relationship to one another. Drawing on findings from research on attribution, implicit theory, and group perception in psychology, I propose a typology that distinguishes between four types of implicit theory in international politics. The typology is derived from two underlying dimensions: entitativity and mutability (see figure 1.1). The first, entitativity, is a concept I borrow from social psychology and responds to the question, how unified is a state?[11] A person who scores high on entitativity assumes states to be unified group agents; a person who scores low on entitativity assumes states to be more akin to aggregates of subgroup agents like leaders. The second, mutability, responds to the question, how changeable is the agent? A person who scores high on mutability assumes that the properties of agents – whether individual or group agents – are changeable; a person who scores low on mutability assumes that the properties of agents are highly static.

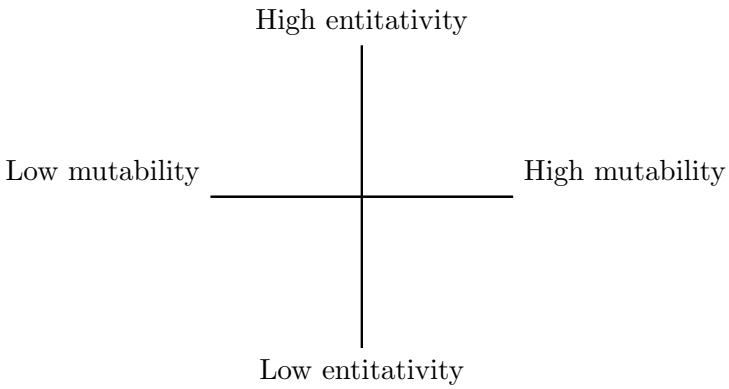


Figure 1.1: Implicit theories in international politics: a typology

Integrating findings from attribution research with scholarship on signaling in international relations, I offer predictions for how each type of implicit theory will shape individuals' attributions of costly signals. Specifically, I argue that differences in implicit theories will lead to differences in the extent to which individuals regard dispositions as properties of states or individual state actors and how responsive individuals are to situational factors shaping a target's behavior, with significant implications for the risk of arms races. Crucially, these differences do not necessarily generate divergent attributions of a state's signal: because implicit theories affect the type of information that individuals attend to in making behavioral ascriptions, the resulting attributions depend on the informational structure of the signal in question.

[11] Campbell 1958.

Put simply, the theory presented in this thesis argues that attribution, and the inference of motives in international politics, depends on the *interaction* between an individual's implicit theory and the informational structure of a state's signal (see figure 1.2). Thereby, it differs from structural theories of signaling that emphasize the role of external cues in assessments of motives. Structural theories tacitly assume that actors respond to external cues uniformly. By contrast, the theory presented here joins broader calls within international relations to recognize actor-level variation in perceptions of, and responses to, signals.[12] At the same time, the theory differs from recipient-driven theories, which explain actor-level variation based on individual beliefs or preferences. These theories tend to emphasize belief consistency, arguing that actors with distinct prior beliefs will diverge in their assessments of a state's motives. By contrast, the theory presented here argues that implicit theories shape the *process* of belief formation rather than its outcome. As a result, it is able to marry the strengths of structural theories with the strengths of recipient-driven theories and explain how beliefs about a state's motives dynamically vary across individuals *and* within individuals over time.

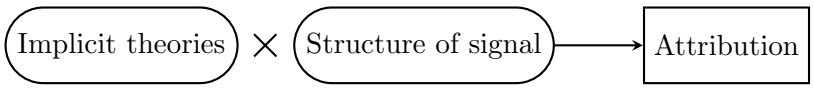


Figure 1.2: Explaining attributional patterns: a theoretical framework

1.2 Theoretical Foundations

In developing a theory to understand how individuals infer states' motives and, crucially, why belief updating about motives varies across individuals, this thesis builds on existing theories and tools to study signaling and perception in international relations. Specifically, the thesis draws together insights from rationalist, psychological, and constructivist approaches to international relations. This is not a mere by-product of academic rigor, nor is it a principled statement on eclecticism. It is a core feature of the theory of attribution proposed here, and reflects a broader argument I make concerning the study of perception and the mind in international relations: as chapter 3 will show, in order to understand

[12] Hafner-Burton et al. 2017; B. C. Rathbun, Kertzer, and Paradis 2017; B. C. Rathbun 2019; Yarhi-Milo 2014.

variation in the perception of signals in international relations, and develop actor-specific assumptions about strategic choice, scholars must integrate rationalist and psychological insights into information processing with constructivist insights into the intersubjective constitution of social reality and the social construction of meaning.

In particular, I build on the insights rationalist approaches offer into the role of uncertainty in bargaining between states, the role of costly signals in reducing this uncertainty, and the basic mechanism that underlies information processing in international politics. A key insight of rationalism is that uncertainty and a lack of information can cause war even when mutually preferable agreements short of war exist.[13] Uncertainty arises from states' incentives to misrepresent or conceal information about their relative military power or intentions (and thereby alter other states' strategic calculus).[14] The chief mechanism to reduce uncertainty, rationalist scholars argue, is costly signaling.[15] Costly signals are actions that incur a significant cost to the sender and are therefore more likely to be taken by sincere actors than insincere ones. By focusing on the question of how actors infer states' unobservable motives from their behavior, this thesis adopts rationalism's emphasis on information, or lack thereof, as a key factor in understanding international outcomes. Moreover, in answering this question, this thesis builds on the logic of costly signaling, theorizing that individuals infer a state's motives from its costly actions. That said, what individuals perceive as costly is not purely a function of external cues, as rationalist approaches suggest.

Along with its emphasis on uncertainty and information, rationalism has provided scholars with a basic mechanism to conceptualize information processing, that of Bayesian updating.[16] Bayesian updating posits that actors process information in accordance with Bayes' rule, updating their beliefs about the world in response to novel information (where 'belief' is defined as the probability of an event or state of the world). More precisely, it holds that actors update their beliefs to reflect the likelihood of novel information *given their prior beliefs*. The theory of attribution I advance incorporates this mechanism, hypothesizing that individuals respond to signals and update their beliefs in light of their prior beliefs about a state. At the same time, as I argue in chapters 2 and 3, forming and

[13] Fearon 1995.
[14] Fearon 1995.
[15] Fearon 1997; Glaser 2010; Kydd 2005.
[16] For example, Kydd 2005.

updating beliefs about a state's motives presumes the existence of mental concepts like 'states', 'leaders', or the 'international realm'. These concepts and their relationship to one another fundamentally structure the way information is processed. Therefore, Bayesian updating may be necessary but is not sufficient for understanding how actors form and update their beliefs about a state's motives in response to costly signals.

I also build on the insights psychological approaches provide into the role of perception in signaling in general and the role of attribution in communication about motives in particular. Unlike structural-realist theories of international relations which focus on macro-level factors,[17] and rationalist theories which stress information,[18] psychological approaches call attention to the important role micro-level, cognitive processes play in generating international outcomes. Proponents of psychological approaches highlight the indeterminacy of structure and information in explaining these outcomes; a theory of cognition is needed to clarify how external cues, like signals, are processed and enter strategic choice, either facilitating or hampering communication between states.[19] Offering such a theory, psychological research on signaling in international relations has imported findings from attribution research in psychology to understand how actors infer states' unobservable motives from their observable behavior.[20] This thesis adopts psychological approaches' focus on perception and takes attribution as its starting point for investigating the microfoundations of signaling about motives in international politics. It theorizes that the way individuals ascribe states' behavior explains their responses to signals and their perceptions of a state's motives. That said, attributional processes are not uniform across individuals or distinct types of targets, as psychological work in international relations tacitly assumes, and extending findings from attribution research in psychology to interstate interactions requires an understanding of the perceivers of signals and the targets of attribution in international politics.

Last but not least, I build on constructivist insights into the intersubjective constitution of social reality and the social construction of meaning. Whereas structural, rationalist, and psychological theories of international relations tend to treat the agents of international politics, notably (nation-)states, as given and fixed, constructivist scholars

[17] For example, Mearsheimer 2014; Waltz 1979.
[18] For example, Fearon 1995; Glaser 2010.
[19] Jervis 1976.
[20] Jervis 1976; Larson 1997a; Rosenberg and Wolfsfeld 1977.

emphasize that these agents, and the international realm as a whole, are intersubjectively constituted and do not exist independently of the social processes that generate and sustain them.[21] This thesis takes this insight seriously and argues that information processing is premised on social facts like the existence of states as group agents in international politics. For example, ascribing a state's decision to increase its military spending to that state's motives presumes the existence of an entity, 'state', that is capable of having intentions. Consequently, as chapter 3 will elaborate, it is necessary to ground theories of perception and behavior in international relations in a domain-specific understanding of the social facts that structure interactions in the international realm.

More broadly, constructivist approaches agree with psychological approaches in arguing that external stimuli like signals do not speak for themselves; actors perceive or interpret them, and we must understand how they do in order to explain foreign policy choices and international outcomes. However, unlike psychological approaches, constructivist scholars focus on the shared meaning and expectations actors attach to particular stimuli rather than individuals' cognitive processes in order to do so.[22] Retracing how actors make and remake this meaning through their interactions with one another, constructivist scholars show that the interpretation of external stimuli like signals is both heterogenous and patterned: it varies across social contexts and actors, yet systematically so.[23] The theory of attribution presented here incorporates this emphasis on socially constructed meaning, arguing that actor-level assumptions about the agents that populate the international realm shape the way individuals attribute a state's behavior and, thus, update their beliefs about the state's motives. That said, for any one individual, the sites of meaning-making are many and overlapping, and this thesis does not assume that interstate interactions are the sole, or even primary, site of meaning-making when it comes to notions of agency in international politics.

1.3 Theoretical Contributions

While this thesis builds on rationalist, psychological, and constructivist insights into signaling and perception in international relations, it also makes novel contributions to each

[21] For example, Jackson and Nexon 1999.
[22] For example, Wendt 1999.
[23] Wendt 1999.

approach. For rationalist scholars, this thesis provides a systematic investigation into the microfoundations of signaling. From a rationalist perspective, the problem of motives, i.e. uncertainty, and its solution, i.e. costly signals, are both grounded in a core feature of international politics: strategic interdependence. Because international outcomes result from the choices of multiple states, the strategic calculus of any single state dynamically depends on the actual and anticipated choices of all others.[24] Linking rationalist arguments to the concept of attribution in psychology, this thesis fleshes out the implications of strategic interdependence for the cognitive processes that underpin costly signaling at the individual level. It argues that the costly signaling mechanism hinges on particular assumptions about how actors attribute behavior and investigates these assumptions empirically. In doing so, this thesis offers a novel, experimental test of the costly signaling hypothesis. The findings contribute to a broader effort within international relations to ground rationalist and game-theoretic models in empirically-validated actor assumptions.

For proponents of psychological approaches to international relations, this thesis offers a new, domain-specific theory of attribution in international politics. Attribution itself is not new to scholarship on signaling. Based on findings from attribution research in psychology, scholars have argued that actors in international politics will tend to overestimate the extent to which a state's malign policies are driven by its motives, and underestimate the extent to which they are driven by fear in response to actors' own policies, increasing the risk of arms races.[25] However, the empirical evidence scholars offer in support of this argument rarely includes a direct test of the microfoundational claims that underpin it.[26] This compounds two deeper issues: the question of how findings from attribution research in psychology extend to international politics remains underdeveloped, as does the question of how these findings bear on established theories of signaling. The theory of attribution presented here makes three significant innovations that address these gaps: drawing on psychological research on group perception, it incorporates the target of attribution in international politics – states – into the theory, offering a domain-specific theory of attribution; integrating attribution research with rationalist arguments about costly signaling, it elaborates the implications of attribution for the costly signaling mechanism;

[24] Morrow 1999; Schelling 1960.
[25] Jervis 1976; Larson 1997a.
[26] Two notable exceptions are Rosenberg and Wolfsfeld 1977 and Dickson 2009. Only the former explicitly examines the attribution of state, as opposed to individual, behavior.

building on psychological research on implicit theories, it theorizes variation in individuals' attributional patterns in international politics. Empirically, the thesis offers a novel, experimental test of whether findings from attribution research in psychology extend to the domain of international politics.

More broadly, this thesis advances recurring and foundational debates about the role of psychology in international relations theory. Whereas some scholars regard psychology as the starting point for a 'behavioral revolution' within the discipline,[27] this thesis shows that the introduction of psychology into international relations is better understood as a reform than a revolution: psychological approaches compete only with a subset of rational choice theories and, where they do, they fundamentally share with rational choice a conception of the mind as computer. Drawing on the philosophy of mind, I discuss the limitations of this conception of the mind for the psychological study of international relations, along with the possibility of addressing them by integrating psychological research with constructivist insights into the social processes that underpin meaning making.

Finally, for constructivist scholars, this thesis develops new conceptual tools to understand how socially-constructed meaning shapes information processing and the way individuals respond to signals in international politics. While constructivist scholarship offers valuable insights into the role that complex systems of meaning, norms, and identities play in shaping actors' perceptions and behavior, calls within the discipline to integrate these insights further with psychological research on the microfoundations of interstate interactions have often met with scepticism from constructivist scholars who reject psychology's tendency to reduce socially-constructed meaning to individual beliefs or preferences – essentially, another form of input into uniform and involuntary mechanisms of information processing. The arguments in this thesis take a different approach, theorizing that implicit theories fundamentally shape the way information is integrated and, therefore, processed. In other words, socially-constructed meaning is more than input; it is the pre-requisite for higher order processes like attribution and the formation of beliefs about states' motives. In this sense, this thesis helps pave the way for a more encompassing dialogue between constructivist scholars and advocates of psychological research in international relations.

[27] Hafner-Burton et al. 2017; Kertzer and Tingley 2018.

1.4 Methods

The methodological approach of this thesis mirrors its theoretical argument: just as a theory of attribution for international politics requires the integration of rationalist, psychological, and constructivist approaches to international relations, so too the empirical study of signaling about motives requires a combination of experimental and qualitative methods. An important obstacle to the empirical study of motives, and to investigating how individuals draw inferences about states' motives in particular, is the difficulty of observing actors' beliefs about states' motives, let alone measuring and isolating possible causes. To overcome this obstacle, I proceed in two steps. First, I conduct a survey experiment investigating responses to sunk-cost signals of resolve among more than 800 adults in the U.S. and 800 adults in the UK.[28] The experimental setting allows me to observe the conclusions individuals draw from sunk-cost signals directly rather than infer them indirectly from verbal utterances made, or behavior displayed, in policy settings. Moreover, by randomly assigning participants to treatment conditions that vary the cost of a state's action and the strategic environment in which the action was taken, the experiment offers a robust investigation into the attributional processes that underpin signaling in international politics and, crucially, the causal role that a state's strategic environment plays in shaping the conclusions recipients draw from a state's actions. Finally, the experimental design allows me to collect data on a range of psychological variables that are known to moderate attributional processes, notably implicit theories and self-construal. Leveraging this data, the experiment offers initial insights into systematic variation in attributional patterns among individuals in international politics.

The results show that a state's strategic environment affects the extent to which individuals attribute a state's actions to its motives: consistent with the costly signaling hypothesis, individuals attribute a state's sunk-cost signals of resolve more to its motives when the state's environment is secure than when the state faces an external threat. Importantly, this finding holds even when the individuals in question are members of the state posing the external threat. However, the results also suggest that individuals' attri-

[28] Throughout this thesis, I use 'signals of resolve' as a shorthand to refer to foreign policy choices that are competitive rather than cooperative. I do not focus on signaling and inferences about resolve *per se* (see section 2.2).

butional patterns systematically vary depending on their implicit theories of individuals and states, as well as their self-construal. In particular, individuals who consider personality to be more static, and states to be more unified, are less responsive to changes in a state's strategic environment. These findings show that canonical results from attribution research in psychology, notably attributional biases found in interpersonal and intergroup settings, do not necessarily extend to international politics, increasing the risk of arms races, as some scholars have argued.[29] Thereby, they highlight the need for a *domain-specific* theory of attribution and the importance of rigorously testing microfoundational claims drawn from other fields of research before applying them to international relations. At the same time, the findings underscore the import of psychology, and research on implicit theories in particular, for recipient-driven theories of signaling.

In a second step, and building on these findings, I begin to develop a domain-specific theory of attribution to account for variation in the inferences individuals draw from costly signals. As an initial test of the theory, I conduct a plausibility probe and qualitatively analyze three elite decision makers' evolving perceptions of Soviet motives during the Carter Administration. In particular, I compare the three decision makers' responses to three instances of sunk-cost signaling by the Soviet Union. While all three instances involve competitive policies, they differ with respect to a key explanatory variable of the theory, namely the signal's informational structure. This approach allows me to pair comparative analysis with process-tracing and test the theory's ability to explain variation in the inference of motives both across individuals and within individuals in response to distinct signals. Taking a qualitative approach to testing the theory has three important advantages. First, the theory's workhorse is the typology of implicit theories I develop to explain systematic variation in individuals' attributional patterns. A qualitative approach allows me to study each decision maker's implicit theory in-depth. This facilitates the process of mapping complex belief systems onto one of the typology's four ideal types while maintaining a rich, contextual understanding of how the typology applies to real-world cases. In other words, the qualitative analysis puts flesh onto the typology's bones, bringing the second key explanatory variable of the theory to life. Second, by tracing how the three decision makers' perceptions of Soviet motives evolved over time, the qualitative

[29] Jervis 1976; Larson 1997a.

analysis allows me to study the explanatory role of implicit theories and attribution in belief change. Finally, by focusing on elite decision makers, the qualitative analysis offers an initial test of the theory among those individuals whose perceptions of signals most directly bear on foreign policy decisions and the trajectory of interstate relations.

The findings provide initial support for the domain-specific theory of attribution I advance in this thesis: in making behavioral attributions, decision makers who regarded individuals or the Soviet Union as a group agent as mutable were more responsive to information about the Soviet Union's strategic environment; in making dispositional attributions, decision makers who considered the Soviet Union to be more entitative attributed Soviet actions more to the Soviet Union as a whole than to the personal dispositions of leading individuals. Decision makers' implicit theories of the Soviet Union can thus explain why their beliefs about Soviet motives dynamically varied during the Carter Administration, sometimes diverging and other times converging in response to the Soviet Union's sunk-cost signals depending on the informational structure of the signal in question. More broadly, the analysis underscores the importance of decision makers' attributions in foreign policy analysis and shows how attribution combines with other, interpersonal and institutional, mechanisms to shape foreign policy choices.

1.5 Plan of the Thesis

The thesis is structured as follows. Chapter 2 surveys scholarship on motives in international relations. Although states' motives, and the question of how actors assess them, are central to international relations theory, empirical research on the communication of motives has long lagged behind theoretical scholarship. I discuss findings from a recent wave of empirical work on the communication of motives and the questions they raise about established theories of signaling, notably the costly signaling hypothesis. Building on rationalist insights into the role of strategic interdependence in interstate interactions, this chapter argues that attribution research can help address these questions. However, unlike previous scholarship in international relations, I argue that findings from attribution research in psychology cannot be applied directly to signaling in international politics; we require a *domain-specific* theory of attribution in order to understand how actors infer states' motives.

Chapters 3, 4, and 5 present three stand-alone papers and form the core of this thesis. While each paper makes a distinct contribution to the study of signaling and perception in international relations, jointly, the papers lay the foundations for, and begin to develop and test, a domain-specific theory of attribution in international relations.

Chapter 3 examines the theoretical foundations of a recent wave of psychological research in international relations. This research is often understood as part of a 'behavioral revolution' that reinserts the mind into the study of international politics. Contrary to this view, this chapter argues that mainstream, rationalist scholarship is not void of the mind and that psychological approaches share with rationalism a conception of the mind as computer. This conception faces severe limitations in understanding variation in the perception of signals because it relegates meaning to the role of individual preferences and beliefs – a type of input into uniform mental processes – sidelining the deeper role of meaning in cognition. To overcome these limitations and understand variation in perception, I argue, it is necessary to integrate rationalist and psychological insights into information processing with constructivist work on socially-constructed meaning. This argument lays the foundations for a domain-specific theory of attribution.

Chapter 4 investigates the microfoundations of costly signaling and presents the results from a novel, experimental test of the costly signaling hypothesis. Rationalist models have long assumed that actors in international politics rely on costly signals to communicate their motives. Building on the concept of attribution from psychology, this chapter shows that the costly signaling mechanism entails particular assumptions about how recipients attribute a sender's behavior. Integrating rationalist arguments with findings from research on attribution in social and cultural psychology, notably the role of implicit theories and self-construal in attribution, the chapter develops hypotheses about who is most likely to attribute sunk-cost signals in accordance with these assumptions and respond to costly signals the way rationalist theories predict. The findings from a cross-national survey experiment with over 1700 participants provide partial support for the costly signaling hypothesis and show that canonical findings from attribution research in psychology do not necessarily extend to interstate interactions as some scholars have suggested.[30] Yet, they also suggest that individuals' attributional patterns vary, with significant implications for

[30] Larson 1997a.

the risk of conflict escalation.

Chapter 5 develops a domain-specific theory of attribution in international politics to explain variation in individuals' attributional patterns and, thus, inferences from sunk-cost signals. Integrating rationalist arguments with attribution research, it theorizes that the attribution of sunk-cost signals depends on the interaction between an individual's implicit theory and the signal's informational structure, notably the salience of situational factors in a state's actions. As an initial test of the theory, the chapter compares the responses of three elite decision makers – President Carter, Secretary of State Cyrus Vance, and National Security Advisor Zbigniew Brzezinski – to three instances of Soviet sunk-cost signaling between 1977 and 1980. The findings support the theory, suggesting that decision makers' inferences from sunk-cost signals dynamically vary depending on their implicit theory and the informational structure of the signal, and shed light on the role of attribution in the deterioration of U.S.-Soviet relations in the late 1970s.

The concluding chapter, chapter 6, summarizes the arguments and findings of this thesis and discusses their implications for the study of signaling in international relations, notably the need to pair experimental methods with interpretive tools to understand how individuals process information and form beliefs. It also identifies several avenues for future research, including a direct test of the causal role of implicit theories in attribution, the theory's extension to other types of signals, notably costly signals of reassurance, and the theory's extension to other targets of attribution in international politics, notably individual state actors, which could help address the level of analysis problem in international relations.

2 Theories of Mind in International Relations

States' motives, and the question of how actors in international politics can communicate them, are central to understanding conflict and cooperation between states. A large literature within the field of international relations attests to this. In this chapter, I review this literature and, in particular, a growing body of empirical work investigating the microfoundations of motives: how actors in international politics do, in fact, convey and infer states' motives. Findings from this body of work raise important questions about established theoretical arguments within the field, notably the costly signaling hypothesis. Building on rationalist insights into the role of strategic interdependence in international politics, I argue that a theory of attribution can help address these questions. Although international relations scholars have previously drawn on attribution research from psychology, I show that a *domain-specific* theory of attribution is essential to advancing a microfoundational understanding of how states communicate about motives.

2.1 Motives in International Relations: Theories and Methods

States' motives, and the question of how actors in international politics assess them, gets to the heart of international relations theory and debates over the prospects for cooperation between states. On the one hand, structural realists maintain that states' motives are all but unknowable. In the absence of a central authority with the capacity to sanction actors, states are therefore doomed to compete for power in order to secure their survival.[1] From this perspective, though states need not be predatory by disposition, the structural imperatives of anarchy render any measure of uncertainty about another's motives too risky, compelling states to distrust one another and adopt strategies that maximize their relative power. As a result, international politics is characterized by perennial conflict.

On the other hand, institutionalist and rationalist scholars assert that states can assess one another's motives, albeit with varying degrees of (un)certainty. To be sure, these scholars recognize the formidable obstacles states face in doing so. Indeed, institutionalist and rationalist scholars have done much to flesh out why the anarchical structure of international politics engenders uncertainty and impedes cooperation.[2] However, they

[1] Mearsheimer 2014; Rosato 2021; Waltz 1979.
[2] Fearon 1995; Keohane 1984.

also identify a range of mechanisms that allow states to reduce this uncertainty. Notably, they argue that a state's behavior can be an important indicator of its motives if it incurs a significant cost to the sender and that international institutions can reduce uncertainty about motives by increasing the flow of information between states.[3] In virtue of these mechanisms, they argue, competition between states is not inevitable. Rather, whether or not states compete depends on the trade-offs they face between varying degrees of uncertainty about motives, the cost of foregoing cooperation, and the cost of competing. In certain circumstances, this trade-off works out in favor of cooperation.

Although institutionalist and rationalist scholars disagree with structural realists about the prospects for cooperation between states, they share with structural realists a focus on anarchy and states' lack of (credible) information as the principal causes of conflict in international politics.[4] In contrast, constructivist scholars agree with institutionalism's and rationalism's optimism about cooperation, yet ground this optimism in a wholly different set of arguments about the international realm. Focusing on meaning in lieu of information, and interpretation in lieu of information processing, they consider anarchy itself to be indeterminate: a state's behavior, they argue, does not speak for itself; rather, what it conveys depends on the systems of shared meaning and expectations that emerge from, and structure, states' interactions with one another.[5] Shared meaning and expectations explain why '500 British nuclear weapons are less threatening to the United States than 5 North Korean nuclear weapons', as Alexander Wendt famously argues.[6] On this view, cooperation is not merely a matter of increasing the quality or quantity of information available to states, but of constructing the very meaning information holds for actors. Where the social relations that underpin states' interactions are cooperative, cooperation is more than one strategic option among many: it becomes the default mode of interaction between those states.[7]

Over the past decade, challenges to American primacy and the liberal world order, and the emergence of China as a new center of power in world politics in particular, have stoked up these enduring, scholarly debates about motives and the prospects for

[3] Fearon 1997; Glaser 2010; Keohane 1984; Kydd 2005.
[4] B. C. Rathbun 2007; Tang 2008.
[5] Wendt 1992; Wendt 1999.
[6] Wendt 1995: 73. Also quoted in Hurd 2008.
[7] Hopf 2010.

cooperation between states.[8] In part, this development has encouraged scholars to revive and elaborate grand-theoretical arguments that reproduce familiar fault lines within international relations. In his recent book, Sebastian Rosato, for example, offers a major restatement of key realist arguments, renewing the debate between realist and rationalist scholars about whether states can, in certain circumstances, surmise another's intentions with some degree of certainty.[9] However, the field has also seen the rapid growth of a new body of empirical work on signaling and the inference of motives that seeks to move beyond these theoretical divides and investigate how actors in international politics do, in fact, communicate about motives.[10] Two broader developments within the discipline have facilitated the emergence, and shaped the trajectory, of this body of work.

First, as Joshua Kertzer observes, 'IR scholarship has taken a microfoundational turn', and scholars are increasingly interested in exploring the lower-level processes that underpin macro-level phenomena in international politics.[11] Insights into these lower-level processes not only add explanatory depth to macro-level claims,[12] but also help scholars identify the scope conditions of those claims. As advocates of microfoundational research point out, many macro-level claims in international relations theory already rely on microfoundational claims.[13] Neta Crawford, for example, argues that the emotion of fear plays a central, if undertheorized, role in producing realism's dire conclusions about the prospects for cooperation between states.[14] Similarly, several scholars have argued that rational choice theory assumes that actors in international politics apply the logic of utility maximization to their decision making and respond to new information about the international environment in accordance with Bayes' rule.[15] By validating these claims empirically, microfoundational research offers insights into the conditions under which competing theoretical arguments and predictions are likely to obtain.[16]

[8] For example Legro 2007; Rosato 2021.
[9] Rosato 2015; Glaser et al. 2016.
[10] A seminal example of this is Yarhi-Milo 2014.
[11] Kertzer 2017, p. 82.
[12] For example, Yarhi-Milo 2018.
[13] Goldgeier and P. E. Tetlock 2001; Kertzer 2016. From a philosophy of science perspective, the realism of certain assumptions or concepts that underlie macro-level theoretical claims is contested. For an overview of the debate as it pertains to rational choice theory, see MacDonald 2003 and chapter 3 of this thesis.
[14] Crawford 2000.
[15] Mercer 2005; B. C. Rathbun 2019.
[16] For example, Kertzer and McGraw 2012; B. C. Rathbun, Kertzer, and Paradis 2017. The emphasis this body of work places on the need to empirically validate actor assumptions tracks with important developments in economics, notably the emergence of behavioral economics, which supplies abstract,

As part of this broader development within international relations, scholarship on motives has begun to systematically test established, theoretical mechanisms of signaling, notably the costly signaling mechanism proposed by rationalist scholars,[17] as well as develop new mechanisms, like assymetric updating, vividness, personal impressions, or social bonding, to understand how states and state actors convey and infer motives.[18] Although the costly signaling mechanism is not inherently located at the individual level, this microfoundational research has focused predominantly on the individual level of analysis and, sometimes, the interpersonal levels of analysis.[19]

Related to the first development in international relations is a second: in the past few years, international relations scholars have increasingly embraced experimental methods as part of their methodological toolkit.[20] Two types of experiments have become particularly popular: lab experiments involving economics-style bargaining games and survey experiments employing vignettes. Whereas the former allow scholars to observe strategic behavior in tightly controlled settings, the latter allow scholars to study perceptions of, and beliefs about, actors in scenarios that, to varying degrees, resemble real-world situations or events.[21] Both types of experiments help scholars overcome two important obstacles to the empirical study of motives in international relations. First, states' motives and actors' perceptions of states' motives are difficult to observe, let alone measure.[22] This problem is particularly severe for signaling mechanisms located at the individual level: whereas historical documents like policy briefs offer a window into governments' past foreign pol-

game-theoretic models with empirically-validated behavioral assumptions (for example, Thaler and Sunstein 2009; Bruni and Sugden 2007).

17 Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Quek 2016; Tingley and Walter 2011; Yarhi-Milo, Kertzer, and Renshon 2018.

18 Hall and Yarhi-Milo 2012; Holmes and Wheeler 2020; Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Yarhi-Milo 2014.

19 The notion of an interpersonal level of analysis is developed by Wheeler 2018, and presents an addition to the three-tiered distinction between the individual, the state, and the interstate or systemic levels that is commonly used in international relations theory (Waltz 1959). The focus of recent empirical work on individuals has revived debates about the 'aggregation problem' in international relations (see Hafner-Burton et al. 2017; Powell 2017). Whereas critics consider the aggregation problem to be insurmountable and, consequently, doubt the utility of insights from the individual level of analysis for international relations theory, recent microfoundational work shows that empirical research can shed light on the question of aggregation, thereby closing the gap between individual-level findings and macro-level processes and outcomes (for example, Kertzer, Holmes, et al. 2022). A notable exception to the literature's focus on the individual level of analysis are Yarhi-Milo 2014, which also examines how intelligence organizations assess states' motives, and Trager 2017.

20 Hyde 2015.

21 Hyde 2015, p. 407.

22 Quek 2016.

icy goals, or assessments of another state's motives,[23] it is difficult to find an equivalent at the individual level.[24] In experiments, scholars can induce a specified goal to study its effects on signaling and employ behavioral or self-report measures to observe participants' perceptions and beliefs of other actors.

Second, it is difficult to isolate signaling mechanisms in real-world settings and, thus, to study when and how they are used, as well as their impact on actors' perceptions, behavior, and the outcome of strategic interactions.[25] In experimental designs, scholars can address this issue through random assignment: by randomly assigning participants to treatment groups that induce distinct goals (in bargaining games) or present distinct signals (in bargaining games or vignette-based experiments) while holding all other features of a game or scenario constant, they can establish the causal link between a sender's motives and the signals they send, as well as the causal link between signals and recipients' perceptions of the sender.[26]

By helping scholars overcome both obstacles, experiments facilitate the empirical study of signaling and increase the internal validity of microfoundational claims about interstate interactions. However, the gains in internal validity that scholars achieve by employing experiments come at a cost: to varying degrees, experimental designs abstract away from important features of international politics, decreasing their external validity.[27] For example, signaling experiments often include only a small number of iterations in strategic interactions, if they include any at all, essentially assuming a tabula rasa environment for strategic interactions.[28] Similarly, they abstract away from the complex decision making processes that characterize foreign policy and strategic decision making within states.[29] To address the resulting trade-offs between internal and external validity, empirical research on signaling and strategic choice tends to combine experiments with qualitative methods.[30]

[23] For example, Trager 2017; Yarhi-Milo 2014.
[24] Rare exceptions include personal diaries kept by leaders that later become accessible to scholars.
[25] Quek 2016: 927; Tingley and Walter 2011: 997.
[26] On the excludability assumption in experiments see Gerber and Green 2012, pp. 39-43.
[27] Hyde 2015. However, contrary to received wisdom, Brutger et al. 2022 suggests that the trade-offs between internal and external validity in experiments depends on the study's research aims and, generally, may be less severe than previously assumed. Also see McDermott 2002 for a discussion of this trade-off and its validity as a criticism of experiments in political science.
[28] Quek 2016.
[29] Kertzer, Holmes, et al. 2022 have recently started to use group experiments to address this issue.
[30] For example, B. C. Rathbun, Kertzer, and Paradis 2017; B. C. Rathbun 2019; Yarhi-Milo 2018.

As section 2.3 discusses, experimental research and qualitative research have jointly yielded a number of important insights into the communication of motives in international politics that challenge received wisdom in international relations theory. However, before we delve into these insights and the questions they raise, it is useful to clarify what 'motives' are – and what they are not.

2.2 Conceptualizing Motives

Despite the centrality of motives in international relations theory, scholars have adopted a variety of terms and concepts to study them. In this thesis, I define a state's motives simply as its goals in international politics.[31] A couple of things are worth noting about this definition. First, motives, thus understood, are distinct from intentions, which Sebastian Rosato defines as 'a state's...ideas about how it intends to behave'.[32] Unlike Rosato, who focuses on intentions to the exclusion of motives, scholarship in international relations has often either conflated the two concepts or at least left the distinction between them unclear.[33] Insofar as the distinction between motives and intentions mirrors that between ends and means, this is understandable: as Arnold Wolfers observes, 'it is impossible to draw a sharp line of demarcation between the two [goals and means]...means can be said to constitute intermediary or proximate goals, and few goals if any can be considered ultimate'.[34] Nonetheless, the distinction is an important one: as Charles Glaser elaborates, '[m]otives...are inherent features of states'.[35] They precede strategic interaction, remaining stable across distinct strategic contexts. By contrast, a state's intentions result from its motives *in interaction* with the incentives posed by its strategic environment; causally, they are one step removed from the state's motives and one step closer to its behavior.[36] For scholars, distinguishing between the two is essential to isolating and studying distinct causes of war.[37] For practitioners, distinguishing between the two is strategically important

[31] For example, Glaser 2010.
[32] Rosato 2021, p. 10.
[33] For example, Edelstein 2002 defines a state's intentions as 'that state's ambitions, how it is likely to act to achieve those ambitions, and the costs it will bear to realize those goals' (p. 3). Similarly, although Yarhi-Milo 2014 notes the distinction between intentions and motives, she defines intentions as an 'adversary's foreign policy goals...with regard to changing or maintaining the status quo, and...the adversary's inclination to use military force to achieve those objectives' (p. ??).
[34] Wolfers 1962, p. 68.
[35] Glaser 2010, p. 38.
[36] Glaser 2010, p. 38.
[37] Frieden 1999.

because motives, unlike intentions, help anticipate how a state will dynamically respond to changes in its strategic environment.[38]

Second, motives are distinct from resolve, which Kertzer defines as a state's willingness to '[maintain] a policy despite contrary inclinations or temptations to back down'.[39] Again, previous scholarship has often conflated the two concepts,[40] sometimes explicitly and deliberately so. In his seminal work on signaling in international relations, Robert Jervis, for example, incorporates a state's resolve into his notion of 'basic intentions', arguing that states should be classified based on their goal in combination with their 'willingness to run risks and make sacrifices to further this goal'.[41] For some analytical purposes, this makes sense: motives and resolve both bear on a state's strategic choices, and, like motives, resolve is at least partly dispositional.[42] However, conceptually, they remain distinct: whether a state is resolved to compete or to cooperate, it may be more or less resolved.

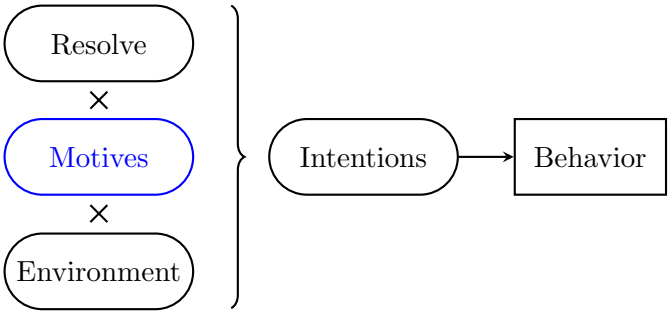


Figure 2.1: Distinguishing motives from related concepts: a schematic depiction

For the purposes of this thesis, these conceptual distinctions are important: clearly, although a state's resolve, motives, and intentions are all essential to understanding its strategic choices, how they relate to those choices – and to one another – differs. Accordingly, it stands to reason that individuals' inferences about them, and how individuals arrive at those inferences, will differ, too. Here, I am interested in how individuals infer states' motives specifically.

[38] Glaser 2010, p. 38.
[39] Kertzer 2016, p. 8.
[40] Kertzer 2016.
[41] Jervis 1976: 56
[42] Kertzer 2016. For Jervis, a state's resolve is distinct from 'its perceptions of costs and dangers' and entirely dispositional (Jervis 1976, p. 58).

2.3 Motives and the Puzzle of Attribution

At the heart of empirical research on motives is a theoretical conundrum: states' motives cannot be observed directly; they are 'private information' held exclusively by one party in an interaction.[43] To assess a state's motives and anticipate its responses to alternative policy options, actors in international politics therefore must rely on more readily observable, yet indirect, indicators, notably a state's behavior. However, the strategic nature of interstate interactions creates incentives for states with distinct motives to behave similarly in certain circumstances.[44] It also creates incentives for states to deliberately mask their true motives in a bid to secure a better deal.[45] Either way, the resulting problem is twofold: for recipients, it can be difficult to tell a benign state apart from a malign one, or a sincere state apart from an insincere one; for senders, it can be difficult to convey their true motives credibly. In other words, actors in international politics face uncertainty about motives at both ends of a strategic interaction. As rationalist scholars show, this uncertainty can have far-reaching consequences for international outcomes.[46] For example, it can lead to so-called security dilemmas, i.e. situations in which policies designed to increase a state's own security backfire, inadvertently reducing another state's security and setting off an escalatory dynamic that increases the cost of security for both states.[47]

As Robert Jervis explains, solving this conundrum and inferring a state's motives from its behavior 'involves the separation of internal from situational components' of the state's behavior:[48] observers must gauge the extent to which the state's behavior reflects its underlying motives or the incentive structures of its strategic environment. Only the former helps observers anticipate a state's behavior *across* strategic environments.[49]

Building on rationalist premises, theoretical scholarship in international relations has long maintained that costly signals are essential to solving the problem of motives. Costly signals are actions that incur a significant cost to the sender and are therefore more likely to be taken by sincere than insincere actors.[50] Accordingly, they serve as credible

[43] Fearon 1995.
[44] Glaser 2010; Jervis 1970; Morrow 1999.
[45] Fearon 1995; Schelling 1960.
[46] Fearon 1995; Glaser 2010.
[47] Jervis 1976; Kydd 1997; Tang 2009.
[48] Jervis 1976, p. 48.
[49] Glaser 2010.
[50] Fearon 1997; Glaser 2010; Quek 2021.

indicators of a state's underlying motives, compelling foes to take threats seriously and reassuring friends that cooperative gestures are genuine.[51]

Following the lead of theoretical scholarship, empirical research on signaling has begun to systematically test the costly signaling mechanism.[52] Focusing on the individual level of analysis, it makes three key findings. First, empirical research finds support for the costly signaling mechanism. Rationalist arguments about costly signals assume that actors form their beliefs about a state in accordance with Bayes' rule, revising them to reflect the probability of a signal given their prior beliefs about the state.[53] Accordingly, they predict that actors will update their beliefs about a state in response to a costly signal. Several studies bear out this prediction. For example, in a recent survey experiment, Kai Quek finds that recipients perceive a signal as more credible when the sender incurs a high cost than when the sender incurs a low cost or no cost at all, and confirms this finding across distinct types of cost.[54]

Second, and contrary to received wisdom, empirical research finds that costless signals can be perceived as credible indicators of motives, too. For example, in lab experiments investigating deterrence, Dustin Tingley and Barbara Walter show that cheap talk can be effective at deterring potential challengers.[55] Support for the idea that costless signals can be credible also comes from qualitative analyses of how decision makers infer states' motives. For example, Keren Yarhi-Milo finds that decision makers consider vivid information, particularly 'impressions from personal interactions with members of the adversary's leadership', to be credible.[56] As Hall and Yarhi-Milo argue, one reason for this is that we have limited control over our spontaneous emotional expressions, like our facial

[51] For example Fuhrmann and Sechser 2014; Kydd 2005; Yoder and Haynes 2021.

[52] Although this thesis focuses on motives in distinction from resolve (see 2.2), I include empirical work on assessments of resolve in my literature review for two reasons. First, in the literature, motives and resolve are sometimes conflated (e.g. Jervis 1976). Second, and more importantly, neither the arguments about costly signals in the theoretical scholarship nor the research designs used to test the costly signaling mechanism in empirical studies are specific to costly signaling about resolve (as opposed to motives). Therefore, insofar as the studies pertain to costly signaling, the findings are relevant to our discussion.

[53] insert ref

[54] Quek 2021. It is worth noting that the findings partly contradict findings from Quek 2016, which did not find support for the costly signaling mechanism on the recipients' end. Other studies that provide support for costly signaling are Kertzer, B. C. Rathbun, and N. S. Rathbun 2020 and Yarhi-Milo, Kertzer, and Renshon 2018. For a more detailed discussion of the experimental studies cited here see chapter 3.

[55] Tingley and Walter 2011. Other studies that find support for the effectiveness or credibility of costless signals are Quek 2016 and Quek 2021.

[56] Yarhi-Milo 2013, p. 12.

expressions, making them less amenable to bluffing and inherently more credible.[57]

Third, empirical research on motives finds significant variation in the inferences individuals draw from costly signals – both across individuals (between-subjects) and within individuals (within-subjects). For example, in a survey experiment investigating perceptions of Iran and support for the Iranian nuclear deal among Americans, participants systematically varied in the extent to which they revised their beliefs about Iran's trustworthiness in response to costly signals of reassurance.[58] Similarly, in a survey experiment exploring perceptions of resolve in naval disputes among public and elite samples, participants systematically varied in the extent to which they responded to costly signals of resolve depending on the type of cost the sender incurred: some participants revised their beliefs more in response to public threats ('tying-hands signals'); other participants revised their beliefs more in response to military mobilizations ('sunk-cost signals').[59] Meanwhile, qualitative analyses find that individual decision makers sometimes respond differently to signals, updating their beliefs about the sender's motives in one instance of signaling, but not another, even when the signals in question incurred the same type of cost.[60]

Whether or not this third finding is compatible with Bayesian updating is debatable. For a long time, psychological theories in international relations were framed as competing with Bayesianism: Bayesian updating was taken as the 'null hypothesis' or normative benchmark against which psychological hypotheses could be tested, and deviations from this benchmark counted as evidence in favor of psychological theories.[61] However, more recent work stresses the compatibility of psychological theories with Bayesianism.[62] Indeed, political scientists and economists have shown that Bayes' Theorem is able to accommodate a wide range of apparent 'deviations' from Bayesian updating.[63] This does not diminish the importance of the finding that individuals' inferences from costly signals vary for international relations theory: even if this finding is compatible with Bayesianism in principle, it nonetheless highlights the need to look beyond current rationalist theories

[57] Hall and Yarhi-Milo 2012.
[58] Kertzer, B. C. Rathbun, and N. S. Rathbun 2020.
[59] Yarhi-Milo, Kertzer, and Renshon 2018.
[60] For example, Yarhi-Milo 2014.
[61] For example, Kaufmann 1994. Also see Stein 2017. In adopting this framing, international relations scholars took their lead from behavioral economics - see chapter 3.
[62] For example, Kertzer, B. C. Rathbun, and N. S. Rathbun 2020.
[63] For example, Bullock 2009; Dixit and Weibull 2007.

and develop recipient-driven theories of signaling.[64]

Existing research broadly offers two approaches to explaining variation in the inferences draw from costly signals and developing such theories: 'consistency theories' and 'personalized theories'. Building on insights from psychology, consistency theories emphasize the role of prior beliefs in shaping responses to costly signals. They argue that individuals seek to reduce or avoid cognitive dissonance.[65] One important way individuals do so is by attending to information that is consistent with their prior beliefs, but discrediting or minimizing information that is not.[66] According to consistency theories, the key distinction in signaling, therefore, is not between costly and costless signals, but between belief-consistent and belief-inconsistent signals: quite apart from its cost, individuals will consider a signal to be more credible if it is consistent with their prior beliefs than if it is not.[67] For example, Kertzer and colleagues argue that the extent to which individuals revise their beliefs about a state's motives in response to costly signals of reassurance depends on whether they hold more cooperative or competitive views of international politics – individuals who hold cooperative views, they argue, are more responsive to costly signals of reassurance than individuals who hold competitive views.[68]

Consistency theories are good at explaining both between-subjects and within-subjects variation in the inferences individuals draw from costly signals. Like constructivist theories, they focus on meaning and how meaning shapes the reception of cues in international politics.[69] Based on variation in meaning, they can explain why responses to costly signals may systematically differ across recipients, depending on recipients' beliefs, and why the same recipient may respond differently to costly signals, depending on the signal's content. At the same time, unlike constructivist theories, particularly systemic constructivism of the 1990s,[70] consistency theories highlight variation in meaning across individuals rather than state-state dyads. Therefore, they are able to explain variation in responses to costly

[64] Yarhi-Milo, Kertzer, and Renshon 2018; Yarhi-Milo 2014.
[65] Festinger 1957.
[66] insert consistency theory refs. The extent to which consistency theories are incompatible with Bayesian updating is debated. For example, Kertzer, B. C. Rathbun, and N. S. Rathbun 2020 compellingly argue that asymmetric updating is, in principle, compatible with a Bayesian framework. Also see Fey and Ramsay 2006.
[67] Holsti 1962; Jervis 1976; Rosati 1988; P. Tetlock 1999; Vertzberger 1990, For example, For more recent examples, see Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Renshon 2008; Yarhi-Milo 2014.
[68] Kertzer, B. C. Rathbun, and N. S. Rathbun 2020.
[69] That said, constructivism's notion of 'meaning' is more encompassing than consistency theories' notion of 'beliefs'.
[70] Wendt 1999.

signals across recipients within a single state.

However, their ability to do so comes at a cost: consistency theories emphasize belief stability rather than change.[71] To be sure, they do not altogether neglect the possibility of belief change: consistency theories do document instances of belief change and, typically, seek to explain them by pointing to external shocks,[72] i.e. dramatic events that cannot be reconciled with an individual's prior beliefs, forcing the individual to reconsider their beliefs in order to reduce dissonance.[73] Yet, the resulting explanations tend to be ad-hoc and descriptive, leaving the mechanisms that drive belief change undertheorized.[74] As a result, they raise additional questions. Notably, why do external shocks not uniformly result in belief change, leading to belief convergence?[75]

Personalized theories offer an alternative approach to explaining variation in the inferences individuals draw from costly signals that is able to address this question. Like consistency theories, personalized theories build on insights from psychology to understand why individuals' responses to external cues in international politics vary. However, they focus on individual differences in cognitive style or beliefs about what constitutes a credible signal in interstate interactions rather than beliefs about a particular state, international politics in general, or both. For example, scholars have argued that individuals vary in how responsive they are to novel information, including costly signals, depending on their 'need for cognitive closure', i.e. their '[...] desire for a firm answer to a question and an aversion toward ambiguity':[76] individuals with a low need for closure are more responsive to novel information, and likely to experience belief change, than individuals with a high need for closure.[77] Alternatively, Keren Yarhi-Milo has argued that belief updating about motives varies across individuals depending on what they subjectively believe to be a credible indicator of a state's motives.[78]

Personalized theories are good at explaining between-subjects variation in the infer-

[71] Renshon 2008. In this regard, consistency theories can be considered closer to psychological than constructivist approaches in international relations: they take meaning as given, leaving the link between subjectively held beliefs and the social processes of meaning construction unclear.
[72] For example, Renshon 2008.
[73] Festinger 1957.
[74] Lebow and Stein 2019.
[75] Also see chapter 5.
[76] Kruglanski and Webster 1996, p. 264.
[77] For example, Aronoff 2014; B. C. Rathbun 2019.
[78] Yarhi-Milo 2013; Yarhi-Milo 2014. Alongside belief consistency, this forms one of several mechanisms that underpin Yarhi-Milo's 'subjective credibility hypothesis'.

ences individuals draw from costly signals, as well as variation in belief change about a state's motives. Unlike consistency theories, which emphasize the role prior beliefs play in belief updating, they focus on the process of belief updating itself and, particularly, the cognitive processes that underpin belief updating. Based on between-subjects variation in these processes, they can explain variation in individuals' responsiveness to costly signals independently of their prior beliefs and, thus, variation in individuals' propensity toward belief stability or change. In other words, personalized theories offer an account of why belief change may occur unevenly. In doing so, they add considerable explanatory power to actor-specific theories of signaling. However, personalized theories are less good at explaining why individuals' responses to costly signals vary *dynamically*, that is, why the timing of belief change, not merely the propensity toward belief change, systematically varies across individuals.[79]

To address this puzzle, I return to the initial problem that motivates theoretical and empirical research on signaling: the problem of inferring a state's unobservable motives from its behavior, which requires 'the separation of internal from situational components' of that state's behavior, as Jervis aptly puts it.[80] Although costly signals appear to solve this problem, recipients must relate a state's behavior to its international environment in order to identify the cost, and thus credibility, of a signal in the first place.[81] In other words, the costly signaling hypothesis presumes that recipients distinguish 'internal from situational components' of a state's behavior; it does not explain how they do.[82] A key claim I make in this thesis is that we can make headway in understanding why individuals vary in the inferences they draw from costly signals by addressing this gap and exploring how individuals *integrate* the distinct cues that constitute costly signals. I take research on attribution – that is, research on how individuals ascribe behavior to underlying, internal or situational, causes – in psychology as a starting point for doing so.

[79] An important exception to this, and to the distinction I make between consistency theories and personalized theories, is Yarhi-Milo, Kertzer, and Renshon 2018 who link foreign policy attitudes to differences in the *type* of costly signal individuals respond to, showing that foreign policy 'hawks' are more responsive to sunk-cost signals and foreign policy 'doves' are more responsive to tying-hands signals.

[80] Jervis 1976, p. 48.

[81] This is implied by rationalist theories, for example, Glaser 2010. I elaborate this point in chapters 4 and 5.

[82] Jervis 1976, p. 48.

Theoretical Approach	**Theoretical Predictions**		**Explanatory Scope**		
	Attribution to state's		Variation in the attribution of costly signals		Belief change
	...motives	...strategic environment	Between-subjects	Within-subjects	
Rationalist theories	Costly signals	Costless signals	No	No	Yes
Consistency theories	Belief-consistent signals	Belief-inconsistent signals	Yes (belief-dependent)	Yes (content-dependent)	No
Personalized theories	Depends on individual cognitive style or preferred indicator		Yes	No	Yes
Theory of attribution	??	??	Yes	Yes	Yes

Table 2.1: Theories of motives and the puzzle of attribution.

Please do not circulate or cite.

2.4 A Theory of Attribution for International Relations

The idea that attribution research in psychology can help scholars understand the microfoundations of signaling in international politics is not new. In the 1970s already, international relations scholars began to draw on attribution research to understand conflict and cooperation among states. Famously, Robert Jervis drew on attribution research in his seminal contribution to scholarship on signaling, *Perception and Misperception in International Politics*.[83] Around the same time, Shawn Rosenberg and Gary Wolfsfeld published an article on the role of attribution in international conflicts, notably the Arab-Israeli conflict.[84] The 1990s saw the publication of two further books that built on attribution research and would become canonical in scholarship on signaling: Deborah Welch Larson's *Anatomy of Mistrust*, which analyzed U.S.-Soviet relations from a psychological vantage point, and Jonathan Mercer's *Reputation and International Politics*, which provides scholars with a theory of reputation formation.[85]

The main thrust of this work has been that attributional processes undermine effective signaling and entrench actors' prior beliefs about a state.[86] In conflictual state dyads, scholars have argued that attributional processes lead actors to underestimate the role of situational factors, notably their own actions, in an adversary's choice to compete and to overestimate the role of internal factors, notably an adversary's malign motives. Conversely, they have argued that actors overestimate the role of situational factors and underestimate the role of internal factors when an adversary chooses to cooperate.[87] As a result, attribution is considered to feed into escalatory dynamics and perpetuate distrust, rendering international conflicts more intractable.[88] By contrast, in allied state dyads, scholars have argued that actors overestimate the role of internal, and underestimate the role of situational, factors in an ally's choice to cooperate, but do the opposite when an ally chooses to compete.[89] In either case, attributional processes interrupt signaling and work to uphold actors' beliefs about a state's motives in strategic interactions.

[83] Jervis 1976.
[84] Rosenberg and Wolfsfeld 1977.
[85] Larson 1997a; Mercer 1996.
[86] For example, Lebow and Stein 2019.
[87] Jervis 1976, particularly chapters 8 and 9; chapter 1 in Larson 1997a; Rosenberg and Wolfsfeld 1977.
[88] Larson 1997a; Rosenberg and Wolfsfeld 1977.
[89] Rosenberg and Wolfsfeld 1977. Mercer 1996 makes a similar argument in relation to inferences about resolve rather than motives.

Applying these arguments about the role of attribution in signaling to a range of historical case studies, notably U.S.-Soviet relations throughout the Cold War,[90] as well as contemporary conflicts in international politics,[91] scholars have shed light on the microfoundations of signaling and, in particular, security dilemmas, where two non-expansionist states find themselves locked into inefficient arms races.[92] However, the way in which scholars have applied these findings to interstate interactions also raises several questions.

First, the question of how findings from attribution research in interpersonal and intergroup contexts extend to interstate interactions remains underdeveloped. International relations scholars who draw on attribution research typically base their arguments on three canonical findings from psychology: the tendency of individuals to attribute a person's behavior to her dispositions, and underestimate the role of situational factors, even when those situational factors are salient – a phenomenon known as the 'fundamental attribution error' or 'correspondence bias';[93] the tendency of individuals to attribute another person's behavior to her dispositions, but to attribute their own behavior to situational factors – a phenomenon psychologists call the 'actor-observer bias';[94] the tendency of individuals to attribute an outgroup member's negative behavior to dispositions more, and an outgroup member's positive behavior to dispositions less, than an ingroup member's behavior negative or positive behavior, respectively – a phenomenon known as the 'ultimate attribution error'.[95] All three findings pertain to the behavior of individuals, not states.[96] By applying them more or less directly to interstate interactions, scholars have tacitly assumed that attributional processes are uniform across targets.[97] However, research in

[90] Larson 1997a.
[91] Rosenberg and Wolfsfeld 1977.
[92] Inefficient in the sense that both states could attain their primary goal – security – more cheaply if they could trust each other (see Glaser 2010).
[93] Ross 1977; Jones and Harris 1967; Gilbert and Malone 1995. For example, Jervis 1976; Larson 1997a; Lebow and Stein 2019.
[94] Edward and Nisbett 1971. For example, Jervis 1976; Larson 1997a and, more recently, Dickson 2009.
[95] Taylor and Jaggi 1974; Pettigrew 1979; Rosenberg and Wolfsfeld 1977. In developing their predictions, Rosenberg and Wolfsfeld 1977 combine insights from attribution research with the cognitive consistency literature in psychology (e.g. Festinger 1957.) The evidence for the ultimate attribution error is mixed (see the review article by Hewstone 1990). In particular, although individuals make more dispositional attributions for positive than negative ingroup behavior, and more dispositional attributions for negative outgroup than ingroup behavior, Hewstone 1990 does not find evidence to support the idea that individuals make more dispositional attributions for negative than positive outgroup behavior – an idea that is commonly cited by international relations scholars, e.g. Larson 1997a, pp. 23-24, and Mercer 1996.
[96] Hewstone 1990, pp. 312-313.
[97] Mercer 1996 has theorized variation in attributional processes across targets depending on their relationship to the recipient (i.e. whether they are an ally or an adversary). However, that the findings from research on *individual* behavior extend to *state* behavior is assumed.

social psychology finds that attribution and impression formation differ for individual and group targets.[98] This is unsurprising if we consider the vital role that mental representation plays in information processing (see chapter 3),[99] yet highlights a critical gap in our understanding of attribution in international politics.

Second, the question of how findings from attribution research integrate with scholarship on signaling, and costly signaling in particular, remains undertheorized. On the one hand, scholars have argued that attributional biases undermine costly signaling, invoking attribution to explain security dilemmas.[100] On the other hand, scholars have resorted to costly signals to explain change in adversarial relationships. For example, citing the fundamental attribution error, Larson argues that elite decision makers will tend to '[...] assume that the enemy's conciliatory actions were forced by external pressures [...]'.[101] However, she also maintains that '[p]olicymakers are more likely to infer that another state's concession is a sincere sign of a desire to improve relations if the concession is relatively costly [...]',[102] noting that '[t]he number of concessions that a state must make to elicit a cooperative response depends on their prior beliefs of the target'.[103] Intuitively, this argument makes sense – afterall, even in adversarial relationships, varying degrees of cooperation are possible, and sometimes conflictual relationships in international relations undergo a lasting transformation. Clearly, attributions can vary and beliefs about motives can change altogether. Yet, Larson's arguments raise the question of whether attributions are merely epiphenomenal, updating in a Bayesian manner along with beliefs about the target, or whether attributional processes are, in fact, causally implicated in the costly signaling mechanism, as advocates of psychological approaches seem to suggest. If they are, how so?

Finally, international relations scholars tacitly assume, not only that attributional processes are uniform across targets, but also that they are uniform across recipients. However, a large literature in psychology has established that recipients systematically vary in the attributions they make. For example, in 1984, Joan Miller conducted a cross-cultural study of attribution, comparing American and Hindu participants' attributions of

[98] Hamilton and Sherman 1996; Hamilton, Sherman, and Castelli 2002.
[99] Also see Morris, Nisbett, and Peng 1996; Moscovici 1988.
[100] Larson 1997a; Larson 1997b; Mercer 1996.
[101] Larson 1997a, p. 24.
[102] Larson 1997a, p. 27.
[103] Larson 1997a, p. 32.

two positive and two negative behaviors, respectively.[104] She found that American adults made more dispositional attributions than Hindu adults, and American adults made more dispositional attributions than American children, suggesting that cultural influences play an important role in attribution.[105]

Since then, several research programs within psychology have investigated variation in attribution further, providing distinct explanations of why it occurs. For example, in cross-cultural and cultural psychology, scholars have argued that variation in attribution arises from differences in the extent to which individuals hold 'individualist' versus 'collectivist' social orientations, engage in 'analytic' versus 'holistic' thinking, or hold an 'independent' versus 'interdependent' conception of the self.[106] Meanwhile, in social and developmental psychology, scholars have identified 'lay' or 'implicit' theories – that is, people's everyday theories of the social world – as an important source of variation in attribution.[107]

What all these explanations share is an emphasis on the role of representation and socially-acquired meaning in understanding why attribution varies across individuals.[108] Taking this insight as a starting point, this thesis sets out to develop a domain-specific theory of attribution that incorporates the target of attribution in international politics, namely states, as well as recipient-driven variation in attribution into its theory, while systematically integrating findings from attribution research with scholarship on signaling. In doing so, it provides scholars with a theoretical framework for understanding why individuals' inferences about states' motives dynamically vary in response to costly signals.

[104] Miller 1984.
[105] Miller 1984.
[106] For the link between individualism-collectivism and attribution, see Choi and Nisbett 1998; Masuda and Kitayama 2004; Morris and Peng 1994; Norenzayan, Choi, and Nisbett 2002. See Krull et al. 1999 for countervailing evidence. For studies on analytical/holistic thinking and attribution, see Nisbett et al. 2001 and chapter 6 in Nisbett 2003. For the link between self-construal and attribution, see Markus and Kitayama 1991. In either case, the empirical evidence cited in support of these explanations tends to be based on studies that observe, rather than randomize, cultural differences (for a discussion of this point, see Varnum et al. 2010). A rare exception to this is Kuehnen et al. 2013, which directly manipulates self-construal, showing that self-construal in causally implicated in attribution.
[107] For a brief introduction to lay/implicit theories, see Ramirez et al. 2013; Plaks, Levy, and Dweck 2009. For findings on the link between implicit theories and attribution, see Dweck, Hong, and Chiu 1993.
[108] For example, Moscovici 1984. Also see Morris, Nisbett, and Peng 1996.

3 Of Two Minds? Rationalism and Psychology in International Relations

Abstract

Proponents of psychology have heralded the recent surge in psychological research in International Relations as a 'behavioral revolution'. This notion is founded upon three common, if tacit, beliefs in the discipline: first, the belief that rational choice treats the mind as a black box; second, the belief that present-day psychology contrasts with twentieth-century behaviorism; third, the belief that psychology supplies a data-driven account of the mind. All three beliefs are overstatements that obscure more than they illuminate rational choice, psychology, and their relationship to one another. A more accurate understanding of rational choice and psychology reveals that psychology only competes with a subset of rational choice applications and, where it does, it shares with rational choice a conception of the mind as computer. The introduction of psychology into International Relations is thus better understood as a reform, not a revolution. Moreover, psychology's reliance on a conception of the mind as computer limits the ability of international relations scholars to pursue a core element of this reform: to theorize systematic variation in information processing. To fulfill its vision, psychological research in International Relations must look beyond mainstream psychology to develop a conception of the mind better suited to understanding how minds create and acquire meaning.

Books and Articles Discussed in This Review Essay

Joshua D. Kertzer, Brian C. Rathbun and Nina Srinivasan Rathbun. 2020. The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations. *International Organization, 74*(1), 95-118.

Brian C. Rathbun. 2019. *Reasoning of State: Realists, Romantics and Rationality in International Relations.* Cambridge: Cambridge University Press.

Keren Yarhi-Milo. 2018. *Who Fights for Reputation: The Psychology of Leaders in International Conflict.* Princeton: Princeton University Press.

3.1 Introduction

Recent years have seen a surge in psychological research in International Relations. Combining experimental methods with in-depth historical analyses, this research is rapidly expanding our empirical knowledge of the individual preferences, beliefs and behaviors that inform decision making in international politics and, crucially, variation therein. Moreover, as a substitute for the abstract assumptions of rational choice and game-theoretic models, this knowledge allows scholars to address hitherto unresolved empirical puzzles in the field.

In light of these contributions, a recent special issue of *International Organization* proclaimed a 'behavioral revolution' in International Relations: like the cognitive revolution in psychology, psychological research in International Relations reappraises the role of the mind in causal theories of behavior and holds out the prospect of an *empirical* account of the mind.[1] Moreover, by raising novel questions about the micro-level dynamics that underpin macro-level outcomes in international politics, it fortifies calls for microfoundational research in International Relations more generally.

In this paper, I discuss three common, if tacit, beliefs that underpin the notion of a 'behavioral revolution' in International Relations: first, the belief that rational choice theory treats the mind as a black box; second, the belief that present-day psychology contrasts with twentieth-century behaviorism; third, the belief that present-day psychology supplies a data-driven account of the mind.

[1] Hafner-Burton et al. 2017; Kertzer and Tingley 2018.

Drawing on recent work in the history of economics, the history of psychology, political theory, and the philosophy of mind, I show how, individually, these beliefs are overstatements that obscure more than they illuminate rational choice and psychology.[2] Jointly, they buttress a flawed conception of rational choice theory and psychology's relationship to one another, and distort psychology's contributions to International Relations in three ways. First, the belief that rational choice theory treats the mind as a black box wrongly suggests that psychology reinserts the mind into International Relations.[3] Second, the belief that present-day psychology contrasts with behaviorism wrongly suggests that psychological research in International Relations is free of the vestiges of behaviorism. Finally, the belief that psychology provides a data-driven account of the mind obscures the theoretical assumptions that guide its inquiry into the mind.

Based on a more accurate understanding of rational choice, psychology, and their relationship to one another, I offer a critical reappraisal of psychology's contribution to International Relations. I advance three arguments. First, the introduction of psychology into International Relations is best understood as a reform, not a revolution: psychological theories compete with a subset of rational choice theories, not rational choice theory as a whole. Moreover, where they do compete with rational choice theories, they offer novel substantive claims about the mind; more fundamentally, however, they share with rational choice a conception of the mind as computer. Second, proponents of psychological research in International Relations champion their ability to theorize 'heterogeneity of preferences, beliefs, and decision-making processes'.[4] However, their continued reliance on a conception of the mind as computer severly limits this ability. In particular, psychological research to date can observe, but not explain, variation in preferences and beliefs. Third, to overcome these limitations, closer collaboration and integration of psychological research with constructivist approaches is needed. However, to succeed, such collaboration and integration must be grounded in a genuinely interdisciplinary dialogue that moves beyond, rather than sustains, the current dichotomy between social and cognitive processes.

The paper is divided into three sections. The first section revisits standard accounts

[2] Unless specified otherwise, 'psychology' is used to refer to those research programs within psychology – primarily cognitive and (American) social psychology – that are grounded in the cognitive revolution.
[3] Kertzer and Tingley 2018, for instance, state that 'the behavioral revolution in IR seeks to bring mental states back in' (Kertzer and Tingley 2018, p. 4).
[4] Hafner-Burton et al. 2017, S4.

of the history of economics and psychology, respectively, and shows how they generate the three beliefs that underpin the notion of a 'behavioral revolution' in International Relations. The second section addresses each belief in turn. It shows that rational choice theory never (quite) lost its mind; that psychology, far from being purely data-driven, relies on a particular conception of the mind that is broadly behavioral; and that this conception reveals a hitherto unexamined, yet important, commonality between rational choice theory and psychology. The final section considers the implications of this revised understanding of rational choice theory, psychology, and their relationship to one another for the emerging subfield of behavioral International Relations. It discusses three significant contributions to this subfield to substantiate its arguments.

3.2 A Tale of Two Minds

The late mathematician and computer scientist John von Neumann is well-known to political scientists and economists alike: in collaboration with Oskar Morgenstern, he proved the minimax theorem and developed an axiomatic version of the expected utility principle.[5] Von Neumann and Morgenstern's seminal contribution, *Theory of Games and Economic Behavior*, laid the foundation for decision theory under uncertainty and shaped the social sciences for decades to come.[6]

Less known to political scientists and economists is von Neumann's legacy to the cognitive sciences: in a series of lectures on *The Computer and the Brain*, written shortly before his death and published posthumously in 1958, von Neumann outlined his ideas on the nervous system.[7] These lectures were emblematic of the cross-disciplinary dialogues, seminars, and conferences that took place throughout the 1930s, 1940s, and 1950s, and to which von Neumann was an avid contributor. Soon, these cross-disciplinary efforts would culminate in the 'mind's new science' – the approach to studying human action that has dominated cognitive and social psychology since.[8]

It is not by chance that political scientists and economists only selectively remember von Neumann's legacy: the standard narrative in the social sciences understands

[5] Weintraub 2002.
[6] Neumann and Morgenstern 1944.
[7] Neumann 1958.
[8] Borrowed from Gardner 1985.

rational choice theory as part of the antipsychological movement of the early twentieth century.[9] On this view, von Neumann and Morgenstern's contribution is the culmination of economics' '45-year long "escape from psychology"'[10] rather than a common point of departure for economics and psychology.

Of course, rational choice theory's core assumption, the assumption of utility maximization, is much older: it dates back at least to the nineteenth century, the heyday of marginal utility theory in economics. In an attempt to emulate mechanical physics, proponents of marginal utility theory postulated that individuals act purposively so as to maximize their utility, i.e. the amount of pleasure they experience in consequence of their actions.[11] From this postulate, marginal utility theorists sought to deduce the laws that govern economic behavior, most famously, the law of diminishing marginal utility.[12]

However, according to the standard narrative, rational choice theory underwent a profound transformation in the first half of the twentieth century when economists came under the influence of positivism and, later, logical positivism.[13] A key tenet of positivism is empirical verificationism: the idea that knowledge claims are true only insofar as they can be empirically verified. Its adoption in economics led to a change in what counted as 'scientific' research. Although marginal utility theorists assumed utility to be measurable in the same way that, say, mass or distance are (i.e. a unit-based understanding of measurement),[14] in practice, they lacked a strategy to measure utility (other than introspection).[15] Since introspection relied on psychological processes which themselves were unobservable, positivism discredited it as a method of empirical investigation, thereby discrediting utility as a basis for economic theory. Therefore, economists proceeded to rid rational choice theory of its unobservable, psychological underpinnings – or so the standard narrative tells us. They did so in several progressive steps.

First, during the so-called ordinal revolution, economists replaced the notion of mea-

[9] For an example of this in International Relations, see Mercer 2005. For an example in economics, see Bruni and Sugden 2007.
[10] Giocoli 2005, p. 186
[11] Giocoli 2003, p. 46. Though note differences in marginal utility theorists' precise conception of 'utility', see Moscati 2019, pp. 63-64.
[12] Moscati 2019, pp. 26-27.
[13] Giocoli 2003, p. 43.
[14] In economic history, this conception of utility is often referred to as 'cardinal utility'. For a critical discussion of this practice see Moscati 2019. To avoid confusion, I reserve the term 'cardinal utility' for the conception of measurable utility that emerged in the twentieth century.
[15] Giocoli 2003, pp. 46-48; Hands 2010, p. 635

surable utility with that of ordered preferences ('ordinal utility') – a process that began with Vilfredo Pareto in the early 1900s and was concluded by John Hicks and Roy Allen in the 1930s.[16] Preferences, though mental, were a step removed from the sensations, like pain or pleasure, that gave rise to them. Moreover, preferences could be rank ordered rather than measured. Of course, to rank order preferences was to forgo the theoretical advantages of knowing *how much* more an individual prefers option a to option b, or *how much* more utility an individual derives from option a compared to another individual.[17] Precisely because preference orderings are less information-rich than measurable utility, however, they are also less empirically demanding. Thus, by demonstrating that the most important results of economic theory could be derived solely from ordered preferences, ordinalists placed the study of economics on a more empirically verifiable, and thus scientific, footing – or so they hoped.[18]

However, like marginal utility theory, ordinalism relied on an unobservable, mental referent: preferences. So long as an agent's preferences (explanans) could not be verified independently of their choices (explanandum), any explanation of the latter based on the former was necessarily tautological and at odds with a positivist understanding of science.[19]

Therefore, at the height of the antipsychological movement in economics in the late 1930s, Paul Samuelson proposed to take rational choice theory's emancipation from psychology one step further and to dispense with the concept of utility altogether. He did so by redefining rationality as *consistency* rather than utility maximization.[20] While, mathematically, the former was equivalent to the latter, substantively, it was not: for Samuelson, 'preferences' no longer referred to an agent's mental states; it merely *described* their choices in the market formally. So long as the resulting description ('preference relation') fulfilled the so-called 'weak axiom of revealed preference', i.e. the agent's choices were consistent across choice options, Samuelson showed that the most important results of economic theory could be derived solely from observable market behavior without knowledge of agents' underlying preferences.[21]

[16] For a critical take on the role typically ascribed to Pareto, see Bruni and Guala 2001.
[17] See Rosenberg 2008 for a discussion of implications Rosenberg 2008, pp. 86-87.
[18] Giocoli 2003, pp. 67-71; Moscati 2019, pp. 79-94.
[19] Giocoli 2003, pp. 99-101.
[20] Giocoli 2005, p. 181.
[21] Giocoli 2005, pp. 186-188.

Samuelson's shift from rationality as maximization to rationality as consistency marked the 'fulfillment of [economics'] 'escape from psychology": the concept of rationality had been emptied of its psychological content and reduced to a purely formal description of market behavior.[22] However, like Pareto, Hicks and Allen, Samuelson focused on riskless decisions. By contrast, von Neumann and Morgenstern offered a theory of decision making under uncertainty. Crucially, the introduction of probabilities over choice options enabled them to resurrect the notion of measurable utility while retaining Samuelson's purely formal conception of rationality. Von Neumann and Morgenstern's theory would become the basis for modern decision theory.

To understand von Neumann and Morgenstern's decision theory – an axiomatic version of Bernoulli's nineteenth-century expected utility theory – it is helpful to understand their theory of measurement. Von Neumann and Morgenstern understood measurement as the possibility of assigning unique numbers to objects, here, choice options. This requirement was fulfilled if the assigned numbers remained unique up to a linear transformation ('cardinal utility').[23] Ordered preferences ($a > b$; $b > c$; $a > c$) alone were insufficient to restrict the set of possible transformations, i.e. derive unique utility numbers. However, von Neumann and Morgenstern argued that they could further restrict the set of transformations by identifying a natural operation of addition – a feat they accomplished by combining preference relations with probabilities over outcomes (that add up to 1).[24] Specifically, they showed that, given a consistent set of choices and probabilities over outcomes, it was possible to derive cardinal utilities from observable choices. In other words, von Neumann and Morgenstern were able to 'measure' utility without reverting to the introspective notions that had marred marginal utility theory.

Importantly, von Neumann and Morgenstern's conclusions assumed an objective notion of probability: probabilities over choice options were understood as 'long run frequencies' that are known.[25] By contrast, Savage developed a version of expected utility theory, 'subjective expected utility theory', in which probabilities were assumed to be subjective: they were defined as 'the confidence that a particular individual has in the truth of a

[22] Giocoli 2005, p. 189.
[23] Moscati 2019, pp. 153-154.
[24] Giocoli 2005, pp. 193-196.
[25] Giocoli 2005, p. 196.

particular proposition', where the true probability of the proposition is unknown.[26] Moreover, Savage defined a rational agent 'as someone who checks whether her beliefs satisfy the axioms [of expected utility theory] and, in the negative case, revises them according to Bayes's rule.'[27] In other words, he extended the consistency requirement of expected utility theory to probabilistic beliefs. Savage thus concluded the progression in economics from a conception of rational agents as utility-maximizing *homines economici* to that of rational agents as *Bayesian statisticians*.[28] Based on this conception, he showed that it was possible to derive *both* cardinal utilities *and* subjective probabilities from observable market behavior without recourse to mental concepts.

The standard narrative of economic theory, then, is one of escape: it emphasizes that, in response to logical positivism, economists hollowed out rational choice theory in favor of a purely formal, mathematical conception of rationality, devoid of any psychological content. Thereby, it gives rise to the belief that rational choice theory treats the mind as a black box, and helps sustain an understanding of rational choice theory and psychology as fundamentally separate and distinct approaches to the study of human behavior. This understanding is reinforced by standard accounts of the history of psychology.

Under the joint influence of logical positivism and American pragmatism, psychology, too, underwent important changes in the early twentieth century: like economists, psychologists increasingly sought to ground the scientific study of human action in observable, and thus objectively verifiable, data. To this end, key figures – among them, J. B. Watson and B. F. Skinner – developed an approach known as 'behaviorism'.[29] Whereas previously psychology included the study of unobservable mental events such as consciousness, memory, thoughts and feelings, behaviorism restricted psychology to the study of observable behavior.[30] Indeed, adopting a materialistic view, Watsonian behaviorism went as far as to deny the independent existence of mental events and to redefine formerly mental concepts in behavioral terms.[31] Observable behavior, in turn, was to be explained and predicted by law-like causal relationships. Adapting Bekhterev and Pavlov's notion of conditioned reflexes, these law-like relationships were generally expressed as stimulus-response (S-R)

[26] Savage in Giocoli 2013, p. 79
[27] Giocoli 2013, pp. 80-81.
[28] Giocoli 2013.
[29] Sometimes also referred to as S-R behaviorism (Moore 1996).
[30] Note that behavior was understood to include verbal utterances.
[31] Farrell 2014, p. 392.

models.[32] S-R models captured the basic idea that any behavioral response could be explained by a set of antecedent environmental stimuli.[33] The mind was thus no longer needed; it was either non-existent or figured as a mere black box.

While the reign of behaviorism in American psychology was near-complete, standard accounts of the discipline's history emphasize that it was relatively short-lived: by mid-century, psychologists had started to reconsider the role of the mind – or lack thereof – in their causal accounts of behavior. Behaviorism championed the S-R model. Yet, how could individuals identify a stimulus as such, and select from among the many stimuli their environment presents them with, if not for some mental construct or mediating process?[34] These and other questions ultimately led psychologists to readmit mental constructs and processes into the empirical study of behavior. Widely considered to be a watershed moment in the history of the discipline, this development subsequently became known as the 'cognitive revolution' of psychology.[35]

Crucially, this account of the history of psychology gives rise to the widely-held belief that present-day psychology, which provides an account of the mind in addition to behavior, contrasts with behaviorism. At the same time, cognitive psychology's ascendance into the ranks of the empirical sciences has led to the belief that, unlike its nineteenth-century precursor, it supplies data-driven, descriptively accurate models of the mind.[36]

Jointly, the standard narratives of economics and psychology paint a picture wherein rational choice theory resembles behaviorism: like behaviorism, rational choice theory treats the mind as an impenetrable black box and denies its causal power.[37] By contrast, present-day psychology is thought to resemble neither, for it readmitted mental states and processes into the empirical study of human behavior. Thus emerges, implicitly or explicitly, the promise of psychology to International Relations: like the cognitive sciences revolutionized psychology over fifty years ago, so psychology will revolutionize International Relations alongside economics today. Or will it?

---

[32] Staddon 2014, pp. 15-34.
[33] Though key figures of behaviorism differed in their precise understanding of the stimulus-response model, a fact often overlooked by standards accounts of psychology – see O'Neil 1995.
[34] For a discussion of these problems, see Rosenberg 2008, pp. 76-80.
[35] Miller 2003.
[36] For example, Mercer 2005.
[37] Mercer 2005, p. 84.

3.3 Mind Metaphors Matter

3.3.1 Rational Choice Theory: The Mind as a Black Box?

The standard narrative of economics gives rise to the belief, common among international relations scholars, that rational choice theory treats the mind as a black box: like behaviorism in psychology, rational choice theory explains behavior without reference to the mind. This view, though not entirely mistaken, is misleading. To be sure, rational choice theory did undergo profound changes in response to positivism as it sought to shed economic theory of its unobservable, metaphysical content. However, the resulting changes were more ambiguous than the standard narrative suggests: until today, the role of the mind in rational choice explanations remains contested among economists, political theorists and philosophers. We can distinguish broadly between three interpretations of rational choice theory: a behavioral interpretation, a mental interpretation and a third interpretation that is neither behavioral nor mental.[38]

A behavioral interpretation of rational choice theory contends that the preference relations, utility functions and subjective probabilities invoked by analysts are 'mere redescriptions of behavioural patterns and perhaps instrumentally useful constructs for organizing and making sense of empirical regularities'.[39] Their use is purely formal and does not imply the existence of mental states. In other words, proponents of this interpretation subscribe to what Dietrich and List call the *'ontological 'revealed preference' thesis'*: the idea that 'the ontological commitments of any theory in economics...are restricted to agents' choices and choice-behavioral patterns and therefore exclude mental states'.[40]

By contrast, a mental interpretation holds that preference relations, utility functions, subjective probabilities or all three are 'representations of mental or psychological phenomena', i.e. they correspond to real mental states.[41] Proponents of this interpretation ground their view in a functionalist conception of mental states: 'mental states,' they

[38] The first and second interpretation roughly correspond to MacDonald's instrumental-empiricist and scientific realist view of rational choice, respectively (MacDonald 2003). For a discussion of a fourth, less common interpretation of rational choice theory as a heuristic model in a Weberian sense see Barkin 2015.
[39] Dietrich and List 2016a, p. 250; also see Okasha 2016, pp. 415-420.
[40] Dietrich and List 2016a, p. 266.
[41] Dietrich and List 2016a, p. 250.

contend, 'are states that play a certain role for an agent'.[42] In other words, preference relations, utility functions and subjective probabilities can be taken to be real because they play the role of preferences and beliefs in causal accounts of behavior. Importantly, this conclusion holds whether or not the specific assumptions made about mental states are, in fact, realistic. Proponents of a mental interpretation, then, endorse the *'epistemological 'revealed preference' thesis'* – the idea that the 'evidence for any theory in economics is restricted to agents' choice behaviour' – but not its ontological counterpart.[43]

In a recent critique of mentalism, Guala advances a third interpretation of rational choice theory. To reject a behavioral interpretation is not to accept a mental one, he argues, for preferences – understood as an agent's propensity to behave in a particular way, given a specific set of circumstances – are *multiply realizable*: they may arise from distinct causal processes, not all of which are psychological. Particularly where aggregate units are concerned, preferences may be the product of non-psychological processes such as voting. On the one hand, then, preferences, in combination with beliefs, *explain* behavior and therefore are irreducibly distinct from it. On the other hand, preferences generalize across causal bases, which themselves may or may not be psychological. Consequently, preferences are not necessarily indicative of mental states in the sense used by Dietrich and List.[44] Guala's argument is especially pertinent to International Relations where the units under consideration tend to be states or groups – a point I return to in the final section of this paper.

The standard narrative of economics promotes a behavioral interpretation of rational choice theory, and thereby sustains the belief that rational choice theory treats the mind as a black box. Unsurprisingly, it is the interpretation most commonly adopted by economists. However, two things are worth noting. First, with von Neumann and Morgenstern's axiomatization of expected utility theory, rational choice theory became increasingly interpreted as a normative rather than a positive theory of behavior.[45] Second, in the original project pursued by Samuelson, the shift from rationality as utility maximization to rationality as consistency implied the abandonment of any attempt at explaining individual behavior (as opposed to aggregate, market-level phenomena). Both

[42] Dietrich and List 2016a, p. 268.
[43] Dietrich and List 2016a, p. 266.
[44] Guala 2019.
[45] Herfeld 2018.

of these shifts set rational choice theory apart from behaviorism in psychology, which sought to explain and predict individual behavior.

Meanwhile, political theorists and philosophers often adopt a mental interpretation of rational choice theory.[46] A prominent example is the philosopher Alexander Rosenberg, who maintains that expected utility theory is essentially 'folk psychology formalized to a very high degree':[47] like folk psychology, rational choice theory explains behavior based on preferences and beliefs, however, it imposes a particular assumption – the axiom of utility maximization – on their relationship to one another. As William Riker, a prominent advocate of rational choice theory in political theory, has argued, this allows analysts to 'work back and forth among outcomes, actions, and intentions to discover the goals people actually have'.[48] In other words, through a process of repeated prediction and revision of original assumptions, rational choice models enable scholars to approximate the unobservable mental states that underlie behavior.[49]

Crucially then, a mental interpretation implies that rational choice theory does *not* treat the mind as a black box, but relies on mental states to explain behavior. Consequently, it differs from behaviorism in psychology. Indeed, for Riker, this difference is a key advantage of rational choice theory: 'it is sometimes said that behaviorism as adapted from psychology can serve as a political theory. But it is just here that behaviorism fails,' he contends, '...the behaviorist cannot do much about revising the generalization because the method forecloses the examination of...intentions.'[50]

Unlike their counterparts in other fields, international relations scholars are rarely explicit about the assumptions that accompany their use of rational choice theory. The extent to which they subscribe to a behavioral versus mental interpretation is therefore unclear. Generally however, rational choice is understood, not as a unified substantive theory, but 'as a methodological approach that explains both individual and collective (social) outcomes in terms of *individual goal-seeking under constraints*...the focus on goal

[46] Satz and Ferejohn 1994, p. 71; Okasha 2016, p. 410
[47] Rosenberg 2008, p. 89.
[48] Riker 1995, p. 26.
[49] Importantly, this does not imply that all behavioral patterns are explainable by rational choice theory. As Dietrich and List 2016b note, there may be behavioral patterns for which no underlying preference relations can be found. Consequently, rational choice theory is falsifiable.
[50] Riker 1995, p. 26. It is interesting to note that Riker has been explicitly cited by international relations scholars who emphasize commonalities between behaviorism and rational choice theory, e.g. Mercer 2005, p. 85.

seeking presumes that explanation should proceed in terms of relevant actors, the goals they seek, and their ability to do so.'[51] As in other fields, then, any interpretation of rational choice theory necessarily hinges on the theoretical and substantive assumptions made in particular contexts.[52]

Fortunately, for the purposes of this paper, we need not settle the debate about behaviorism versus mentalism in International Relations in general; it is entirely sufficient to argue that applications of rational choice theory to questions addressed by psychological research imply mentalism. Adopting a functionalist conception of mental states, this is a relatively straightforward case to make: psychological theories of International Relations generally seeks to *explain* either the behavior or the beliefs of *individuals*. Where rational choice theory explains the behavior of individuals, preference relations and subjective probabilities play the role of preferences and beliefs in causal accounts of behavior. Moreover, because we can assume individual's preferences to be generated by psychological processes, Guala's objections to a mental interpretation of preferences do not apply. Preference relations and subjective probabilities can thus be taken to correspond to real mental states. Where rational choice theory explains the beliefs of individuals, it presupposes the existence of the mental states in question. Moreover, Bayes' rule plays the role of a mental process in causal accounts of beliefs. In other words, both can be taken to correspond to mental phenomena.

The belief that rational choice theory treats the mind as a black box and explains behavior without reference to the mind, then, is misleading: rational choice theorists are not of one mind when it comes to behaviorism versus mentalism, and the applications most relevant to debates about psychological theories of International Relations evidently have a mind of their own.

3.3.2 Psychology: The Mind as Computer

Standard accounts of psychology portray the readmission of mental states and processes into the scientific study of behavior as a watershed moment in the history of the discipline. Most international relations scholars have uncritically imported this narrative, along with the belief that present-day psychology contrasts with behaviorism. At the same time,

[51] Snidal 2013, p. 87.
[52] Herfeld 2020.

they tend to emphasize that psychology, unlike rational choice theory, offers an *empirical* account of the mind.[53] While the latter dovetails with psychology's own claim to being a 'hard' science, a discussion of how psychology squares this claim with the study of unobservable mental states and processes remains crucially absent in International Relations. Addressing this gap reveals that, far from being purely data-driven, the 'mind's new science' relies on a particular conception of the mind – a conception that is 'broadly behavioral' and, importantly, informs psychology's substantive research focus.[54]

Standard accounts emphasize the cognitive revolution's key innovation: the reintroduction of mental states into the study of psychology. Unlike their behaviorist counterparts, proponents of the 'mind's new science' posit the existence of mental states with *causal powers*, i.e. real internal states that explain how sensory input is translated into behavioral output. Crucially, these mental states depend on, but are irreducible to, the physical properties of the mind.[55] Moreover, they can take other mental states, in addition to observable physical stimuli, as their input and have other mental states, in addition to observable physical responses, as their output.[56] The existence of mental states with causal powers underpins present-day psychology's claim to being autonomous of the physical and biological sciences.[57]

This emphasis on the cognitive revolution's disruptiveness obscures important continuities between present-day psychology and early twentieth-century behaviorism. Most fundamentally, present-day psychology shares with behaviorism its claim to being a hard science. This claim is grounded in *causal determinism*, i.e. the assumption that mental phenomena, like physical phenomena, are governed by causal regularities.[58] More importantly for the purposes of this paper, present-day psychology shares with behaviorism a commitment to *methodological behaviorism*: it seeks to discover these causal regularities solely based on observable behavioral data,[59] i.e. data that do not feature a

[53] Hafner-Burton et al. 2017; Mercer 2005.
[54] Gardner 1985; Kim 2011, p. 136
[55] Gardner 1985, pp. 38-40; Kim 2011
[56] Kim 2011.
[57] Kim 1992.
[58] Crane 2003, p. 4. Note, however, that causal determinism does not imply physicalism.
[59] Note that this includes the observable behavior of organisms, i.e. data on physical states. Crucially, methodological behaviorism presumes a more fundamental commonality between present-day psychology and the hard sciences: the assumption that mental phenomena are governed by causal regularities in much the same way as physical phenomena are and that these regularities are discoverable through science – an assumption known as *causal determinism* (Crane 2003, p. 4).

'first-person/third-person assymetry of epistemic access'.[60]

To understand how psychology squares the study of unobservable mental states with its claim to being a hard science,[61] and with methodological behaviorism in particular, we must understand its conception of the mind. This conception is rooted in early twentieth-century developments in mathematical logic, computation and engineering, and the emergence of one concept in particular: digital computation.

The principal ideas that underlie digital computation can be traced back to the British mathematician and logician Alan Turing. In the 1930s, Turing described a simple machine, known as 'Turing machine', that manipulates 'sequences of digits according to general rules defined over the digits'.[62] Specifically, the machine consists of a set of instructions – 'computations' – that translate input sequences of digits into output sequences of digits. The Turing machine's output depends on the input sequences of digits, i.e. the identity and location of a digit within a sequence of digits, and any internal states of the machine, i.e. 'states [from among a finite number of states] that the system can distinguish from one another'.[63] Turing famously demonstrated that 'any computation performed by any computer can be done by a Turing machine' (aka Church-Turing thesis).[64]

Crucially, a Turing machine – or digital computation – does not presuppose any particular physical medium or 'realizer': the rules governing the translation of input into output, the sequences of digits they manipulate and the machine's internal states can be defined independently of the physical properties of the system that implements them, and can be implemented by any physical system whose causal mechanisms are isomorphic to the computational relations among input, internal states and output.[65] In principle, then, Turing machines, or computers, can be compared to minds on purely functional grounds. Indeed, Turing himself proposed such an analogy: in 1950 he devised the so-called 'Turing test', which a machine would be said to have passed if its responses were indistinguishable from those of a human being. A machine's ability to pass, in turn, would support Turing's

[60] Costall 2004; Kim 2011, p. 86; Moore 2001

[61] It is worth noting that the dominant view in psychology, the cognitive sciences and philosophy of mind today considers the mind to be irreducible to the brain (Kim 2011). Thus, while technological advances in recent decades such as functional magnetic resonance imaging (fMRI) provide novel data on physical states of the brain, it does not follow that the *mind* per se has become more observable.

[62] Piccinini 2012.

[63] Piccinini 2012.

[64] Kim 2011, p. 143.

[65] Kim 2011, p. 145; Piccinini and Scarantino 2011, p. 8

notion that machines could think.[66]

Turing's notion of digital computation influenced the work of well-known neurophysiologists, mathematicians and engineers, among them Warren McCulloch and Walter Pitts.[67] In an influential paper, published in 1943, McCulloch and Pitts advanced the idea that nerves, due to their all-or-none property (they either fire or do not fire), could be thought of as propositional statements in logic. Neural processes, in turn, could be thought of as sequences of discontinuous events.[68] Pairing these ideas about the mathematical properties of neural networks with Turing's notion of computation, McCulloch and Pitts took Turing's analogy between minds and machines one step further: not only could machines be compared to minds, but minds could also be compared to machines. Like machines, they argued, minds perform digital computations (where the discontinuous events are represented by '1's and '0's).[69]

McCulloch and Pitts' theory of cognition laid the foundation for two important conceptual developments in the philosophy of mind and psychology. First, it gave rise to the idea that mental states, like machine states, are multiply realizable: the same mental state can be implemented by different physical-biological mechanisms.[70] By implication, mental states, like machine states, are *functional states*: what distinguishes them from one another – 'individuates' them – are not their physical (or other intrinsic) properties, but their causal-functional roles, i.e. their 'causal relations to inputs, outputs, and other functional states'.[71] This understanding of mental states is known as 'functionalism' in philosophy.

Second, McCulloch and Pitts' theory of cognition gave rise to the idea that minds not only are *like* machines; they in fact *are* machines. Specifically, it spawned the development of 'computationalism' in philosophy and cognitive science, which holds that the functional relations among input, output and mental states are computational: input and mental states are transformed into output via computations, i.e 'algorithmic (and perhaps heuristic) rules',[72] which operate on both in virtue of their syntactic properties, i.e.

[66] Gigerenzer and Goldstein 1996, p. 135.
[67] Piccinini 2004b.
[68] Gardner 1985; Piccinini and Scarantino 2011.
[69] Piccinini 2012.
[70] Kim 2011.
[71] Piccinini 2010, p. 270; also see Farrell 2014, pp. 399-400
[72] Crane 2003, p. 133.

their physical shape. According to computationalism, then, computation does not merely *describe* the mind; it *explains* it.[73] While functionalism does not imply computationalism,[74] computational functionalism is the dominant conception of mind in philosophy and psychology today.

Present-day psychology typically makes a further assumption about the mind. It assumes mental states to be *symbols* that have semantic, in addition to syntactic, properties: they are structured entities that can be distinguished from one another based on their physical shape *and* content-bearing representations that refer to particular objects or states of affairs in the world, whether or not these objects actually exist (e.g. Santa Claus may not exist, but still be represented in the mind) or these states of affairs are in fact true (e.g. it may or may not, in fact, be snowing).[75] To give an example, the symbols 'U-S-A' not only have a recognizable physical shape, but also represent three letters from the latin alphabet, which, in sequence, represent a country, the United States of America. The idea that mental states are symbols is known as representationalism.[76]

Whether computationalism implies representationalism or merely allows it remains controversial within the philosophy of mind.[77] The key point here is that present-day psychology typically assumes both, computationalism and representationalism. Indeed, the idea that mental states have representational content lies at the heart of intentionality, the notion that our thoughts, beliefs and desires 'are *about*, or are *directed upon*, objects that may or may not exist or have contents that may or may not be true'.[78] To return to our earlier example, it is the intentional quality of our mind that allows the symbols 'U-S-A' to stand for a country and enables us to have thoughts or beliefs about, and desires toward, that country. This quality is generally considered to be 'one of the distinctive and fundamental features of the human mind',[79] and is central, not only to the philosophy of mind and psychology, but also to microfoundational research in International Relations – a point we return to later.

[73] Kim 2011. Note that to claim the mind is computable is not to claim that it is programmable.
[74] Piccinini 2004b.
[75] Egan 2014; Kim 2011; Piccinini 2004a.
[76] Though note that computational processes operate on symbols in virtue of their syntactic, not their semantic, properties. In other words, it is their physical shape, not their content, that endows them with causal power.
[77] For example, Egan 2010; Piccinini 2004a; Piccinini 2008.
[78] Kim 2011, p. 24; Smith 2003, p. 32.
[79] Rupert 2008.

To conclude, many international relations scholars emphasize that psychology's account of the mind is empirically-grounded. However, this emphasis obscures psychology's reliance on a particular conception of the mind: the mind as computer and, specifically, as a symbol-processing machine.[80] This conception is vital to the empirical study of the mind: by defining mental states in relation to sensory input and behavioral output, it renders the mind amenable to empirical verification. In other words, computational functionalism's 'broadly behavioral approach to mentality' allows researchers to study the mind without abandoning methodological behaviorism.[81] Moreover, by specifying how the mind relates to sensory input and behavioral output, computational functionalism provides a general template for causal explanations of the mind and, thereby, informs the types of questions asked by psychologists. This is evident in psychology's tendency to focus either on the algorithms that enable us to solve particular cognitive tasks or on the syntactic properties of mental states, but not on the semantic properties of mental states.[82] As we will see, this has important implications for International Relations.

3.3.3 Of One Mind: Rational Choice and Psychology

By highlighting differences, rather than commonalities, between rational choice theory and psychological theories of International Relations, the 'tale of two minds' has served defenders and critics of rational choice theory alike: the belief that rational choice treats the mind as a black box renders it immune to psychological critiques, while the belief that psychology offers a data-driven account of the mind obscures its shared methodological and conceptual roots with rationalism. As the preceding sections show, however, neither belief is correct – at least as far as applications of rational choice to questions about the beliefs and behavior of individuals are concerned. This unearths an important commonality between rational choice and psychological theories that, to the best of my knowledge, remains unexamined in International Relations. The 'mind as intuitive statistician', an example of the mind as computer that has made deep inroads into the social sciences, helps illuminate this commonality.

Alongside computers, the cognitive revolution in psychology saw statistical programs

[80] Bruner 1990; Egan 2014; Gigerenzer and Goldstein 1996.
[81] Kim 2011, p. 136.
[82] Davies 2007.

emerge as an important analogy for the mind. Facilitated by the introduction of new statistical tools and methods into the scientific practice of psychology, the analogy was widely adopted by researchers and gave rise to the conception of mind as 'intuitive statistician'.[83] This conception can be considered a more specific instance of the mind as computer: the idea is that the algorithms involved in translating sensory input and mental states into mental or behavioral output can be described by probability theory. Its uses in psychology have been twofold. As a descriptive model, the intuitive statistician offers an explanation of how the mind actually accomplishes particular tasks. As a normative model, it offers a benchmark for 'rational' probability judgments and other outcomes of interest to experimenters, and thereby helps draw inferences about the actual rules that govern the mind from experimental data.

An example of the first is found in research on causal reasoning. In the 1940s, psychologists had already proposed statistical models as a way to conceptualize causal reasoning. In the late 1960s, Harold Kelley popularized this idea in his seminal contribution on 'attribution theory', which argued that the mental processes involved in causal reasoning can be described by Ronald Fisher's ANOVA models, a type of covariation model commonly used in psychological studies. Just as psychologists use ANOVA models to identify the relationship between two or more variables so the mind relies on ANOVA models to identify the relationship between two or more objects or events, Kelley suggested.[84] Subsequent studies failed to confirm Kelley's theory as a descriptive model of causal reasoning. However, it nonetheless became entrenched in the literature as a normative model, leading, among others, to the discovery of the so-called 'fundamental attribution error'.[85] The fundamental attribution error which scholars have applied to questions of International Relations.[86]

Like early research on causal reasoning in psychology, applications of rational choice theory to questions about individual beliefs in the social sciences rely on the intuitive statistician as a descriptive model of the mind: in accordance with Savage's subjective expected utility theory, they take individuals to be Bayesian statisticians who, in response

[83] Gigerenzer 1991. Crucially, Gigerenzer shows how the precise instantiation of the metaphor was shaped by the practical context within which it was adopted.
[84] Gigerenzer 1991.
[85] Ross 1977.
[86] For example, Jervis 1976; Larson 1997.

to new information, update their beliefs according to Bayes' rule (see section 2). In other words, rational choice theory assumes Bayes' rule to describe the mental processes involved in translating input into output. To give an example, imagine a foreign policy decision maker in 2018 who believed that the US was very likely to cooperate with Iran (mental state at t_0). Contrary to their belief, they soon received news that the US had decided to retract from the nuclear deal with Iran (input). Rational choice theory leads us to expect that the decision maker subsequently revised their belief to reflect the news, and now believes that the US is very unlikely to cooperate with Iran (output; mental state at t_1). Bayes' rule explains the transition from the decision maker's initial belief (mental state at t_0) to their subsequent belief (mental state at t_1).

Like research on causal reasoning, subsequent studies failed to confirm the intuitive statistician as a descriptive model of the mind and probability judgments in particular. This insight lies at the heart of Kahneman and Tversky's 'judgment under uncertainty' program, which relies on the intuitive statistician as a normative model of the mind: experiments in this program typically compare the responses of participants to reasoning problems involving uncertainty with the solutions generated by Bayesian statistics. Any discrepancies between the former and the latter are taken as evidence of 'heuristics and biases', which in turn characterize the actual mental processes involved in probability judgments. Well-known examples of heuristics and biases include overconfidence, the base rate fallacy and confirmation bias.[87]

While the 'judgment under uncertainty' program has gained considerable traction in economics and political science over the past two decades, its reception in International Relations has been mixed.[88] Proponents maintain that the program offers an empirically-grounded, and therefore potent, alternative to rational choice. By contrast, critics have either altogether dismissed psychological research as unfeasible or highlighted more specific problems that arise from the program's reliance on Bayesian statistics as a normative model of the mind in particular. Jonathan Mercer's widely-cited article, 'Rationality and Psychology in International Politics', offers a powerful example of the latter: it discusses how Kahneman and Tversky's more or less arbitrary choice of a Bayesian benchmark '[...] for problems that lack unique logical or statistical solutions' informs the conclusions

[87] For example, Tversky and Kahneman 1973; Lichtenstein, Fischhoff, and Phillips 1982.
[88] For example, Larson 1997; Khong 1992; Johnson 2004; McDermott 1998.

they draw about the actual mental processes that govern probability judgments from experimental data. More broadly, the article shows how psychology's reliance on rational choice as a normative model of the mind unnecessarily limits the scope of psychological research in International Relations.[89]

However, absent from all these perspectives, including Mercer's, is a recognition that psychological research can only rely on the intuitive statistician as a normative model of the mind to begin with, because, fundamentally, it shares with rational choice theory a conception of the mind as computer. On a mental interpretation, rational choice theory, like psychology, adopts a functionalist conception of mental states: preference relations and utility functions are taken to be real mental states, because they *play the role of* preferences and beliefs in causal accounts of behavior. Moreover, rational choice theory assumes the causal relations among preferences, utility functions and behavior to be computational.

Applications of rational choice theory to questions about individual beliefs illustrate this well: as we saw, they rely on a specific instance of the mind as computer, namely, the intuitive statistician, to describe the mental processes involved in the updating of beliefs. Similarly, applications of rational choice theory to questions about individual behavior can be understood as relying on a particular instance of computationalism. Here, the consistency requirement, not Bayes' rule, explains the translation of input into output: across choice contexts, rational choice theory expects individuals to choose the option (output) that is most consistent with their preferences and beliefs (mental states; input). Where this requirement is fulfilled, an individual's preference relations can be said to explain their behavior.[90]

To conclude, proponents of psychological research in International Relations highlight differences between rational choice and psychology – and rightly so: research on causal reasoning and the judgment under uncertainty program offer novel insights into the mental processes that transform sensory input into behavioral output. A growing body of research in International Relations attests to the explanatory power that can be gained from replacing the intuitive statistician or consistency requirement with these insights.

[89] Mercer 2005. Note that probability theory does not offer unique solutions to any problem, see Daston 1988.

[90] See Dietrich and List 2016b.

However, burried in controversies about the merits and demerits of psychological theories vis-à-vis rational choice theory lies an unacknowledged commonality between the two: fundamentally, both conceptualize the mind as computer. Consequently, rational choice theory and psychological theories are quite literally cast in the same mold.

3.4 Implications for International Relations

3.4.1 Reform, Not Revolution

Earlier, we saw that the interpretation of rational choice theory remains contested, even among its proponents. In particular, scholars in economics, political theory, and philosophy disagree on whether rational choice explanations imply the existence of, and rely on claims about, unobservable mental states or not. At stake in this debate are not only the ontological foundations of rational choice theory, but also rational choice theory's relationship to insights from psychology: whereas a behavioral or other non-mental interpretation renders rational choice theory immune to psychological critiques, a mental interpretation leaves it open to them. Despite the obvious relevance of this debate to International Relations, international relations scholars have rarely, if ever, explicitly engaged with it. Doing so provides a more accurate understanding of rational choice theory and of psychology's contribution to International Relations.

To begin, the previous section noted that rational choice theory is not a unified substantive theory, but a methodological approach, and, consequently, any interpretation of the theory hinges on the theoretical and substantive assumptions made in particular contexts.[91] Whereas rational choice theory's application to questions about individual beliefs and behavior in international politics implies mentalism, its application to questions about aggregate outcomes like war might plausibly be interpreted as behavioral: here, the 'preference relations' invoked by scholars to explain the outcome might not refer to preferences *per se*, but merely *describe* behavioral patterns at lower levels of analysis. Alternatively, where rational choice theory is applied to questions about group behavior, like foreign policy decisions, it might be neither mental nor behavioral: the preference relations invoked by scholars might *explain* behavior, but emerge from non-psychological processes

[91] Herfeld 2020.

like voting and, therefore, not reflect the preference relations of any one individual within the group.

Consequently, the extent to which rational choice theory depends on microfoundational claims, and the kind of microfoundational claims it depends on, varies along with its context of application: if we understand a microfoundational approach broadly as an 'analytic commitment to exploring [causal] mechanisms at a lower level of analysis', then, as Joshua Kertzer notes, 'there is nothing inherent about a microfoundational approach that demands fealty to the first image'.[92] Only where rational choice explanations rely on microfoundational claims at the individual level do they imply mentalism and vice versa.

This has several important implications for debates about rational choice theory and psychology in International Relations and our understanding of psychology's contribution to the discipline. First, psychological theories of International Relations should be understood as competing with a *subset of rational choice theories*, not rational choice theory as a whole. Where the application of rational choice theory to questions of international politics can be interpreted as behavioral, or as neither behavioral nor mental, psychological critiques miss the mark, for the rational choice explanations in question do not hinge on mental concepts, i.e. microfoundational claims at the individual level, to begin with.[93]

Second, proponents of psychological research sometimes suggest that the 'aggregation problem' is no more or less severe for psychological theories of International Relations than it is for other theories that rely on 'simplified abstractions at higher levels of analysis'.[94] Janice Gross Stein, for instance, contends that 'rational choice theorists routinely assume a unitary decision maker and go up the ladder of analysis to attribute these properties [the individual properties of information processing and probability estimation] to the state'.[95] However, this line of defense presumes a mental interpretation of rational choice theory. By contrast, where a behavioral or other non-mental interpretation is more plausible, the use of simplified abstractions is grounded in a purely formal conception of rationality as consistency (see section 2); it does not rely on mental concepts. In these instances, rational choice models may travel across levels of analysis in ways that psychological

[92] Kertzer 2017, p. 84.
[93] This is not to suggest that rational choice theory offers a compelling explanation in those instances – it may well fail to convince on other grounds.
[94] Gildea 2020.
[95] Stein 2017.

theories cannot – provided that the consistency requirement is met.

Finally, where psychological theories do compete with rational choice theory, they do not 'bring mental states back in',[96] as some proponents have suggested: the rational choice explanations in question rely no less on mental concepts than psychological theories do. To be sure, the behavioral revolution does draw attention to the role that individual beliefs, preferences and behavior play in causal accounts of international politics. Given the discipline's longstanding tradition of state-centered theorizing, this is no small feat. However, the key difference between rational choice and psychological theories of International Relations is not their ability to theorize this role. Nor is it their ability, in principle, to empirically ground their theories: although proponents of psychological research emphasize that psychology, unlike rational choice theory, supplies an empirically-validated account of the mind,[97] there is nothing inherent in rational choice theory that necessitates a reliance upon abstract assumptions. To the contrary, its core hypotheses – the idea that behavior is explainable by preference relations and that beliefs are updated in accordance with Bayes' rule – are falsifiable.[98]

Rather, the key difference between rational choice theory and psychological theories lies in their substantive claims about the mind. However, even this difference is less profound than the notion of a 'behavioral revolution' might suggest: as the previous section showed, rational choice theory's and psychology's substantive claims are grounded in a common conception of the mind as computer. As we saw, this is true for psychological theories that rely on the intuitive statistician as a normative model of the mind: these theories simply replace the intuitive statistician with alternative instantiations of the mind as computer. Research on causal reasoning and the judgment under uncertainty program offer prominent examples. It is yet more apparent in recent work on psychology in International Relations, which often continues to rely on, rather than replace, the intuitive statistician and the consistency requirement as descriptive models of the mind. Three important contributions help illustrate this.

The first is Keren Yarhi-Milo's book, *Who Fights for Reputation*, which seeks to understand whether and when leaders are willing to use military force to maintain their

[96] Kertzer and Tingley 2018, p. 4.
[97] Hafner-Burton et al. 2017.
[98] Dietrich and List 2016b.

reputation for resolve in international politics. Focusing on international crises, Yarhi-Milo argues that a leader's willingness to deploy force will depend on two things: first, their 'self-monitoring' dispositions, and, second, their 'beliefs about the efficacy of military force'.[99] 'Self-monitoring' is a concept borrowed from social psychology and describes the 'extent to which individuals strategically cultivate their public appearances'.[100] Integrating this concept with literature on reputational concerns in international politics, Yarhi-Milo hypothesizes that individuals with a disposition toward self-monitoring will be more concerned about their reputation for resolve, and therefore more likely to support public displays of military force in international crises, than individuals with little or no disposition toward self-monitoring.

Beliefs about the efficacy of military force tap into a familiar distinction in International Relations: the distinction between 'hawks', who tend to regard military force as an effective tool of statecraft in foreign policy, and 'doves', who tend to believe military force does more harm than good. Combining this distinction with self-monitoring dispositions, Yarhi-Milo derives four ideal-type leaders to explain variation in decision makers' propensity to fight for reputation. Two survey experiments conducted with American and Israeli participants, respectively, a large-N observational study on all militarized interstate disputes the U.S. was involved in between 1945 and 2008, and three historical case studies on former presidents of the U.S. lend strong support to her theory.[101]

Yarhi-Milo's substantive claims about reputational concerns and the willingness to use force in international crises depart from the claims of rationalist models. Most notably, they predict individual-level variation in responses to international crises. More fundamentally, however, Yarhi-Milo's theory is in keeping with rationalist claims about the mind: it does not contest the consistency requirement, i.e. the idea that decisions to use military force (behavioral output) are explainable by individuals' preference relations in combination with their beliefs (mental input). It merely posits actor-specific variation in preference relations due to relatively stable, individual-level traits or dispositions.

In another notable contribution to the recent literature on psychology in International Relations, Joshua Kertzer, Brian Rathbun and Nina Rathbun investigate whether and

[99] Yarhi-Milo 2018, p. 19.
[100] Yarhi-Milo 2018, p. 20.
[101] Yarhi-Milo 2018.

how individuals respond to costly signals in international politics. Drawing on political psychology, the authors hypothesize that individuals' responses to signals vary depending on their prior beliefs, particularly their foreign policy beliefs. Specifically, focusing on signals of reassurance, they argue that individuals with cooperative views of international politics, and a benign image of the adversary, will be more responsive to costly signals than individuals with competitive views and a malign image. In other words, only some individuals will respond to costly signals in the way rationalist models tell us they should. The authors tested their theory in an online survey experiment conducted with over 1800 American participants. In the experiment, each participant received a hypothetical Iranian proposal for the 2015 nuclear deal. The proposals specified the concessions Iran would be willing to make in order to reach an agreement with the negotiating parties, and mainly differed with respect to the costs they incurred to Iran. After reading the proposal, participants were asked to estimate Iran's trustworthiness. The results lend strong support to the author's theory.[102]

As in the previous example, Kertzer, Rathbun and Rathbun's predictions differ from those of rationalist models, but do not fundamentally contest rationalist claims about the mind: in response to incoming information (here, the hypothetical Iranian proposal), individuals are thought to update their beliefs about Iranian intentions in accordance with Bayes' rule. The key difference to canonical models is that prior beliefs are hypothesized to assign weights to incoming information. Holding information constant, Bayesian updating therefore leads to divergence rather than convergence in individuals' beliefs about Iranian intentions.[103] In other words, Kertzer, Rathbun and Rathbun refine rather than replace the conception of mind as intuitive statistician.

Finally, Brian Rathbun's book, *Reasoning of State*, interrogates the psychological underpinnings of instrumental rationality, i.e. rational behavior. Rational behavior, Rathbun argues, requires procedural rationality, i.e. rational thinking, on the part of individuals. Rational thinking in turn is marked by objectivity and deliberation and, as Rathbun shows, exists along a continuum: individuals can think more or less rationally. Drawing on insights from psychology and behavioral economics, Rathbun characterizes this

[102] Kertzer, B. C. Rathbun, and N. S. Rathbun 2020.
[103] For a more technical explanation of this point see Kertzer, B. C. Rathbun, and N. S. Rathbun 2020: supplementary appendix.

variation in terms of the widely-cited distinction between two modes of cognitive processing, System I and System II. Whereas System I relies on heuristics and biases, is quick, unconscious and intuitive, rational thinking is associated with System II, which is more deliberative, slow and conscious. While all individuals rely on both, System I and System II, to process information, some individuals rely more on System II than others. Consequently, they think more rationally. Rathbun hypothesizes that this variation in rational thinking across individuals is driven by variation in epistemic motivation, i.e. individuals' commitment to 'effortful cognitive activity'.[104] A laboratory experiment conducted in collaboration with Kertzer, and the in-depth historical analysis of four cases – Bismarck, Richelieu, Churchill and Reagan – lend strong support to Rathbun's theory.[105]

By positing systematic variation in rationality among individuals, and identifying a distinctly nonrational way of thinking, Rathbun's theory more clearly departs from rational choice theory than the previous two examples: it relies on the intuitive statistician and consistency requirement as descriptive models for individuals who draw more on System II, but not for so-called 'romantics' who primarily draw on System I. All the more, however, Rathbun's theory highlights the fundamental commonality that underlies substantive differences between rational choice theory and psychological theories: Rathbun's characterization of romantics draws heavily on insights from the judgement under uncertainty program. By implication, his theory of variation in rational thinking continues to rely on, rather than replace, rationalist claims about the mind – if not as a descriptive model, then as a normative model (see section 3.3).

All three contributions challenge rational choice theory, both substantively and theoretically. Substantively, they highlight the role actor-specific variation plays in foreign policy decisions, even where we might expect structural pressures in the international system to override individual-level factors (e.g. during the Cold War).[106] The empirical evidence in support of all three theories attests to the explanatory power that is gained from understanding this variation. Theoretically, they make good on Mercer's claim that 'no single approach has a lock on understanding rationality':[107] they show how psycho-

[104] B. C. Rathbun 2019, p. 22.
[105] B. C. Rathbun 2019; also see B. C. Rathbun, Kertzer, and Paradis 2017.
[106] Hafner-Burton et al. 2017 themselves do not seem to come down clearly on one side or the other in the debate about how broad the challenge posed by behavioral International Relations is.
[107] Mercer 2005, p. 77.

logical theories are able to explain both, outcomes that are consistent with and outcomes that deviate from, the predictions of rationalist models. Actor-specific theories of international politics thus stand as an example of how efforts to dismantle the dichotomy between rationality and psychology, and to decouple rationality from rational choice theory, have been successful in International Relations.

At the same time, the emphasis many international relations scholars place on differences between rational choice and psychological theories obscures the fact that rationalist predictions can only be subsumed under psychological theories because the two approaches are, quite literally, of one mind: they share a conception of the mind as computer and, as Yarhi-Milo's and Kertzer, Rathbun and Rathbun's research shows, often even the more specific conception of the mind as intuitive statistician. The introduction of psychology into International Relations is therefore better understood as a reform, not a revolution.

This is not a matter of mere semantics: how we construe the contribution of psychological theories to International Relations, and situate them in relation to other approaches in the discipline, impacts the vision we develop for the emerging subfield of psychological International Relations. The foregoing analysis suggests that, more than an empirical account of the mind *per se*, psychology's core contribution to the discipline lies in its ability to theorize *individual-level variation* in mental states and processes and, thereby, explain individual-level variation in behavior. This claim dovetails with Hafner-Burton et al.'s introduction to the 2017 special issue of *International Organization*, which singles out 'heterogeneity of preferences, beliefs, and decision-making processes even across similarly situated individuals' as an emerging theme in behavioral International Relations.[108] If the ability to theorize variation in mental states and processes is a core component of our vision for a psychology of international politics, then it is imperative that we understand how present-day psychology's reliance upon a conception of the mind as computer limits its realization.

3.4.2 Variation in Mind

A key strength of psychological research in International Relations is its ability to explain variation in behavior across similarly situated individuals. It does so based on variation

[108] Emphasis removed, Hafner-Burton et al. 2017. The authors also identify a second theme, the aggregation issue, which is not discussed here.

in preferences and beliefs, i.e. mental states, variation in mental processes, or both.[109] Focusing on preferences and beliefs, the following discusses how present-day psychology's reliance on a conception of the mind as computer limits its ability to theorize this variation. Specifically, it shows that, insofar as psychological research relies on the mind as computer, it can observe, but not explain, variation in the *propositional content* of preferences and beliefs. Simply put, the psychological research international relations scholars most heavily rely on is ill-equipped to address the central question of how and why meaning, i.e. the content of our thoughts, varies across individuals. However, this variation promises to be especially important for a psychology of international politics.

Like rational choice theory, psychological theories of International Relations rely on preferences (or 'desires') and beliefs, in addition to mental processes, to explain behavior. Preferences and beliefs generally consist of two elements: first, an attitude such as *hoping*, *believing*, *wanting* or *expecting*; second, a situation or state of affairs in the world that is represented in the mind.[110] In the examples from International Relations discussed above, these situations include a country having a reputation for resolve, military force being an efficacious foreign policy tool, and Iran being trustworthy. In other words, preferences and beliefs are what philosophers of mind call 'propositional attitudes': an individual might *believe* that Iran is trustworthy or might *doubt* that Iran is trustworthy. In either case, the individual in question has a particular attitude toward the proposition 'Iran is trustworthy'.[111] This is possible only if we assume mental states to have representational content, that is, to represent objects and states of affairs in the world (also see section 3.2). Not only, then, do psychological theories of International Relations rely on computationalism; they also assume representationalism about the mind.

However, computation does not account for representational, or 'mental', content. The philosopher John Searle famously advanced this argument in his 'Chinese room' thought experiment. In brief, the thought experiment involves a person who has no knowledge of the Chinese language and is locked in a room. The person receives two Chinese texts, 'the script' and 'the story', along with a set of rules that specify how to translate one string of Chinese symbols into another string of Chinese symbols using the two texts. The rules

[109] Hafner-Burton et al. 2017.
[110] Crane 2003.
[111] Crane 2003.

are stated in English and, like computational processes, can be implemented solely based on the syntactic properties of the symbols, not their semantic properties. Aided by these rules, the script, and the story, the person in the room could therefore 'output' a string of Chinese symbols in response to an input string of Chinese symbols without understanding a single word of Chinese. According to Searle, the thought experiment demonstrates that computation – the 'rule-governed manipulations of symbols based on their syntactic shapes' – is insufficient to explain a key feature of human cognition: the ability to ascribe and understand representational content, i.e. to acquire semantic knowledge.[112]

One need not refute computationalism, as Searle does, to agree with his basic conclusion: many, if not most, proponents of computationalism today recognize the limitations of computation once highlighted by Searle. This is true for both, proponents who adopt and proponents who reject a 'semantic view' of computation. Proponents who adopt a semantic view of computation take computational processes to operate on a particular type of input, namely, symbols. On their view, computation *presupposes* representation (see section 3.2). Thus, it cannot explain representation on pain of circularity.[113] Proponents who reject a semantic view do not take computation to pressuppose representation. The philosopher of mind Gualtiero Piccinini, for example, demonstrates that 'computation ascription alone is insufficient for the ascription of mental content'.[114] Therefore, he argues, 'the question of computation is independent of the question of content'.[115] This renders computationalism immune to critiques of representationalism. At the same time, Piccinini asserts that 'representations are a staple of mainstream psychology and neuroscience [...] they are not likely to be eliminated from our explanations of cognition'.[116] Thus, a complete account of cognition requires a theory of content in addition to a computational theory of mind, and a non-computational mechanism that maps the semantic properties of mental states onto the computational relations among them to link the two.[117]

More generally, then, we can take Searle's thought experiment to demonstrate, not the inability of computing machines to acquire semantic knowledge, but the inability of computation to explain how they do. In other words, the thought experiment does not so

[112] Kim 2011, p. 162; also see the original paper and peer commentary, Searle 1980.
[113] Crane 2003.
[114] Piccinini 2004a, p. 388.
[115] Piccinini 2004a, p. 405.
[116] Piccinini 2009, p. 526.
[117] Piccinini 2004a.

much render the conception of mind as computer implausible or wrong as it reveals it to be heavily incomplete.[118] As the prominent philosopher of mind Jaegwon Kim argues, Searle's thought experiment also suggests why it is incomplete: it demonstrates that, absent any interaction with the outside world, the person in the room has no way of acquiring semantic knowledge, i.e. no way of learning what the Chinese symbols mean. Put simply, it shows that 'syntactic symbol manipulation alone will not yield such [semantic] knowledge; only real-world experience will'.[119]

This has important consequences for psychological research in International Relations, and for efforts to theorize heterogeneity in individual preferences, beliefs and behavior in particular: it implies that, insofar as psychological research relies on the mind as computer, it can observe, but not explain, variation in preferences and beliefs. That is, psychological research can explain variation in belief updating and behavior based on variation in the attitudes individuals hold toward a given proposition or variation in how they interpret that proposition.[120] However, it cannot explain variation in how individuals interpret a given proposition. To illustrate, psychological theories can explain variation in responses to costly signals of reassurance from Iran based on whether or not an individual believes that Iran is trustworthy, but they cannot explain how or why individuals understand the proposition 'Iran is trustworthy' the way they do.

There are at least two reasons why this limitation is particularly severe for psychological research in International Relations. First, the set of behaviors that constitute international politics depend on representational content: they presume the existence of particular categories, like 'states', 'international organizations', 'rebel groups', 'international law', 'diplomacy' or 'war', and norms to define and structure interactions in international politics. Therefore, the question of how representational content emerges and changes – that is, how it varies over time – is inherently interesting for international relations scholars who seek to understand the individual-level mechanisms that underpin stability and change in international order.

Second, international politics involves the interaction of individuals from diverse social, cultural and institutional contexts. If semantic knowledge emerges from individuals'

[118] Kim 2011.
[119] Kim 2011, p. 163.
[120] Although note that, technically, the non-semantic properties of mental states – here, preferences and beliefs – do the explanatory work, not the semantic properties.

interactions with their environment, as philosophers of mind suggest it does, then the semantic knowledge of individuals in international politics will vary accordingly. Given that 'things only have meaning in a field, that is, in relation to the meanings of other things',[121] this variation, in turn, likely impacts their understanding of propositional statements in, and about, international politics. In other words, individuals in international politics are, if anything, particularly likely to vary, not only in the attitudes they hold toward a given propositional statement, but in their very understanding of the statement itself. Indeed, existing scholarship suggests as much. In their experimental study on reputation costs, for instance, Ryan Brutger and Joshua Kertzer find that doves and hawks vary, not only in the importance they ascribe to reputation, but also in their understanding of what reputation costs are.[122]

Insofar as psychological research in International Relations continues to rely on mainstream psychology's conception of the mind as computer, it can catalogue variation in representational content within individuals over time or across individuals from distinct contexts. However, it cannot theorize this variation: it cannot move beyond idiosyncratic observations to understand how representational content emerges and, relatedly, how it is patterned, i.e. how it varies *systematically*.

3.4.3 A Meeting of Minds

Of course, we might consider this state of affairs to be part of a healthy division of labor within the discipline: whereas psychological research explains individual belief updating and behavior *given representational content*, constructivist scholars address themselves to the social processes that drive meaning construction – a division akin, perhaps, to Robert Cox's distinction between problem-solving and critical theories of world order, but applied to the mind.[123] Indeed, many international relations scholars seem to tacitly assume such a division of labor. This is apparent in the contributions to psychological research cited earlier, which assume rather than explain the representational content that underlies variation in individual preferences and beliefs. It is also apparent in calls for collaboration between psychological and constructivist scholarship, which generally emphasize their

[121] Taylor 1971, p. 11.
[122] Brutger and Kertzer 2018.
[123] Cox 1981.

complementarity. In the introductory chapter to his co-edited volume on *Psychology and Constructivisim in International Relations*, for example, Vaughn Shannon argues that '[...] the virtues of constructivism set macrostructural boundaries on the perceptions of values and possible responses based on prevailing norms and identity, while psychology provides microfoundations for the motives behind normative behavior and identity change'.[124] And, while the volume's contributions seek, not only to 'address different questions [...] in a complementary fashion', but to generate '[...] insights neither constructivists nor political psychologists would have been likely to achieve independently', ultimately, they nonetheless treat individual cognition as distinct and prior to the social processes of meaning construction.

However, if we take seriously the importance of representational content to cognition, it becomes clear that such a division of labor is untenable: it sustains a disjuncture between individual cognition on the one hand and the social processes of meaning construction on the other. Only if we understand the processes by which meaning is continuously made and remade at the individual level can psychological research engage the 'highly variegated, internally contradictory, loosely integrated, and deeply interpenetrated' systems of meaning and practice within which individuals are embedded, and can constructivist research develop a microfoundational account of stability and change in international politics.[125] To this end, international relations scholars must look beyond cognitive and social psychology to develop a conception of the mind better suited to understanding how minds create and acquire meaning and, more generally, how they are shaped by social interaction. Emanuel Adler's recently published seminal contribution, *World Ordering*, offers a first example of this type of research: although it focuses on practice, not on representation, as a starting point for developing an alternative conception of the mind, it understands cognition as socially constituted.[126]

3.5 Conclusion

Proponents of psychology have heralded the recent surge in psychological research in International Relations as a 'behavioral revolution': like the cognitive revolution in psychology,

[124] Shannon and Kowert 2012, p. 14.
[125] Reus-Smit 2017, p. 856.
[126] Adler 2019.

the behavioral revolution 'seeks to bring mental states back in' and offer an empirical account of the mind.[127] If we pursue this narrative further, we might wager that, a decade or two from now, students of International Relations will be taught a story similar to that taught to undergraduates in psychology today: like the cognitive sciences lit psychology's way out of the 'Dark Ages' of behaviorism,[128] so too does psychology light International Relation's way out of the 'Dark Ages' of rationalism. As this paper demonstrates, such a narrative would be deeply flawed: the notion of a behavioral revolution is founded upon widely held, yet false, beliefs about rational choice, psychology and their relationship to one another. A more accurate understanding of both reveals that rational choice is not free of the mind and psychology relies on a conception of the mind that is informed by behaviorism and rationalism. The introduction of psychology into International Relations is therefore better understood as a reform, not a revolution.

International relations scholars have much to gain from this admission: to understand the conception of the mind that underpins psychological research is also to understand its limits. As this paper discusses, the conception of the mind as computer is unsuited to theorizing how minds create and acquire meaning. However, doing so is key to understanding systematic variation in individual beliefs, preferences and behavior, and how this variation links to the social processes that drive macro-level changes in international politics. Therefore, this paper joins calls for closer collaboration between psychological and constructivist approaches. At the same time, it argues that such collaboration must move beyond the current dichotomy between cognitive and social processes. To this end, future research must engage in a genuine interdisciplinary dialogue that recognizes distinct approaches within psychology and the philosophy of mind, including theories of content, and critically interrogates their assumptions. By identifying the core assumptions that underpin psychological research in International Relations to date, and situating them in relation to rational choice theories, this paper takes a first step in the proposed direction.

[127] Kertzer and Tingley 2018.
[128] Roediger 2004.

References

Adler, Emanuel (2019). *World Ordering: A Social Theory of Cognitive Evolution*. Cambridge University Press.

Barkin, J. Samuel (2015). "On the Heuristic Use of Formal Models in International Relations Theory." In: *International Studies Review* 17 (4), pp. 617–634.

Bruner, Jerome S. (1990). *Acts of Meaning*. Harvard University Press.

Bruni, Luigino and Francesco Guala (2001). "Vilfredo Pareto and the Epistemological Foundations of Choice Theory." In: *History of Political Economy* 33 (1), pp. 21–49.

Bruni, Luigino and Robert Sugden (2007). "The Road Not Taken: How Psychology Was Removed From Economics, and How It Might Be Brought Back." In: *The Economic Journal* 117 (516), pp. 146–173.

Brutger, Ryan and Joshua D. Kertzer (2018). "A Dispositional Theory of Reputation Costs." In: *International Organization* 72 (3), pp. 693–724.

Costall, Alan (2004). "From Darwin to Watson (and Cognitivism) and Back Again: The Principle of Animal-Environment Mutuality." In: *Behavior and Philosophy* 32 (1), pp. 179–195.

Cox, Robert W. (1981). "Social Forces, States and World Orders: Beyond International Relations Theory." In: *Millennium* 10 (2), pp. 126–155.

Crane, Tim (2003). *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*. 2nd ed. Routledge.

Daston, Lorraine (1988). *Classical Probability in the Enlightenment*. Princeton University Press.

Davies, M. (2007). "Cognitive Science." In: *The Oxford Handbook of Contemporary Philosophy*. Ed. by Frank Jackson and Michael Smith. Oxford University Press, pp. 358–394.

Dietrich, Franz and Christian List (2016a). "Mentalism Versus Behaviourism in Economics: A Philosophy-of-Science Perspective." In: *Economics and Philosophy* 32 (2), pp. 249–281.

Dietrich, Franz and Christian List (2016b). "Reason-Based Choice and Context-Dependence: An Explanatory Framework." In: *Economics and Philosophy* 32 (2), pp. 175–229.

Egan, Frances (2010). "Computational Models: A Modest Role for Content." In: *Studies in History and Philosophy of Science* 41 (3), pp. 253–259.

Egan, Frances (2014). "How to Think About Mental Content." In: *Philosophical Studies* 170 (1), pp. 115–135.

Farrell, Martin (2014). *Historical and Philosophical Foundations of Psychology*. Cambridge University Press.

Gardner, Howard (1985). *The Mind's New Science: A History of the Cognitive Revolution*. Basic Books.

Gigerenzer, Gerd (1991). "From Tools to Theories." In: *Psychological Review* 98 (2), pp. 254–267.

Gigerenzer, Gerd and Daniel Goldstein (1996). "Mind as Computer: Birth of a Metaphor." In: *Creativity Research Journal* 9 (2), pp. 131–144.

Gildea, Ross James (2020). "Psychology and Aggregation in International Relations." In: *European Journal of International Relations* 26 (S1), pp. 166–183.

Giocoli, Nicola (2003). *Modeling Rational Agents: From Interwar Economics to Early Modern Game Theory*. Elgar.

Giocoli, Nicola (2005). "Modeling Rational Agents the Consistency View of Rationality and the Changing Image of Neoclassical Economics." In: *Cahiers d'Économie Politique* 49 (2), pp. 177–208.

Giocoli, Nicola (2013). "From Wald to Savage: Homo Economicus Becomes a Bayesian Statistician." In: *Journal of the History of the Behavioral Sciences* 49 (1), pp. 63–95.
Guala, Francesco (2019). "Preferences: Neither Behavioural Nor Mental." In: *Economics and philosophy* 35 (3), pp. 383–401.
Hafner-Burton, Emilie M. et al. (2017). "The Behavioral Revolution and International Relations." In: *International Organization* 71 (S1), S1–S31.
Hands, D. W. (2010). "Economics, Psychology and the History of Consumer Choice Theory." In: *Cambridge Journal of Economics* 34 (4), pp. 633–648.
Herfeld, Catherine (2018). "From Theories of Human Behavior to Rules of Rational Choice." In: *History of Political Economy* 50 (1), pp. 1–48.
Herfeld, Catherine (2020). "The Diversity of Rational Choice Theory: A Review Note." In: *Topoi* 39 (2), pp. 329–347.
Jervis, Robert (1976). *Perception and Misperception in International Politics*. Princeton University Press.
Johnson, Dominic D. P. (2004). *Overconfidence and War: The Havoc and Glory of Positive Illusions*. Harvard University Press.
Kertzer, Joshua D. (2017). "Microfoundations in International Relations." In: *Conflict Management and Peace Science* 34 (1), pp. 81–97.
Kertzer, Joshua D., Brian C. Rathbun, and Nina Srinivasan Rathbun (2020). "The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations." In: *International Organization* 74 (1), pp. 95–118.
Kertzer, Joshua D. and Dustin H. Tingley (2018). "Political Psychology in International Relations: Beyond the Paradigms." In: *Annual review of political science* 21 (1), pp. 319–339.
Khong, Yuen Foong (1992). *Analogies at war: Korea, Munich, Dien Bien Phu, and the Vietnam decisions of 1965*. Princeton University Press.
Kim, Jaegwon (1992). ""Downward Causation" in Emergentism and Nonreductive Physicalism." In: *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. Ed. by Ansgar Beckermann, H. Flohr, and Jaegwon Kim. W. de Gruyter.
Kim, Jaegwon (2011). *Philosophy of Mind*. 3rd ed. Westview Press.
Larson, Deborah Welch (1997). *Anatomy of Mistrust : U.S.-Soviet Relations During the Cold War*. Cornell University Press.
Lichtenstein, S., B. Fischhoff, and L. D. Phillips (1982). "Calibration of Probabilities: The State of the Art to 1980." In: *Judgment Under Uncertainty: Heuristics and Biases*. Ed. by Daniel Kahnemann, Paul Slovic, and Amos Tversky. Cambridge University Press, pp. 306–334.
MacDonald, Paul K. (2003). "Useful Fiction or Miracle Maker: The Competing Epistemological Foundations of Rational Choice Theory." In: *American Political Science Review* 97 (4), pp. 551–565.
McDermott, Rose (1998). *Risk-Taking in International Politics: Prospect Theory in American Foreign Policy*. University of Michigan Press.
Mercer, Jonathan (2005). "Rationality and Psychology in International Politics." In: *International Organization* 59 (1), pp. 77–106.
Miller, George A. (2003). "The Cognitive Revolution: A Historical Perspective." In: *Trends in Cognitive Sciences* 7 (3), pp. 141–144.
Moore, Jay (1996). "On the Relation Between Behaviorism and Cognitive Psychology." In: *The Journal of Mind and Behavior* 17 (4), pp. 345–367.
Moore, Jay (2001). "On Distinguishing Methodological from Radical Behaviorism." In: *European Journal of Behavior Analysis* 2 (2), pp. 221–244.

Moscati, Ivan (2019). *Measuring Utility: From the Marginal Revolution to Behavioral Economics*. Oxford University Press.
Neumann, John Von (1958). *The Computer and the Brain*. Yale University Press.
Neumann, John Von and Oskar Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton University Press.
Okasha, Samir (2016). "On the Interpretation of Decision Theory." In: *Economics and Philosophy* 32 (3), pp. 409–433.
O'Neil, W. M. (1995). "American Behaviorism." In: *Theory & Psychology* 5 (2), pp. 285–305.
Piccinini, Gualtiero (2004a). "Functionalism, Computationalism, and Mental Contents." In: *Canadian Journal of Philosophy* 34 (3), pp. 375–410.
Piccinini, Gualtiero (2004b). "Functionalism, Computationalism, and Mental States." In: *Studies in History and Philosophy of Science* 35 (4), pp. 811–833.
Piccinini, Gualtiero (2008). "Computation without Representation." In: *Philosophical Studies* 137 (2), pp. 205–241.
Piccinini, Gualtiero (2009). "Computationalism in the Philosophy of Mind." In: *Philosophy Compass* 4 (3), pp. 515–532.
Piccinini, Gualtiero (2010). "The Mind as Neural Software? Understanding Functionalism, Computationalism, and Computational Functionalism." In: *Philosophy and Phenomenological Research* 81 (2), pp. 269–311.
Piccinini, Gualtiero (2012). "Computationalism." In: *The Oxford Handbook of Philosophy of Cognitive Science*. Ed. by Eric Margolis, Richard Samuels, and Stephen P. Stich. Oxford University Press, pp. 222–249.
Piccinini, Gualtiero and Andrea Scarantino (2011). "Information Processing, Computation, and Cognition." In: *Journal of Biological Physics* 37 (1), pp. 1–38.
Rathbun, Brian C. (2019). *Reasoning of State: Realists, Romantics and Rationality in International Relations*. Cambridge University Press.
Rathbun, Brian C., Joshua D. Kertzer, and Mark Paradis (2017). "Homo Diplomaticus: Mixed-Method Evidence of Variation in Strategic Rationality." In: *International Organization* 71 (S1), S33–S60.
Reus-Smit, Christian (2017). "Cultural Diversity and International Order." In: *International Organization* 71 (4), pp. 851–885.
Riker, William H. (1995). "The Political Psychology of Rational Choice Theory." In: *Political Psychology* 16 (1), pp. 23–44.
Roediger, Henry L. III (2004). *What Happened to Behaviorism*. URL: https://www.psychologicalscience.org/observer/what-happened-to-behaviorism.
Rosenberg, Alexander (2008). *Philosophy of Social Science*. 3rd ed. Westview.
Ross, Lee (1977). "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process." In: *Advances in Experimental Social Psychology* 10, p. 174.
Rupert, Robert D. (2008). "Causal Theories of Mental Content." In: *Philosophy Compass* 3 (2), pp. 353–380.
Satz, D. and J. Ferejohn (1994). "Rational Choice and Social Theory." In: *The Journal of Philosophy* 91 (2), pp. 71–87.
Searle, John R. (1980). "Minds, Brains, and Programs." In: *The Behavioral and Brain Sciences* 3 (3), pp. 417–424.
Shannon, Vaughn P. and Paul Kowert (2012). *Psychology and Constructivism in International Relations: An Ideational Alliance*. University of Michigan Press.
Smith, Brian C. (2003). "The Foundations of Computing." In: *Computationalism: New Directions*. Ed. by Matthias Scheutz. MIT Press, pp. 23–58.

Snidal, Duncan (2013). "Rational Choice and International Relations." In: *Handbook of International Relations*. Ed. by Walter Carlsnaes, Thomas Risse, and Beth A. Simmons. SAGE, pp. 85–111.

Staddon, J. E. R. (2014). *The New Behaviorism*. 2nd ed. Psychology Press.

Stein, Janice Gross (2017). "The Micro-Foundations of International Relations Theory: Psychology and Behavioral Economics." In: *International Organization* 71 (S1), S249–S263.

Taylor, Charles (1971). "Interpretation and the Sciences of Man." In: *The Review of Metaphysics* 25 (1), pp. 3–51.

Tversky, Amos and Daniel Kahneman (1973). "Availability: A Heuristic for Judging Frequency and Probability." In: *Cognitive Psychology* 5 (2), pp. 207–232.

Weintraub, E. Roy (2002). *How Economics Became a Mathematical Science*. Duke University Press.

Yarhi-Milo, Keren (2018). *Who Fights for Reputation: The Psychology of Leaders in International Conflict*. Princeton University Press.

4 Can a Leopard Change Its Spots? Implicit Theories and Signaling in International Politics

Abstract

Rationalist scholarship has long assumed that actors in international politics rely on costly signals to reduce uncertainty about motives in international politics and achieve bargaining outcomes short of war. Recently, scholars have begun to test this assumption experimentally. Generally, they induce variation in the cost of a signal by varying the sender's action. However, rationalist scholarship implies that the cost of a signal does not inhere in the sender's actions: it depends on the strategic environment in which the sender's actions are taken. Consequently, to infer motives from signals the way rationalist theories predict, recipients must relate a signal to information about the sender's strategic environment. Integrating rationalist arguments with findings from research on attribution in psychology, I hypothesize that a sender's strategic context impacts the inferences about motives recipients draw from that sender's action, but that the extent to which it does varies depending on recipients' assumptions about agents in international politics and their conceptions of the self. The findings from a survey experiment with over 800 participants from the U.S. and UK, respectively, lend partial support to the costly signaling hypothesis and show that attributional biases found in psychology do not necessarily extend to international politics. However, they also suggest that recipients' attributional patterns systematically vary, with significant implications for the risk of conflict escalation.

4.1 Introduction

A key insight of rationalism is that uncertainty, including uncertainty about motives, plays a central role in determining the outcome of strategic encounters in international politics.[1] In the presence of uncertainty, states that might otherwise choose to cooperate will compete; in the absence of uncertainty, states that might otherwise choose to compete will cooperate.[2] Identifying sources of variation in uncertainty about motives is therefore key to understanding bargaining success and failure in international politics.

Scholarship on signaling has long assumed that actors rely on costly signals to reduce this uncertainty.[3] Costly signals are actions that are less costly for, and thus more likely to be taken by, one type of actor (for example, a greedy state; a resolved state) than another (for example, a security seeker; an unresolved state).[4] By implication, costly signals are inherently credible and help 'type separate' between cooperative and competitive actors. Given incentives for states to conceal or misrepresent information in international politics, including information about their motives, actors must rely on costly signals to find mutually-acceptable agreements and avoid costly arms races or military conflicts.[5]

The costly signaling argument assumes uncertainty about motives to be a property of the international environment and assumes actors to process information about their environment accurately.[6] Harnessing the methodological strengths of experimental designs, scholars have recently begun to investigate this assumption empirically. Drawing on public and elite samples, they systematically test whether individuals respond to costly signals the way that rationalist theories predict. The findings from this emerging body of work provide mixed support for the costly signaling hypothesis. On the one hand, several studies show that costly signals affect individuals' beliefs about a state's motives or resolve the way rationalist theories predict.[7] On the other hand, studies also find that the extent to which they do varies across individuals, depending on their prior beliefs about international politics, the state in question, or the credibility of distinct types of cost.[8] This

1 Glaser 2010; B. C. Rathbun 2007.
2 Glaser 2010; Fearon 1995.
3 Fearon 1997.
4 Of course, this conclusion only holds if the sender is assumed to be a rational utility maximizer.
5 Fearon 1995; Glaser 2010; Kydd 2005.
6 Kertzer, B. C. Rathbun, and N. S. Rathbun 2020.
7 For example, Quek 2021; Kertzer, B. C. Rathbun, and N. S. Rathbun 2020.
8 Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Yarhi-Milo, Kertzer, and Renshon 2018.

suggests that costly signals may be necessary, but are not sufficient, for understanding how actors in international politics form and update their beliefs about a state's motives and reduce uncertainty in strategic interactions.

Indeed, this is the key insight that propels recipient-driven theories of signaling. Recipient-driven theories maintain that actors respond to information about motives selectively, based on their prior beliefs.[9] On this view, uncertainty about motives is a property of the recipient rather than the international environment: whether an individual perceives a signal as credible or not depends, not only on (the cost of) the signal, but also on the individual's beliefs about the sender or about actors in international politics generally. Signals that bolster a recipient's prior beliefs will be perceived as more credible than signals that challenge those beliefs.[10]

However, experimental studies investigating the microfoundations of signaling provide only a partial test of the costly signaling hypothesis: they induce variation in the cost of a signal by varying the sender's actions, yet rationalist arguments about signaling imply that the cost of a signal does not inhere in the sender's actions; it results from a sender's actions *in light of its strategic environment* and is relative to the cost of the sender's alternative courses of action. Consequently, recipients must relate a sender's actions to information about the sender's strategic environment in order to gauge a signal's credibility and update their beliefs about a state's motives in accordance with rationalist predictions. Experimental studies to date do not test whether they do.

In this paper, I address this gap and make two contributions to scholarship on the microfoundations of signaling in international politics. First, I offer a novel experimental test of the costly signaling hypothesis that varies both the sender's actions *and* the sender's strategic environment, shedding light on the attributional processes that underpin costly signaling at the individual level. Second, integrating rationalist arguments with findings from attribution research in social and cultural psychology, I develop hypotheses about who is most likely to respond to information about a sender's strategic environment the way rationalist theories predict, identifying two novel sources of variation in costly signaling – recipients' assumptions about agents in international politics and their conceptions

[9] Holsti 1962; Jervis 1976; Yarhi-Milo 2014.

[10] For recent examples of recent-driven theories and empirical evidence to support them, see Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Yarhi-Milo, Kertzer, and Renshon 2018; Yarhi-Milo 2014.

of the self.

To this end, I conduct a survey experiment investigating individuals' beliefs about a state's motives with respect to a contested body of water. Unlike previous experiments, the vignettes in this experiment induce variation in the cost of the state's signal by varying both its actions (i.e. whether it makes a small or large increase in its naval presence in the body of water) and the strategic environment in which the actions are taken (i.e. whether or not the state faces a threat). This allows me to examine the impact that each constituent component of a signal's cost has on recipients' beliefs about the sender's motives. Since there is no 'view from nowhere' in international politics, and recipients' own role in the scenario may affect their response to the sender's signal, I ran the experiment in two locations: the U.S. and the UK. Whereas the U.S. was party to the scenario described in the vignette, the UK was a third-party observer.

The findings from the experiment show that a state's strategic environment affects the extent to which recipients ascribe the state's sunk-cost signals to its underlying motives: in line with rationalist predictions, recipients ascribed a state's signal of resolve less to expansionist motives, and more to defensive motives, when the state faced an external threat than when its strategic environment was secure. This finding holds even when recipients are members of the state posing the external threat, demonstrating that attributional biases found in interpersonal and intergroup settings in psychology do not necessarily extend to interstate interactions, as some scholars have argued.[11] Yet, the results also suggest that recipients' responsiveness to information about a sender's strategic environment systematically varies depending on their assumptions about agents in international politics and their conceptions of the self, with important implications for the risk of conflict escalation. This finding underscores the import of psychological research on the role of 'implicit theories' in attribution for recipient-driven theories of signaling in international relations.

[11] Jervis 1976; Larson 1997a.

4.2 Costly Signaling and Attribution

Rationalist scholarship has long argued that actors in international politics rely on costly signals to reduce uncertainty about private information, including motives, in international politics and achieve bargaining outcomes short of war.[12] Costly signals are actions that are more likely to be taken by one type of actor than another, and are therefore inherently credible.[13] They solve the problem of 'type separation' that arises from incentives for actors to conceal or misrepresent information in order to negotiate more favorable bargains.[14]

Recently, international relations scholars have begun to test the costly signaling hypothesis systematically using experimental methods. They do so by varying the level or type of cost incurred to the sender. To manipulate the level or type of cost incurred by a signal, experimental studies to date vary the sender's actions. For example, Yarhi-Milo, Kertzer, and Renshon studied the effect of costly signals on individuals' beliefs about an adversary's resolve in a hypothetical dispute scenario. The treatment for this study was the *type of cost* incurred by the adversary's actions: in one treatment group, subjects read that the adversary had mobilized its military, sending a sunk-cost signal which incurs costs ex ante; in the other, subjects read that the adversary had made a public threat, sending a tying-hands signal which incurs costs ex post.[15]

By contrast, in a series of experiments investigating the effect of cost on the perceived credibility of a signal, Quek held a signal's type of cost constant, but varied the *level of cost* incurred by the sender's actions: in one treatment group, the sender sent a signal that incurred no cost or a small cost (for example, a 'not very costly' military mobilization in experiment 3); in the other, the sender sent a signal that incurred a large cost (e.g. a 'very costly' military mobilization in experiment 3).[16] Similarly, in a second series of experiments, Quek examined the effect that the level of cost has on the perceived credibility of four distinct types of signals: sunk-cost signals, tying-hands signals, installment-cost signals, and reducible-cost signals.[17]

[12] Fearon 1995; Glaser 2010.
[13] Fearon 1997; Quek 2021.
[14] Fearon 1995; Morrow 1999.
[15] Yarhi-Milo, Kertzer, and Renshon 2018; Fearon 1997.
[16] Quek 2016.
[17] Quek 2021.

Yet other studies devise treatments that combine different levels and types of cost.[18] The resulting treatments are often more realistic, resembling real-world signals, yet render it impossible to isolate the mechanism that drives the signaling effect.

So far, the results from this growing body of experimental research on the costly signaling hypothesis paint a mixed picture. In accordance with rationalist predictions, several studies find that, compared to no signal, costly signals of reassurance or resolve are effective.[19] This conclusion holds for tying-hands signals and for sunk-cost signals.[20] However, the results from studies examining the effect of different levels of cost are less conclusive: Quek's first series of experiments finds no support for the effect of cost on the perceived credibility of a signal;[21] Quek's more recent series of experiments provides strong support for the conclusion that high-cost signals are more effective than low-cost or costless signals, and confirms this finding across four distinct types of signals.[22] At the same time, and contrary to rationalist predictions, both series of studies suggest that even costless signals of resolve are perceived as credible and can deter challengers effectively.[23] This finding is consistent with the results from a laboratory experiment investigating the effect of cheap talk on potential entrants in entry-deterrence games[24] and recent qualitative and formal research on signaling, which shows the importance of non-costly signals.[25] Finally, several studies find significant heterogeneity in the effect of costly signals, depending on recipients' beliefs about international politics.[26]

These experimental studies make significant headway in understanding the microfoundations of signaling in international politics. However, they offer only a partial test of the costly signaling hypothesis: they manipulate the cost of a signal by varying the sender's actions, yet rationalist arguments about signaling imply that the cost of a signal does not inhere in those actions; it results from a sender's actions *in conjunction with its strategic environment* and is relative to the cost of the sender's alternative courses of action.[27]

[18] For example, Albuyeh and Paradis 2018; Kertzer, B. C. Rathbun, and N. S. Rathbun 2020.
[19] Kertzer, Renshon, and Yarhi-Milo 2019; Yarhi-Milo, Kertzer, and Renshon 2018; Kertzer, B. C. Rathbun, and N. S. Rathbun 2020.
[20] Kertzer, Renshon, and Yarhi-Milo 2019; Yarhi-Milo, Kertzer, and Renshon 2018.
[21] Quek 2016.
[22] Quek 2021.
[23] Quek 2021; Quek 2016.
[24] Tingley and Walter 2011.
[25] Hall and Yarhi-Milo 2012; Kurizaki 2007; Trager 2010; Trager 2017.
[26] Yarhi-Milo, Kertzer, and Renshon 2018; Kertzer, B. C. Rathbun, and N. S. Rathbun 2020.
[27] Glaser 2010; Morrow 1999.

A classic example, the choice to build arms, illustrates this point. Absent an external threat, a greedy state who seeks to challenge the status quo is more likely to bear the cost of building arms than a security seeker who is content with the status quo, and the choice to build arms is a costly signal of resolve that affords conclusions about an actor's motives and helps type separate, i.e. distinguish a greedy state from a security seeker. By contrast, when an external threat is present, a greedy state may be no more or less likely to build arms than a security seeker who fears an attack.[28] In this case, the choice to build arms may signal resolve, but does not type separate between greedy states and security seekers. Crucially, this difference in signaling effect results, not from differences in the sender's course of action itself, but from differences in the sender's strategic environment that alter the cost of that course of action *relative to its alternatives*: faced with a threat, the cost of arming relative to not-arming decreases for security seekers, rendering it difficult or impossible to tell them apart from greedy states. The resulting uncertainty can precipitate downward spirals or what international relations scholars call 'security dilemmas', i.e. situations in which two security seekers engage in costly arms races that could be avoided if they could signal their type credibly.[29]

As this example illustrates, a state's strategic environment is essential to gauging the cost, and thus credibility, of its signals as indicators of the state's motives. By implication, the costly signaling hypothesis assumes that recipients will be responsive to information about a state's strategic environment and will relate the state's actions to this information uniformly, updating their beliefs about a state's motives in response to actions that are costly for malign or benign states *given the state's strategic environment*, but not otherwise. Experimental studies of costly signaling to date do not test whether recipients do.

4.2.1 Attribution and Heterogeneity in Costly Signaling

Relating a state's actions to information about its strategic environment involves what psychologists call 'attribution' – the ascription of behavior to underlying causes.[30] Research on attribution in psychology gives us reason to doubt the realism of the assumption that

[28] Glaser 2010; Kydd 1997.
[29] Jervis 1978; Kydd 2005; Tang 2009.
[30] Heider 1958; Hewstone 1989.

recipients are responsive to information about a sender's strategic environment and relate the sender's actions to that information in the way costly signaling arguments assume they will. In particular, psychologists have shown that individuals tend to make dispositional attributions of behavior, ascribing the behavior to an actor's enduring dispositions, even when situational factors are highly salient. For example, in a famous experiment on attribution, the renown psychologists Edward Jones and Victor Harris found that participants ascribed the political opinion in an article to the author even when the author had no choice in the opinion they expressed.[31] This phenomenon subsequently became known as the 'fundamental attribution error' or 'correspondence bias'.[32]

The correspondence bias has potentially far-reaching implications for international politics: it suggests that recipients may be less responsive to information about a sender's strategic context than rationalist theories assume, undermining costly signals as a mechanism to reduce uncertainty about motives in strategic interactions. As several scholars have argued, this can increase the risk of security dilemmas and conflict escalation between states.[33] For example, Deborah Larson has argued that the fundamental attribution error hampered efforts to increase superpower cooperation during the Cold War by (mis)leading each superpower to attribute the other's actions to mal-intent rather than fear.[34]

However, the tendency to underestimate the role of situational factors is not universal: a large literature in psychology shows that individuals' attributional patterns systematically vary. For example, one classic finding in attribution research known as the 'actor-observer bias' is the tendency for individuals to attribute another person's behavior to her dispositions, but to attribute their own behavior to situational factors.[35] In other words, the correspondence bias is a matter of perspective: individuals are susceptible to it when they are the observer, but not when they are the actor. As Robert Jervis and Deborah Larson have pointed out, this can exacerbate the risk of conflict escalation in international politics because actors may overestimate an adversary's awareness of the situational factors shaping its own policies and underestimate the threat those policies pose to the adversary.[36] Findings from intergroup psychology further compound this concern:

[31] Jones and Harris 1967.
[32] Gilbert and Malone 1995; Ross 1977.
[33] Jervis 1976; Larson 1997a; Rosenberg and Wolfsfeld 1977.
[34] Larson 1997a; Larson 1997b.
[35] Edward and Nisbett 1971.
[36] Jervis 1976; Larson 1997a. Dickson 2009 has recently investigated the role of actor-observer differ-

for example, Donald Taylor and Vaishna Jaggi have shown that participants attribute positive behavior to dispositions, and negative behavior to situational factors, for members of the ingroup, but display the reverse pattern for members of an outgroup.[37] Thomas Pettigrew subsequently dubbed this phenomenon the 'ultimate attribution error'.[38]

Attributional patterns vary, not only within individuals depending on their group membership or perspective, but also across individuals depending on their basic conceptions of social entities. Specifically, psychologists have shown that individuals' attributional patterns are shaped by their self-construal, i.e. their conceptions of the self in relation to others, and what psychologists call 'implicit theories', i.e. individuals' tacit assumptions about the social world.[39] Research on the role of self-construal in attribution originates in studies comparing the behavioral attributions of participants from non-Western and Western societies (e.g. from the U.S. and Japan) in cross-cultural and cultural psychology. These studies find that participants from non-Western societies are less susceptible to the correspondence bias than participants from Western societies.[40] Scholars have argued that these differences are culturally-induced,[41] arising from differences in conceptions of the self in relation to others, among others:[42] interdependent conceptions that emphasize connectedness to others induce greater attention to an actor's social context; independent conceptions that emphasize separateness to others induce greater attention to an actor's enduring dispositions.[43] Scholars have theorized that the correspondence bias is thus attenuated for individuals who hold an interdependent conception of the self compared to individuals who hold an independent conception.[44] Although most studies investigating cultural differences in attribution observe rather than manipu-

ences in attribution in bargaining games. Although his findings support the actor-observer bias, his hypothesized mechanism differs from common explanations of the actor-observer bias in psychology, with unclear implications for Jervis' and Larson's arguments.

37 Taylor and Jaggi 1974.

38 Pettigrew 1979. Note that there is mixed evidence of the error, see Hewstone 1990.

39 Morris, Nisbett, and Peng 1996. Implicit theories are also known as 'lay theories' in psychology (Levy, Chiu, and Hong 2006).

40 Choi and Nisbett 1998; Masuda and Kitayama 2004; Miller 1984; Miyamoto and Kitayama 2002; Morris and Peng 1994; Norenzayan, Choi, and Nisbett 2002.

41 For example, Nisbett 2003.

42 Although differences in attributional patterns and, more broadly, information processing have often been studied by comparing samples from different countries, it is important to note that the culturally-induced differences are not inherently bound to any particular unit and can occur both across and within countries (for example, Henrich, Heine, and Norenzayan 2010; Talhelm et al. 2015; Talhelm 2018).

43 Markus and Kitayama 1991.

44 Markus and Kitayama 1991.

late differences in self-construal, Ulrich Kuehnen and colleagues offer a direct test of the causal role of self-construal in attribution.[45] Their findings substantiate the theoretical claim that conceptions of the self explain systematic variation in individuals' attributional patterns.[46]

A second source of variation in attributional patterns are individuals' 'implicit theories', i.e. their tacit assumptions about the social world.[47] In particular, Carol Dweck and colleagues have shown that individuals' assumptions about the malleability of individual traits shape the way they attribute behavior and form impressions of an actor: individuals who assume traits to be fixed (i.e. who hold an 'entity theory' of personality) are less responsive to situational factors, and make more correspondent inferences even when situational factors are salient, than individuals who assume traits to be mutable (i.e. who hold an 'incremental theory' of personality).[48] While research on the role of implicit theories in attribution has predominantly focused on individuals' assumptions about the individual as an entity, scholars have also used the concept to explore individuals' assumptions about groups.[49] These scholars argue that individuals vary in their tendency to perceive groups as 'entitative', i.e. as 'real', unified and coherent, entities.[50] The extent to which individuals perceive groups as entitative in turn bears on the attribution of group behavior: psychologists have found that the correspondence bias vanishes for aggregates of individuals, but extends to groups perceived as entitative.[51]

In contrast to the actor-observer bias and ultimate attribution error, which give us further reason to doubt the assumption that recipients will respond to information about a sender's strategic environment in the way rationalist theories predict, research on the role of self-construal and implicit theories about individuals and groups in attribution add nuance to this concern: they suggest that the extent to which individuals meet this assumptions will vary depending on their self-construal and implicit theories about individuals and groups. Individuals with an interdependent self-construal, an incremental view of personality, and a less entitative view of states will update their beliefs about

[45] Kuehnen et al. 2013.
[46] Kuehnen et al. 2013.
[47] Molden and Dweck 2006.
[48] Chiu, Hong, and Dweck 1997; Dweck, Hong, and Chiu 1993; Molden, Plaks, and Dweck 2006.
[49] See Levy, Chiu, and Hong 2006; Plaks, Levy, Dweck, and Stroessner 2004; Rydell et al. 2007.
[50] Campbell 1958; Levy, Chiu, and Hong 2006; Rydell et al. 2007.
[51] Yzerbyt, Rogier, and Fiske 1998.

motives more in accordance with the costly signaling hypothesis than individuals with an independent self-construal, an entity view of personality, and an entitative view of states. To examine whether the conclusions individuals draw from signals vary depending on the strategic environment of the sender, and whether individuals vary in how responsive they are to information about a sender's strategic environment, I conducted a survey experiment.

4.3 Research Design

Recent years have seen a surge in empirical scholarship on signaling and belief updating about motives and resolve in international politics. Generally, this scholarship employs either qualitative methods, closely examining the assessment of motives among small groups of elite decision makers,[52] or experiments with large samples drawn from the general public. While the former are well-suited to investigating the explanatory power and scope conditions of microfoundational claims about signaling in real-world settings and to integrating individual-level mechanisms with mechanisms at other levels of analysis in international relations, the latter facilitate the rigorous testing of these microfoundational claims and mechanisms in the first place. As a first direct test of the attributional processes that underpin costly signaling, I therefore conducted an experiment.[53]

The aim of the experiment was twofold: to test whether a sender's strategic environment affects the inferences recipients draw about the sender's motives from the sender's actions and to test whether the impact of a sender's strategic environment varies across individuals depending on recipients' implicit theories and self-construal. Since the study focuses on recipients' perceptions (of signals) and beliefs (about motives) rather than their behavior, I employed a vignette-based survey experiment.

The survey flow is shown in figure 4.1. Subjects first filled in a pre-treatment questionnaire assessing a range of dispositional variables, including their conceptions of the self, implicit theories of individuals and states, as well as a battery of control variables (see 4.4.3).[54] Following the pre-treatment questionnaire, subjects were told that they would

[52] For example, Hall and Yarhi-Milo 2012; Yarhi-Milo 2014; B. C. Rathbun 2019.

[53] The design, hypotheses, and analysis for this study were pre-registered on OSF on February 18, 2022. An amendment was submitted on February 24, 2022, and another minor amendment on February 28, 2022. Pilot data collected prior to February 24 are not included in the analysis.

[54] To assess the dispositional variables, I rely on standard measures from psychology. Wherever possible I

be asked to read a vignette describing a scenario in international politics that could occur in the future. Following this introduction, complete randomization was used to allocate subjects to one of four treatment groups.

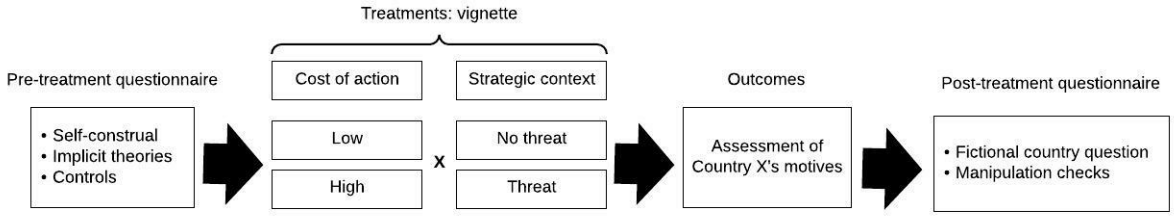


Figure 4.1: Design and survey flow

All four treatment groups read the same basic scenario. To abstract away from current events or actors in international and domestic politics, the scenario was set in the future, in 2035,[55] and described a situation involving the U.S. and 'Country X', who share access to 'a body of water'. Keeping the scenario abstract and general allowed me to manipulate the U.S.' and Country X's actions plausibly while anchoring subjects' beliefs about theoretically salient features of the scenario. In particular, both countries were described as having strong strategic and economic interests, and maintaining a strong naval presence, in the body of water to anchor subjects' beliefs about each party's stake in the issue, as well as the local balance of power. The vignette also notes that the U.S. and Country X 'regularly face tensions over other issues in international politics', implying an adversarial relationship. At the same time, the status quo with respect to the body of water was described as stable: 'neither country has changed its naval deployment to the body of water in past years'. This element in the description is important because it establishes a status quo that minimizes incentives for security seekers – whether trusting or fearful – to build arms.[56] Thereby, it helps generate clear predictions about the inferences 'rational' subjects will draw from competitive and costly actions on either side: if one of the two countries initiates a challenge to the status quo by increasing its military presence in the body of water, it is likely to be greedy.

use short versions that have been validated in previous studies. Specifically, I used an updated version of Singelis 1994's original measure to assess self-construal (the adapted version was used in Na et al. 2010 and is available at `https://uwaterloo.ca/wisdom-and-culture-lab/measures`); I used Levy, Stroessner, and Dweck 1998's three-item measure to assess subjects' implicit theories of personality; I adapted M. B. Brewer, Hong, and Li 2004's very general, two-item measure of entitativity to assess subjects' implicit theories of states.

[55] This strategy has been used in previous experiments, for example, Mattes and Weeks 2018.

[56] Kydd 1997.

To test whether a sender's strategic environment affects the inferences recipients draw from its actions in the way costly signaling predicts it will, the experiment crossed two treatments with two levels, respectively:

1. Cost of Country X's action: whether Country X made a small or large increase in its naval deployment to the body of water (low-cost versus high-cost condition)

2. Strategic context: whether Country X initiated a challenge to the status quo or responded to a challenge by the U.S. (no-threat versus threat condition)

This resulted in a 2x2 factorial design. By varying the cost of Country X's action, in addition to the strategic environment in which Country X took the action, I was able to include a replication element in the design that increases the experiment's comparability to previous studies on costly signaling,[57] and offers a more complete test of rationalist predictions.

The treatments were induced by adding one or two sentences to the vignettes, respectively. In the no-threat conditions, the vignette concluded with the sentence: 'Yesterday, Country X decided to make a small [large] increase in its naval deployment to the body of water'. In the threat conditions, this sentence was preceded by another: 'In recent weeks, the U.S. has decided to make a large increase in its naval deployment to the body of water' (see appendix 4.7.2). To isolate the effects of cost and strategic context on respondents' assessments of motives, it was important to hold the type of signal constant across all treatment groups. As an instance of sunk-cost signaling, naval deployments fulfilled this criteria: unlike a public threat, for example, the cost of a naval deployment can easily be altered without altering the type of the signal or otherwise adding complexity, and introducing possible confounds, to the treatment.[58] Moreover, sunk-cost signals have previously been shown to impact assessments' of an adversary's resolve or the credibility of a threat.[59] If the experiment can replicate this finding, it provides a clear baseline for assessing the moderating role of strategic context in costly signaling.

[57] In particular, this design increases the comparability to Quek 2021, which compared the effect of high-cost to low-cost signals.

[58] Strictly speaking, naval deployments and other forms of mobilization are not pure sunk-cost signals: they combine sinking costs with tying hands, see Fearon 1997; Slantchev 2005. However, the key point here is that the type of signal must be constant, not pure. Since real-world signals are rarely pure, this offers the additional advantage of allowing us to use a more concrete and tangible example.

[59] Kertzer, Renshon, and Yarhi-Milo 2019; Quek 2021; Yarhi-Milo, Kertzer, and Renshon 2018.

One final feature of the scenario is worth noting: unlike 'Country X', the actor identity of the second country, the U.S., was specified rather than abstracted. This choice was made because there is no 'view from nowhere' in international politics: respondents necessarily belong to a state, and psychological research suggests that group membership, as well as recipients' perspective as an actor or observer, may affect the way individuals attribute behavior.[60] Making this fact an explicit feature of the experimental design has two advantages. First, it allows me to test for the role of group membership and actor-observer differences in the attributional processes that underpin costly signaling by running the experiment in two locations, the U.S. (primary sample) and the UK. Whereas subjects in the U.S. sample were party to the dispute described in the scenario, subjects in the UK sample were a third-party observers. Both the actor-observer bias and ultimate attribution error lead us to expect that subjects from the U.S. will respond to changes in Country X's strategic environment less than subjects from the UK, because the U.S.'s actions are the source of threat to Country X in the threat condition (see section 4.3.1). Second, this design choice increases the substantive relevance of the experiment by incorporating a key feature of security dilemmas in international politics: the recipients of signals (here, subjects in the U.S. sample) are also party to the dispute. Investigating how this affects signaling is thus vital to understanding the microfoundations of conflict escalation between states.

After reading the vignette, subjects were asked to answer several questions relating to Country X's behavior:

- 'Given the information available, how likely or unlikely do you think it is that Country X's behavior is driven by an underlying desire to expand its access to, and control over, the body of water?'

- 'Given the information available, how likely or unlikely do you think it is that Country X's behavior is driven by a desire to defend its access to, and control over, the body of water?'

- 'Given the information available, how likely or unlikely do you think it is that Country X would increase its naval deployment to the body of water if the U.S.

[60] Hewstone 1990; Edward and Nisbett 1971; Rosenberg and Wolfsfeld 1977.

made a large decrease in its naval deployment?'

All three outcome measures assessed subjects' beliefs about Country X's motives based on its behavior in the scenario, providing insights into subjects' attributions. Specifically, the first question measures the extent to which subjects make correspondent inferences, attributing observed behavior to underlying motives; the second question measures the extent to which subjects make situational inferences, attributing observed behavior to the strategic context;[61] the third question measures the extent to which subjects predict Country X to behave the same way across situations, again implying a correspondent inference. A 7-point scale ranging from 0 to 6 was used to measure all three items.

Additionally, I included a fourth outcome measure assessing subjects' beliefs about Country X's resolve: 'Given the information available, how likely or unlikely do you think it is that Country X will stand firm in a dispute over access to the body of water with the U.S.?' This question was included as a replication element that increases the comparability of results from this study to previous studies.[62] However, it is important to note that this question is broader than the outcomes of primary interest and could be more susceptible to confounding.[63]

Finally, following the outcome measures, subjects were asked to answer three manipulation/attention check questions, as well as indicate how closely they have been following the evolving situation between Russia and Ukraine (see appendix section 4.7.6).

[61] Within psychology, the dispositional-situational dichotomy that underpins attribution research has been criticised. One critique is that the dichotomy is superficial: for every situational cause, there exists a dispositional counterpart (for example, wanting to avoid embarassment or punishment could be internal reasons for receptivity to external pressure, see Sabini, Siepmann, and Stein 2001). This would not necessarily invalidate the findings from attribution research, but it would change how we interpret them substantively. Here, a dispositional wording of the situational attribution is adopted to avoid this potential concern.

[62] Experiments designed to test the costly signaling hypothesis to date have focused on resolve or the credibility of a threat (Kertzer, Renshon, and Yarhi-Milo 2019; Quek 2016; Quek 2021; Yarhi-Milo, Kertzer, and Renshon 2018)

[63] While beliefs about motives is an important factor in assessments of resolve, so are material variables that affect the cost of war or probability of victory (Glaser 2010; Kertzer 2016). As sunk-cost signals can have both an informational effect, leading to updated beliefs about a sender's motives, as well as a functional effect, resulting in changes to the balance of power (Slantchev 2005), subjects' beliefs about the cost of war or probability of victory might change along with beliefs about the sender's motives. If so, the assumption of information equivalence would be violated (see Dafoe, Zhang, and Caughey 2018).

4.3.1 Hypotheses

Treatment Effect

Rationalist theories tell us that the effect signals have on recipients' beliefs about the sender's motives depends on the cost they incur to the sender. A signal's cost in turn depends on the sender's actions *and* the strategic context in which those actions are taken, giving rise to the following hypotheses:

Hypothesis 1 *Respondents will estimate the likelihood that Country X's behavior is driven by expansionism, and is consistent across strategic contexts, to be higher when Country X initiates a challenge than when it responds to a challenge of the status quo (i.e.* $\mu_{1,2} > \mu_{3,4}$*).*

Hypothesis 2 *Respondents will estimate the likelihood that Country X's behavior is driven by defensiveness to be higher when Country X responds to a challenge than when it initiates a challenge of the status quo (i.e.* $\mu_{3,4} > \mu_{1,2}$*).*

Hypothesis 3 *In the absence of a threat, respondents will estimate the likelihood that Country X's behavior is driven by expansionism, and is consistent across strategic contexts, to be higher when the cost of Country X's action is higher (i.e.* $\mu_2 > \mu_1$*).*

Hypothesis 4 *The effect of cost of Country X's action on respondents' estimates of the likelihood that Country X's behavior is driven by expansionism, and is consistent across strategic contexts, will be greater when Country X initiates a challenge than when it responds to a challenge of the status quo (i.e.* $\mu_2 - \mu_1 > \mu_4 - \mu_3$*).*

While a state's strategic environment should alter the cost of adopting competitive military policies for benign versus malign states, we would not expect it to have the same effect on the strategic calculus of resolved versus unresolved states, since both benign and malign states can be more or less resolved. The hypothesis for the replication element of this study thus is:

Hypothesis 5 *Respondents will estimate Country X's resolve to be higher when the cost of Country X's action is higher (i.e.* $\mu_{2,4} > \mu_{1,3}$*).*

Attribution research in psychology suggests that recipients may not respond to information about a sender's strategic environment the way rationalist theories predict, particularly when recipients' own state poses the external threat in question (and recipients thus attribute the threat, but not the response, to situational factors). This gives rise to the following hypothesis:

Hypothesis 6 *The effect of strategic context will be greater for respondents from the UK than for respondents from the U.S.*

Heterogeneity in Treatment Effect

In contrast to rationalist theories, attribution research in psychology suggests that recipients will vary in how responsive to a sender's strategic environment they are depending on their implicit theories and self-construal. This gives rise to the following hypotheses:

Hypothesis 7 *The effect of strategic context will be greater for respondents low on entity theory than for respondents high on entity theory.*

Hypothesis 8 *The effect of strategic context will be greater for respondents with a less entitative theory of countries than respondents with a more entitative theory.*

Hypothesis 9 *The effect of strategic context will be greater for respondents who report a more interdependent self-construal.*

4.4 Results

4.4.1 Sample

The experiment was fielded between February 24 and March 18 2022 in the U.S. and between April 24 and May 4 2022 in the UK.[64] It was administered via the online survey platform Qualtrics, and subjects were recruited via the recruitment platform Lucid. All subjects who currently live in the U.S. or UK, respectively, and are 18 years or older were eligible to participate. The survey took 10 to 15 minutes to complete.[65] Subjects were reimbursed in accordance with guidelines of the recruitment platform.

[64] For both samples, data collection included a 'soft launch' with approximately 10% of the final sample (N=86 in the U.S. and N=114 in the UK).

[65] The median completion time was 12.33 minutes in the U.S. and 10.88 minutes in the UK.

With the growth of experimental research in political science, online convenience samples have become an increasingly popular and important tool for scholars. In the present study, the use of non-probability samples was justified on two grounds. First, the aim of the study is to establish the relationship between variables – the causal effect of the treatment on outcomes as well as the moderating effect of pre-treatment covariates – rather than to draw descriptive inferences about the population.[66] Second, the theory advanced here should, in principle, apply to all adult humans. Consequently, the two samples under investigation are theoretically relevant.[67] At the same time, the theory hypothesizes treatment effect heterogeneity. Online convenience samples are therefore especially well-suited for testing the theory:[68] albeit less representative than national probability samples, they are more heterogeneous than local convenience or student samples, which have long been the backbone of experimental research in political science and psychology.[69] Lucid in particular automatically applies quotas for age, gender, region and ethnicity to obtain a more heterogeneous and representative sample.[70]

Since the aim of this study is to establish the relationship between variables, the internal validity of the study is key to its conclusiveness. The use of online convenience samples raises two potential concerns about internal validity. One concern is that subjects might participate in a study more than once, thereby violating treatment assignment. I relied on a range of technical tools available on Lucid and Qualtrics, respectively, to automatically detect and prevent repeat survey-taking during data collection. A second concern is that subjects may be inattentive and not read instructions or questions carefully before responding, thereby reducing data quality. Indeed, several scholars have reported a recent decline in attentiveness among Lucid subjects.[71] As a precaution, I therefore included an attention check at the beginning of the survey.[72] Subjects who failed the attention check were not eligible to participate. In the U.S. sample, 29.57% of the subjects

[66] Baker et al. 2013.
[67] Coppock and McClellan 2019.
[68] Druckman and Kam 2011.
[69] Berinsky, Huber, and Lenz 2012; Henrich, Heine, and Norenzayan 2010.
[70] Quotas for ethnicity are only applied in the U.S.
[71] Ternovski and Orr 2022; Peyton, Huber, and Coppock 2021. More generally, there have been concerns about the impact of the global pandemic on survey responses. Investigating these concerns empirically, Peyton, Huber, and Coppock 2021 replicate 12 pre-pandemic experimental designs with COVID era samples.
[72] The wording of the attention check item was adapted from Shamon and Berning 2020. The item read: 'On a scale from 0 to 6, please indicate whether the following statement is true: I was born before 1920. Your response reassures us that you are following closely.'

who agreed to participate in the experiment failed the initial attention check (see appendix, figure 4.10). In the UK, the percentage was somewhat lower (see appendix, figure 4.11). Additionally, subjects were asked to recall two key features of the scenario at the end of the survey.

The target sample size for this study was 836 participants for each country.[73] In order to obtain sufficiently large samples, I had to slightly oversample. The actual sample sizes, and number of subjects assigned to each treatment condition, are shown in tables 4.1 and 4.2. The probability of assignment to treatment condition was the same across countries.

	No-threat	Threat
Low-cost signal	225 (209)	225 (209)
High-cost signal	225 (209)	225 (209)

Table 4.1: U.S. sample size and randomization. Target shown in brackets.

	No-threat	Threat
Low-cost signal	254 (209)	255 (209)
High-cost signal	255 (209)	255 (209)

Table 4.2: UK sample size and randomization. Target shown in brackets.

Demographically, the U.S. sample tracks national probability samples well: the sample was 52.44% female (compared to 52% in the ANES 2020), 73.5% white (compared to 66% in the ANES 2020), 14.22% Hispanic (compared to 14% in the ANES 2020), 36.67% of subjects were from the South (compared to 31% in the ANES 2020), 26% identified as Republican and 36.11% as Democrat (compared to 32% and 35% in the ANES 2020, respectively), 35.88% at least had a Bachelor's degree (compared to 37% in the ANES 2020), and the mean age was 46.8 years.[74]

4.4.2 Randomization and Balance Checks

Simple randomization was used to allocate subjects to treatment groups. The advantage of randomization is identification: if subjects are allocated randomly, we can assume treatment groups to be identical prior to treatment and, consequently, ascribe any difference in

[73] I used G*Power to conduct a power analysis for a one-tailed independent samples t-test (Faul et al. 2007). My goal was to obtain .80 power and to detect an effect size d of 0.244 at the alpha error probability of .05. The estimate for the effect size is based on a total of six survey experiments that examine the effect of costly signals on assessments of credibility in international politics (Quek 2016; Quek 2021). Crucially, unlike other experimental studies of costly signaling, these six experiments compare the effect of high-cost to low-cost signals. Thus, the effect sizes are most comparable to the ones under investigation in this study. Based on these numbers, the estimated sample size is 209 respondents per group.

[74] American National Election Studies 2021. The UK sample was 49.8% female, 90.5% white, 34.1% identified as Conservative, 43.6% as Labour and 7.9% as Liberal, 34.25% at least had a Bachelor's degree, and the mean age was 45.

the average outcome across groups to the treatment.[75] However, despite randomization, imbalances on demographic traits or pre-treatment variables can occur due to (i) chance, (ii) attrition (if a non-random subgroup drops out of the survey), or (iii) error in the implementation of the randomization procedure. Therefore, it is useful to conduct balance checks.[76] Balance tables revealed only small imbalances across treatment groups in both samples. For the U.S. sample, the largest imbalances occurred for militant internationalism ($SDM = .23$) and independent self-construal ($SDM = .21$). A visual summary of all results from the balance tables is included in appendix section 4.7.4.[77] Additionally, following recommendations by Alan Gerber and Donald Green, I used randomization inference to test the hypothesis that the observed imbalance was not greater than we would expect by chance.[78] For both samples, the results provide no evidence that the observed imbalance was greater than we would expect by chance.[79] Since randomization is implemented automatically in Qualtrics and attrition rates across treatment groups were similar (see appendix section 4.7.3), this result is unsurprising.[80]

4.4.3 Analysis

The analysis proceeds in two parts. The first part examines the effect that strategic context and the cost of Country X's actions have on respondents' inferences about Country X's motives and resolve (hypotheses 1-6). Because strategic context and the cost of Country X's actions are randomized (treatment-by-treatment design), confounds are ruled out by design and we can rely on simple difference-in-means analysis to identify the hypothesized effects.[81] The second part investigates the moderating effect of implicit theories and self-construal on respondents' assessments of Country X's motives (heterogeneity in treatment effect; hypotheses 7-9). Because implicit theories and self-construal are observed rather than randomized (treatment-by-covariate design), this part is quasi-experimental and the analysis more closely follows the logic of observational research, where confounds must be

[75] However, randomization does not eliminate concerns about omitted variable bias in all cases (e.g. Dafoe, Zhang, and Caughey 2018).
[76] A. S. Gerber and Green 2012.
[77] All data analyses were conducted in R (R Core Team 2013).
[78] A. S. Gerber and Green 2012.
[79] See appendix section 4.7.4, figures 4.16 and 4.17.
[80] Mutz and Pemantle 2015.
[81] I use a regression model to test for the hypothesized interaction between the cost of the sender's action and the sender's strategic context.

controlled for.[82] For ease of interpretation, I include visuals summarizing key findings from the analysis. All visuals are based on linear regression models. Any divergences between the regression results and results from the difference-in-means analysis are flagged. Unless explicitly noted, the results reported in the main text are based on the U.S. sample.

Experimental Analysis

Responses to the outcome variables and the manipulation check were not normally distributed (p ¡.05 for all Shapiro-Wilk normality tests). Therefore, a two-sample Wilcoxon test was used to compare group means. First, I compared responses to the manipulation check question, 'In the scenario you just read, how constrained was Country X in its strategic options?', in the threat versus no-threat conditions. The difference between groups is highly significant ($p < .01, r = 0.138$).[83] This finding suggests that the second treatment, strategic context, succeeded at manipulating respondents' perception of the situational constraints faced by Country X.[84]

Next, I examined the effect of strategic context on the inferences about motives recipients draw from Country X's signal. A Wilcoxon test indicated that subjects randomly assigned to the no-threat conditions perceived Country X's behavior to be driven by expanionism more than subjects assigned to the threat conditions ($p < .0001, r = 0.228$). Mirroring this result, subjects assigned to the threat conditions perceived Country X's behavior to be driven by defensive motives more than subjects assigned to the no-threat conditions ($p < .0001, r = 0.163$; see figure 4.4a). Finally, subjects in the no-threat conditions predicted greater cross-situational consistency in Country X's behavior than subjects in the threat conditions ($p < .01, r = 0.0802$). These findings provide strong support for hypotheses 1 and 2, and show that the strategic context faced by a sender impacts the inferences recipients draw from the sender's actions: in line with rationalist predictions, recipients attribute a sender's sunk-cost signal to underlying motives more when the sender initiates a challenge to the status quo than when it responds to a challenge.

At the core of this study lies the rationalist argument that the costly signaling mechanism hinges on the interaction between a sender's action and a sender's strategic envi-

[82] Kam and Trussler 2017.
[83] All Wilcoxon tests are one-tailed.
[84] This finding was replicated with the UK sample ($p < .001, r = 0.133$).

ronment: since the cost of a signal does not inhere in the sender's actions but is relative to the cost of alternative courses of action, recipients must relate the sender's actions to its strategic environment in order to draw inferences about the sender's motives. In other words, the strategic constraints faced by a sender should *moderate* the inferences recipients draw from the sender's actions (hypotheses 3 and 4).

To test this, it is useful, first, to show that the present study can replicate a key finding from previous experimental research on costly signaling: the finding that the cost of a sender's actions impacts recipients' beliefs about the sender's motives and resolve.[85] Focusing on resolve, this study replicates this finding (hypothesis 5): a Wilcoxon test indicated that subjects assigned to the high-cost conditions perceived Country X as more resolved than subjects assigned to the low-cost conditions (see figure 4.4b). Although the effect is small, it is highly significant ($p < .01, r = 0.0835$).[86] This result is important: it adds to the growing body of evidence on costly signaling, suggesting that the effect costly signals have on perceptions of resolve is robust to differences in the precise scenario and wording of treatments across experimental designs, and it increases our confidence in the successful manipulation of the first treatment in the present design: the cost of Country X's signal.

With this in mind, we can turn to the moderating effect of strategic context on the relationship between the cost of an action and assessments of the sender's motives. In line with hypothesis 3, the experiment finds that subjects in the no-threat conditions attributed Country X's behavior more to underlying expansionism, and predicted Country X to behave more similarly across strategic contexts, when Country X's action carried a high cost than when it carried a low cost ($p < .0001, r = 0.197$ for Country X's expansionism; $p < .05, r = 0.0797$ for predicted consistency). In line with the theoretical prediction from hypothesis 4, however, this effect turned insignificant in the threat conditions ($p > .05, r = 0.0132$ for perceived expansionism; $p > .05, r = 0.015$ for predicted, cross-situational consistency of Country X's behavior). These findings show that the conclusions recipients draw about a sender's motives crucially depend on the strategic context in which a costly

[85] Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Quek 2021.

[86] For replication purposes, this study focuses on resolve, because it provides the closest – and therefore most informative – point of comparison between this study and previous studies. An additional advantage in the present set-up is that, unlike motives, we would expect perceptions of resolve to be impacted by the cost of a signal, but not by strategic context.

action is taken.

To gain a more complete picture of the interaction between the cost of a sender's action and strategic context, I additionally examined the moderating effect of cost on the relationship between strategic context and assessments of a sender's motives. A Wilcoxon test indicated that subjects in the no-threat conditions attributed Country X's behavior to expansionism more than subjects in the threat conditions, both when Country X's behavior carried a high cost *and* when it carried a low cost. However, the effect size almost doubled when Country X's behavior incurred a high cost compared to when it incurred a low cost ($p < .01, r = 0.148$ in the low-cost conditions compared with $p < .0001, r = 0.304$ in the high-cost conditions): in the low-cost conditions, a strategic threat decreased attributions of Country X's actions to underlying expansionism by 7.23%; in the high-cost conditions, it decreased them by 14.9%. Interestingly, the effect of strategic context on subjects' behavioral predictions was only significant when Country X's action carried a high cost ($p < .01, r = 0.125$), not when it carried a low cost ($p > .05, r = 0.0338$). This highlights the interaction between cost and strategic context. Figures 4.2 and 4.3 visually summarize these findings.[87] Adjusting for pre-treatment covariates did not substantially change the results (see table 4.3).

Two things are worth noting. First, while the effects of an action's cost and strategic context on recipients' behavioral predictions are broadly consistent with hypotheses 1, 3 and 4, they are substantially smaller than the effects of cost and strategic context on perceptions of Country X's expansionism. For example, in the absence of a threat, an action that carries a high cost increases perceived expansionism by 8.66% compared to an action that carries a low cost, but only increases the estimated likelihood that the sender would behave similarly across strategic contexts by 4.41%. Similarly, when Country X sends a high-cost signal, recipients perceive Country X to be 14.9% less expansionist when it responds to a threat than when it initiates a threat, but only estimate it to be 8.07% less likely to behave similarly across strategic contexts.

Second, for recipients' behavioral predictions, the results from the multiplicative regression models shown in table 4.3 differed slightly from the Wilcoxon test results: whereas the latter indicated that the cost of Country X's action significantly increases subjects'

[87] Interaction models cannot be interpreted readily based on regression tables alone, see Brambor, Clark, and Golder 2006.

predictions of cross-situational consistency in the no-threat conditions,[88] the former suggested that the marginal effect of cost on behavioral predictions is insignificant in the threat *and* no-threat conditions (also see figure 4.3).[89] Given the small effect of cost on recipients' behavioral predictions in the no-threat conditions, this difference is likely due to differences in power between the means-difference analysis and multiplicative regression model.[90] Given the significant Wilcoxon test result, as well as the substantively and statistically significant moderating effect of cost on the relationship between strategic context and behavioral predictions in both models, the findings nonetheless suggest a theoretically-relevant interaction between the cost of a sender's action and the sender's strategic environment.

Taken together, the results from the U.S. sample replicate an important finding from previous experimental research on signaling: recipients are more responsive to a sender's signal when it incurs a high cost than when it incurs a low cost. Importantly, however, this study shows that the conclusions recipients draw from the sender's actions depend on the strategic context in which they are taken: recipients attribute a sender's actions to underlying motives more when the sender initiates a challenge than when the sender responds to a challenge, even when their own country poses that challenge. This finding is consistent with the costly signaling hypothesis and suggests that attributional biases found in interpersonal and intergroup settings in psychology do not necessarily extend to strategic interactions in international politics.

That said, the experimental findings from the UK sample partially support the idea that the effect of strategic context on recipients' inferences about Country X's motives depend on their group membership and/or role in relation to the dispute scenario (hypothesis 6): although the effect of strategic context on behavioral predictions was similar across samples ($p < .001$, $r = 0.0974$ for the UK sample), its effect on perceptions of Country X's expansionism ($p < .001$, $r = 0.283$) and defensiveness ($p < .001$, $r = 0.241$) was slightly larger in the UK sample than in the U.S. sample.[91] Interestingly, and contrary

[88] Using randomization inference, I additionally calculated a p-value for the relevant Wilcoxon test statistic. The distribution from 10,000 simulations of the randomization procedure suggest that we can reject the null hypothesis of no effect (under the sharp null of null effect, the likelihood of obtaining a statistic at least as low as the observed statistic is ¡.05).

[89] Although note that it only narrowly misses significance at the .05 level for a one-tailed test and .1 level for a two-tailed test ($t = 1.636$.)

[90] Results from a simple t-test are consistent with the Wilcoxon test results.

[91] A two-sample Wilcoxon test comparing responses to the manipulation check question for UK subjects

to the costly signaling hypothesis, the UK study did not replicate the finding that the cost of Country X's action bears on recipients' assessments of Country X's resolve or motives: the cost of Country X's action had no significant effect (see appendix section 4.7.5).[92] This finding mirrors the results from previous studies, which provide mixed support for the effect of costly signals on perceptions of resolve in international politics.[93]

in the threat versus no-threat conditions indicated that Country X's strategic context was successfully manipulated ($p < .001$, $r = 0.133$), replicating the finding from the U.S. sample.

[92] Unsurprisingly, given this result, the interaction models did not perform better than the additive models for perceived expansionism and predicted cross-situational consistency (see appendix section 4.7.5). However, it is interesting to note that the effect of strategic context on perceptions of expansionism was only larger in the UK sample than the U.S. sample when the cost of Country X's action was low, not when it was high (see appendix, figure 4.20).

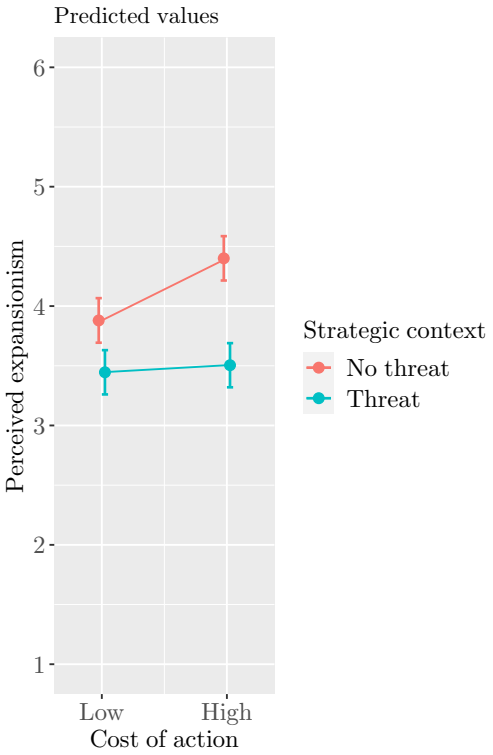
[93] Quek 2016; Quek 2021.

Please do not circulate or cite.

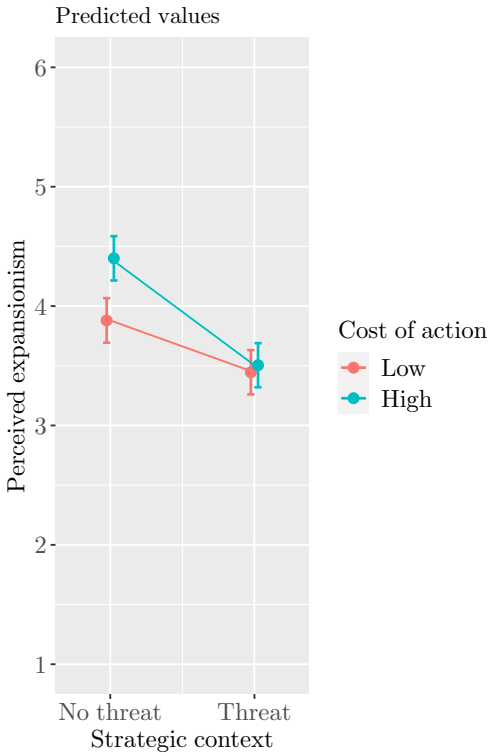
Table 4.3: Treatment effects with and without covariate adjustment

	Perceived expansionism		Predicted consistency		Perceived defensiveness		Perceived resolve	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sunk cost: high	0.520***	0.467***	0.265	0.217	0.302***	0.251***	0.238***	0.200**
	(0.134)	(0.128)	(0.162)	(0.157)	(0.088)	(0.085)	(0.088)	(0.086)
Context: threat	−0.434***	−0.437***	−0.124	−0.146	0.420***	0.415***	0.047	0.051
	(0.134)	(0.128)	(0.162)	(0.157)	(0.088)	(0.085)	(0.088)	(0.086)
Militant internationalism		0.777***		0.608**		0.625***		0.461**
		(0.238)		(0.290)		(0.222)		(0.225)
Political ideology: social		0.222		0.394*		0.284*		0.415***
		(0.165)		(0.202)		(0.154)		(0.156)
International trust		0.649***		1.019***		0.418***		0.382**
		(0.170)		(0.208)		(0.159)		(0.161)
Right-wing authoritarianism		0.500*		0.253		−0.215		−0.058
		(0.275)		(0.336)		(0.257)		(0.260)
Need for closure		1.448***		1.297***		1.728***		1.345***
		(0.317)		(0.387)		(0.296)		(0.300)
Need for cognition		0.637**		0.482		0.906***		0.882***
		(0.260)		(0.317)		(0.242)		(0.245)
High cost*threat	−0.461**	−0.465**	−0.357	−0.349				
	(0.189)	(0.181)	(0.228)	(0.221)				
Constant	3.880***	1.508***	3.716***	1.500***	3.765***	1.590***	4.122***	2.156***
	(0.095)	(0.291)	(0.115)	(0.356)	(0.077)	(0.269)	(0.077)	(0.272)
N	841	841	841	841	841	841	841	841
R^2	0.072	0.162	0.012	0.081	0.039	0.129	0.009	0.081
Adjusted R^2	0.068	0.153	0.008	0.071	0.037	0.121	0.007	0.072

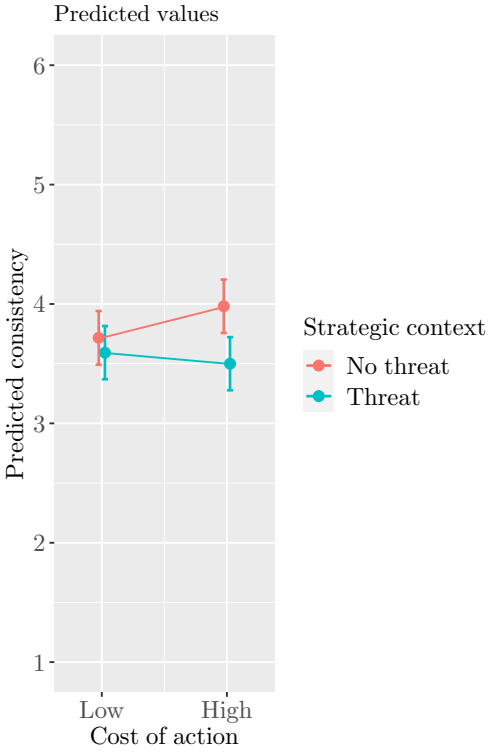
*p < .1; **p < .05; ***p < .01



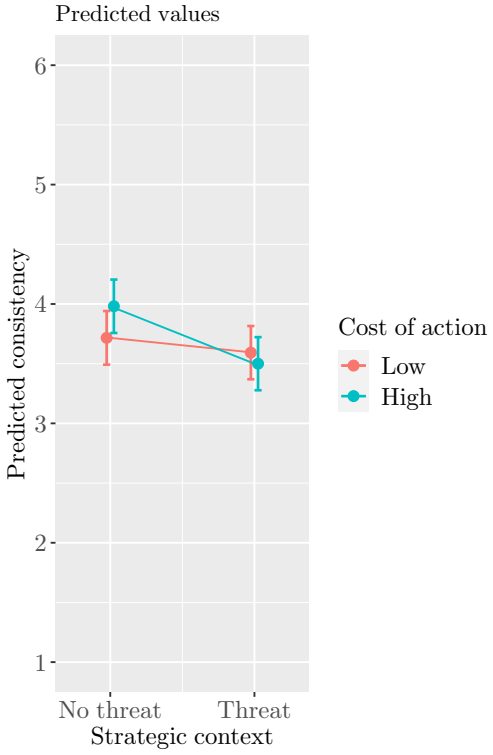
(a) The effect of cost of action on perceived expansionism, by strategic context.



(b) The effect of strategic context on perceived expansionism, by costliness of action.

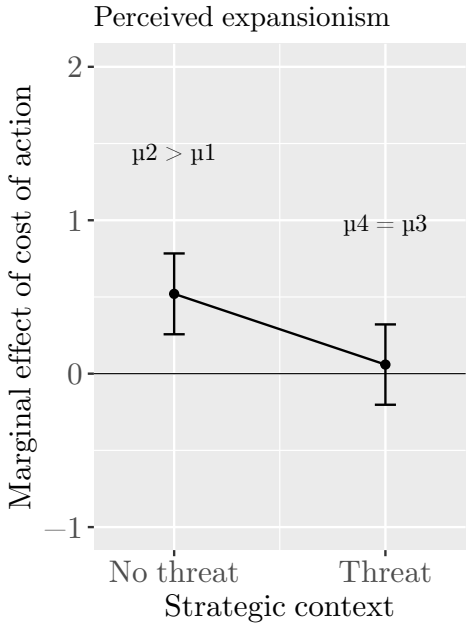


(c) The effect of cost of action on predicted consistency, by strategic context.

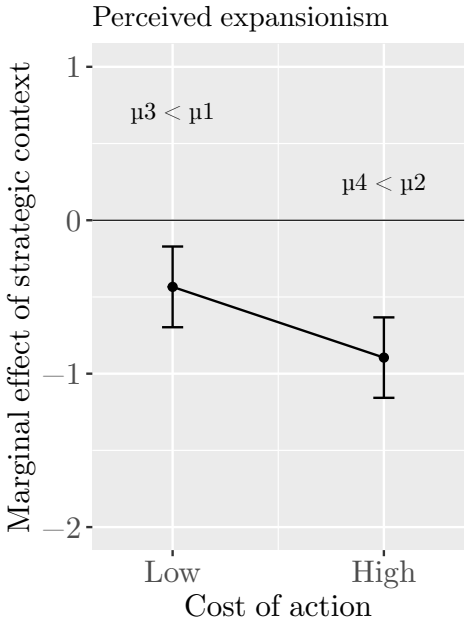


(d) The effect of strategic context on predicted consistency, by costliness of action.

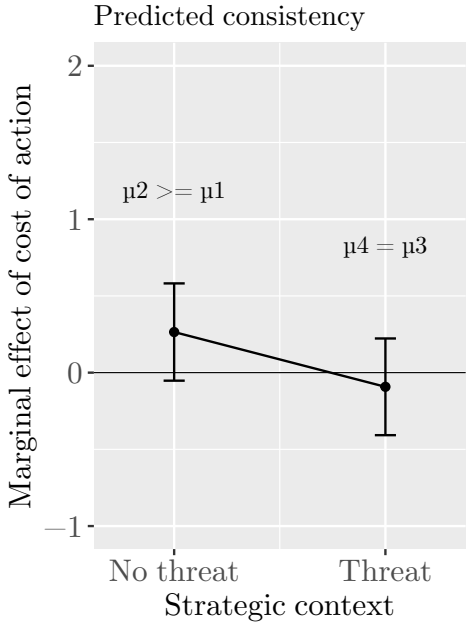
Figure 4.2: The effect of cost of action and strategic context on perceived expansionism and behavioral predictions based on the models without covariate adjustment shown in table 4.3.



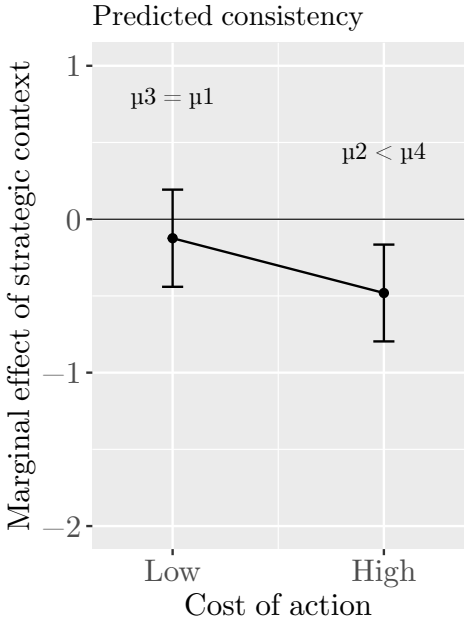
(a) The effect of cost on perceived expansionism is positive and moderated by strategic context.



(b) The effect of strategic threat on perceived expansionism is negative and moderated by cost.

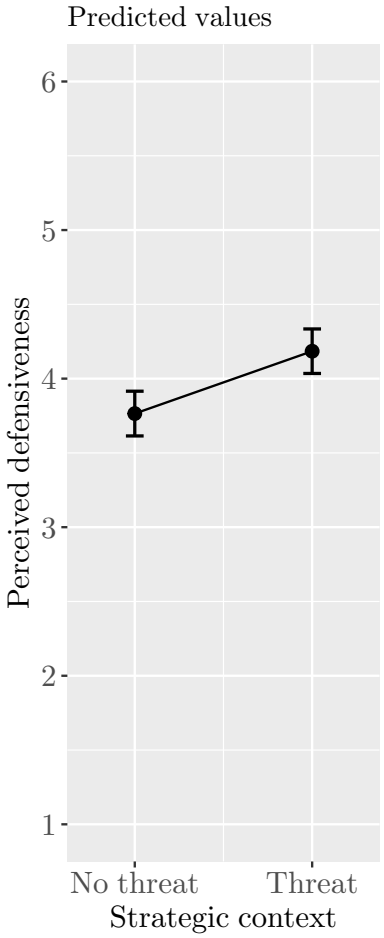


(c) The effect of cost on predicted consistency is positive and moderated by strategic context.

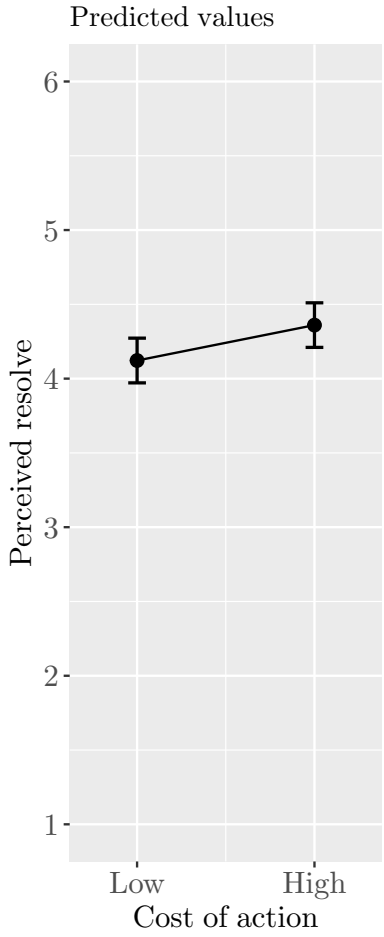


(d) The effect of strategic threat on predicted consistency is negative and moderated by cost.

Figure 4.3: The marginal effect of each treatment on perceived expansionism and behavioral predictions based on the models without covariate adjustment (see table 4.3). The error bars indicate two-tailed, 95% confidence intervals. The superscripts link each estimate to the hypotheses in section 4.3.1. Note that the result in (c) deviates from the result reported in the main text: in the multiplicative regression model, the effect of cost is insignificant in the threat *and* no-threat conditions; in the one-tailed Wilcoxon test, the effect of cost in the no-threat conditions is small but significant at the 95% confidence level.



(a) The effect of strategic threat on perceptions of Country X's defensiveness.



(b) The effect of sunk cost on perceptions of Country X's resolve.

Figure 4.4: The effect of strategic context on perceived defensiveness and resolve based on the models without covariate adjustment shown in table 4.3.

Experimental-Observational Analysis

To investigate implicit theories and self-construal as sources of heterogeneity in signaling (hypotheses 7-9), I include a three-way interaction between the two treatments and entity theory, entitativity, and self-construal, respectively, in the linear regression models estimating perceptions of Country X's expansionism and behavioral consistency; I include a two-way interaction between strategic context and entity theory, entitativity, and self-construal, respectively, in the linear regression model estimating perceptions of Country X's defensiveness.[94] Since implicit theories and self-construal are observed rather than randomly assigned (treatment-by-covariate design), it is important to include any known predictors of responses to signals that might correlate with implicit theories and self-construal as control variables in the regression models.[95] Based on previous research in international relations and political psychology, I include militant internationalism, political ideology, international trust, right-wing authoritarianism, and two common measures of cognitive style, need for closure and need for cognition.[96]

The results from interactions, particularly interactions with continuous variables or higher-order interactions, cannot be read off regression tables directly.[97] To interpret the results, I therefore calculated the marginal effect of strategic context across a range of values of the relevant covariate. For the three-way interactions, I held the other treatment (the cost of Country X's action) constant at 0 and 1, respectively. The 95% confidence intervals for the marginal effects are calculated from the variance-covariance matrix of the respective model.[98]

Models 1 and 2 from table 4.4 present the results from including a three-way interac-

[94] Interactions are not the only, but arguably the most compelling, statistical tool for subgroup analysis (Pocock et al. 2002).

[95] Kam and Trussler 2017.

[96] For studies linking these variables to the perception of signals, see Albuyeh and Paradis 2018; Aronoff 2014; Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; B. C. Rathbun, Kertzer, and Paradis 2017; B. C. Rathbun 2019; Yarhi-Milo, Kertzer, and Renshon 2018; Yarhi-Milo 2014. D'Agostino and Fincherkiefer 1992 and Webster and Kruglanski 1994 also link cognitive style to attribution. To reduce survey length, international relations scholars often employ a subset of items from validated psychological measures. However, this can render studies less comparable. To avoid this problem, I used validated short versions where possible (i.e. NCS-6 for need for cognition, see Holanda Coelho, Hanel, and Wolf 2020; NFCS-15 for need for closure, see Roets and Hiel 2010; and the VSA scale proposed in Bizumic and Duckitt 2018 for right-wing authoritarianism). For international trust, I used the now standard, single-item measure proposed in P. R. Brewer et al. 2004. The social and economic dimensions of political ideology were assessed separately (see Feldman and Johnston 2014; Talhelm et al. 2015).

[97] Brambor, Clark, and Golder 2006.

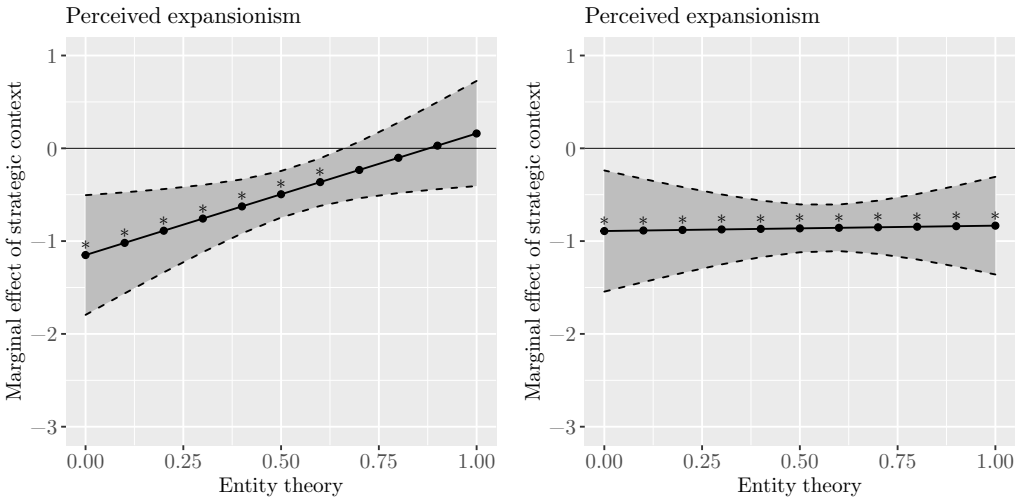
[98] Brambor, Clark, and Golder 2006.

tion between the two treatments and entity theory or entitativity in the linear regression model estimating perceptions of Country X's expansionism. Consistent with the experimental findings described in section 4.4.3, a strategic threat reduces recipients' perceptions of Country X's expansionism in the low-cost (figure 4.5a and 4.5c) and high-cost conditions (figure 4.5b and 4.5d). However, consistent with hypotheses 7 and 8, the extent to which it does varies depending on recipients' implicit theories: in the low-cost conditions, a strategic threat reduces perceived expansionism by 19.17% for recipients who believe personality to be malleable, by 6.07% for recipients who believe personality to be more static than malleable, and has no effect at all for recipients who believe personality to be static (figure 4.5a).[99] Even more strikingly, in the high-cost conditions, a strategic threat decreases perceived expansionism by 26.5% for recipients who do not view countries as unified, coherent wholes, but only by 9.08% for recipients who do view countries as relatively unified, coherent wholes (figure 4.5d). For the highest values of entitativity, strategic context has no effect on recipients' perceptions of Country X's expansionism. Interestingly, and contrary to hypotheses 7 and 8, these interaction effects do not hold across varying levels of cost: the interaction with entity theory vanishes in the high-cost conditions and the interaction with entitativity is relatively weak in the low-cost conditions (figures 4.5b and 4.5c).[100]

Model 1 from table 4.5 presents the results from including a three-way interaction between the two treatments and interdependent self-construal in the regression model estimating perceptions of Country X's expansionism. Again, consistent with the findings from section 4.4.3, the presence of a strategic threat reduces recipients' attribution of Country X's action to underlying expansionism in the low-cost and high-cost conditions. Moreover, consistent with hypothesis 9, the size of this effect varies depending on recipients' self-construal. However, contrary to hypothesis 9, the effect of strategic context is larger when interdependent self-construal is lower, not higher. For example, when the cost of Country X's action is low, a strategic threat reduces perceived expansionism by 12.93% for recipients who score low on interdependence, but only by 5.97% for recipients who score high on interdependence (figure 4.6a). For the highest values of interdependence,

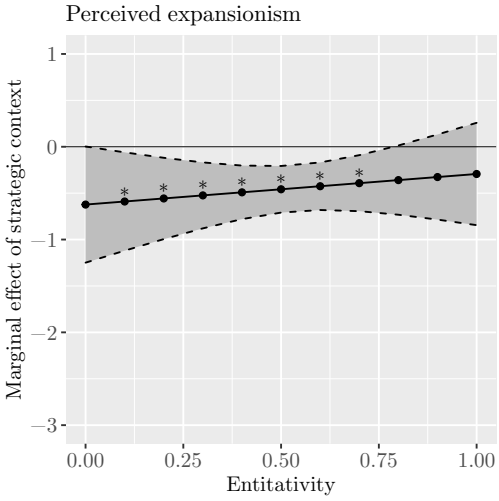
[99] All percentages reported in this section are based on the highest or lowest marginal effect that reaches statistical significance.

[100] The difference in effect between the lowest and highest score of entitativity for which the marginal effect of strategic context is significant is 3.29%.

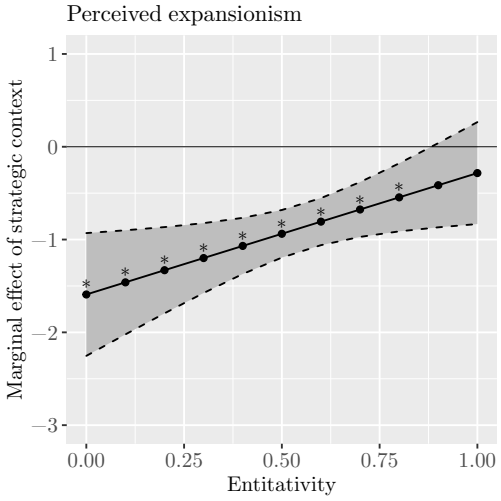


(a) Cost of action low

(b) Cost of action high



(c) Cost of action low



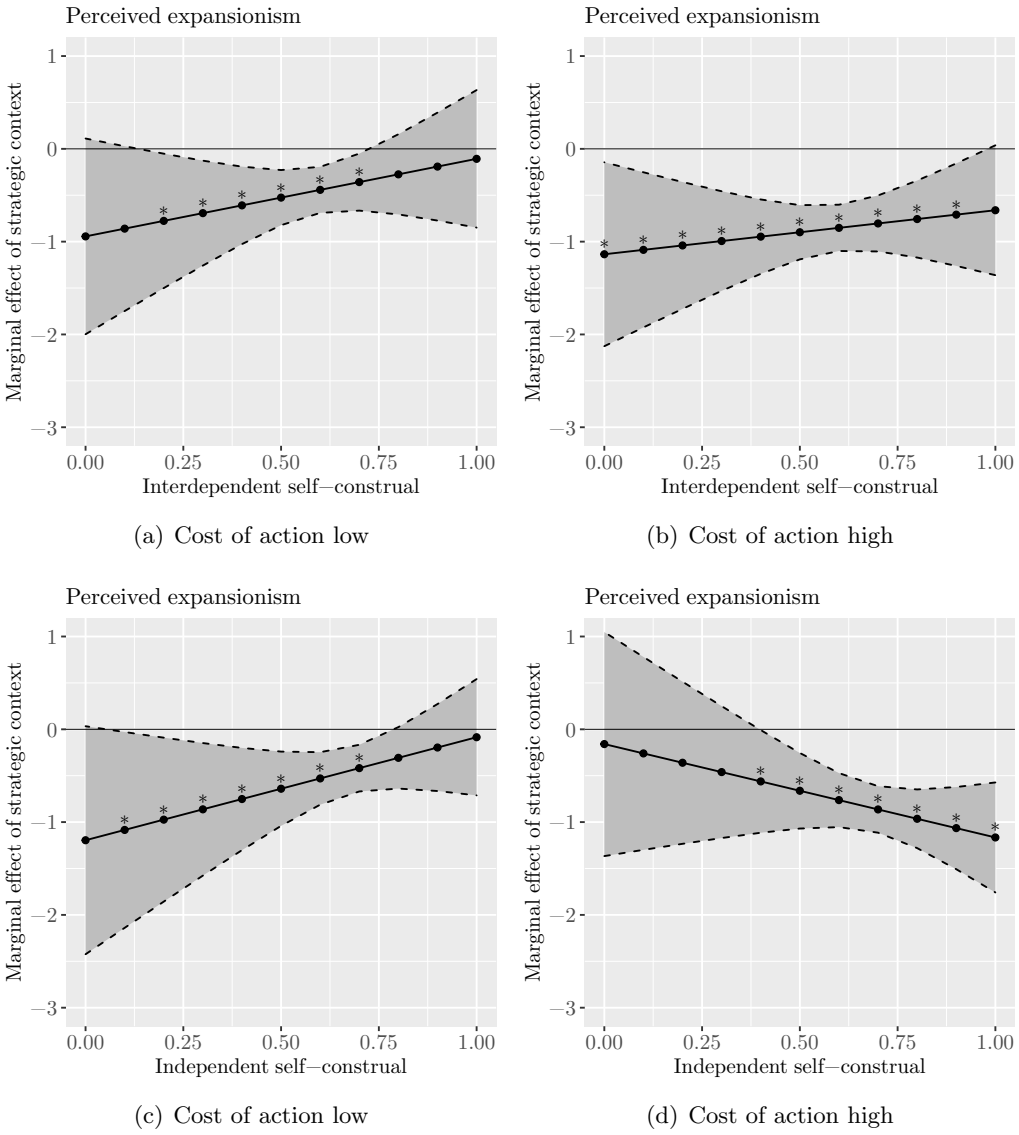
(d) Cost of action high

Figure 4.5: The marginal effect of strategic context on perceived expansionism across values of entity theory and entitativity based on model estimates from table 4.4. Asteriske indicate significance based on the critical $t-$value for a two-tailed, 95% significance test in each model.

Table 4.4: Implicit theories and self-construal as moderators of signaling

	Perceived expansionism		Predicted consistency	
	(1)	(2)	(3)	(4)
Sunk cost: high	0.511	0.599*	0.521	0.159
	(0.322)	(0.317)	(0.396)	(0.389)
Context: threat	−1.150***	−0.623*	−0.175	−0.423
	(0.329)	(0.319)	(0.405)	(0.392)
Entity theory	−0.015	0.298	0.905*	0.586**
	(0.418)	(0.222)	(0.514)	(0.273)
Entitativity	0.244	0.037	0.117	−0.069
	(0.220)	(0.403)	(0.270)	(0.495)
Interdependence	0.629*	0.681*	0.150	0.152
	(0.362)	(0.363)	(0.445)	(0.446)
Independence	1.338***	1.158***	1.674***	1.659***
	(0.398)	(0.400)	(0.490)	(0.492)
Militant internationalism	0.290	0.308	0.132	0.112
	(0.251)	(0.250)	(0.308)	(0.307)
Political ideology: social	0.223	0.228	0.389*	0.389*
	(0.163)	(0.163)	(0.200)	(0.201)
International trust	0.484***	0.491***	0.903***	0.896***
	(0.175)	(0.176)	(0.215)	(0.216)
Right-wing authoritarianism	0.541**	0.509*	0.353	0.367
	(0.275)	(0.275)	(0.338)	(0.338)
Need for closure	0.887***	0.814**	0.645	0.605
	(0.341)	(0.340)	(0.419)	(0.417)
Need for cognition	0.218	0.219	−0.009	−0.015
	(0.276)	(0.277)	(0.340)	(0.340)
High cost*threat	0.259	−0.969**	−0.219	−0.127
	(0.469)	(0.464)	(0.576)	(0.569)
High cost*entity theory	−0.105		−0.607	
	(0.538)		(0.662)	
Threat*entity theory	1.310**		0.040	
	(0.564)		(0.694)	
High cost*threat*entity theory	−1.252		−0.139	
	(0.784)		(0.964)	
High cost*entitativity		−0.276		0.055
		(0.544)		(0.668)
Threat*entitativity		0.330		0.495
		(0.545)		(0.669)
High cost*threat*entitativity		0.979		−0.315
		(0.782)		(0.960)
Constant	1.035***	1.110***	0.698	1.013**
	(0.361)	(0.369)	(0.444)	(0.454)
N	841	841	841	841
R^2	0.198	0.195	0.109	0.108
Adjusted R^2	0.182	0.179	0.092	0.090

*p < .1; **p < .05; ***p < .01



(a) Cost of action low

(b) Cost of action high

(c) Cost of action low

(d) Cost of action high

Figure 4.6: The marginal effect of strategic context on perceived expansionism across values of interdependence and independence based on model estimates from table 4.5.

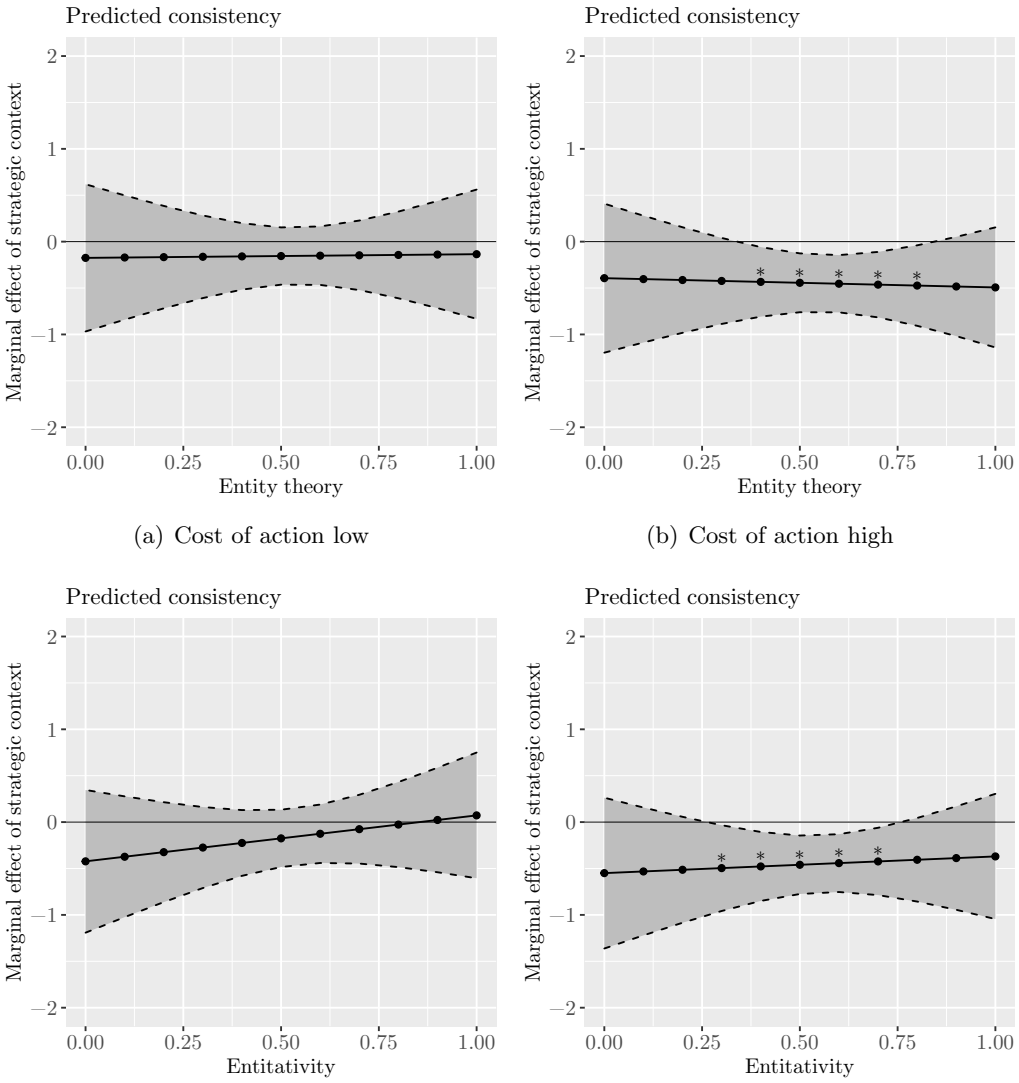
the effect of strategic context turns insignificant.

To gain a more complete picture of the relationship between signaling and self-construal, I re-ran the regression model with a three-way interaction between the two treatments and independent self-construal (table 4.5, model 2). Interestingly, in the low-cost conditions, the results are consistent with hypothesis 9: a strategic threat decreases perceived expansionism by 18.07% for recipients' who score low on independence, but only by 6.97% for recipients who report a very independent self-construal (figure 4.6c). For the highest values of independence, the effect of strategic context is insignificant. By contrast, in the high-cost conditions, the results mirror the results from the three-way interaction with

Table 4.5: Implicit theories and self-construal as moderators of signaling

	Perceived expansionism		Predicted consistency	
	(1)	(2)	(3)	(4)
Sunk cost: high	−0.208	−0.820	0.771	0.285
	(0.510)	(0.631)	(0.624)	(0.773)
Context: threat	−0.943*	−1.195*	1.013	0.676
	(0.538)	(0.627)	(0.659)	(0.768)
Entity theory	−0.260	0.578	1.030	0.116
	(0.639)	(0.364)	(0.783)	(0.446)
Entitativity	0.278	0.295	0.588**	0.576**
	(0.222)	(0.222)	(0.272)	(0.273)
Interdependence	0.269	0.280	0.144	0.168
	(0.220)	(0.221)	(0.270)	(0.271)
Independence	1.250***	0.355	1.701***	2.420***
	(0.399)	(0.665)	(0.488)	(0.815)
Militant internationalism	0.292	0.319	0.125	0.119
	(0.251)	(0.251)	(0.308)	(0.307)
Political ideology: social	0.246	0.228	0.397**	0.392*
	(0.164)	(0.163)	(0.201)	(0.200)
International trust	0.494***	0.497***	0.912***	0.904***
	(0.176)	(0.175)	(0.215)	(0.215)
Right-wing authoritarianism	0.567**	0.530*	0.385	0.378
	(0.275)	(0.275)	(0.337)	(0.336)
Need for closure	0.819**	0.783**	0.548	0.586
	(0.341)	(0.340)	(0.418)	(0.417)
Need for cognition	0.268	0.235	−0.014	0.002
	(0.278)	(0.277)	(0.341)	(0.339)
High cost*threat	−0.192	1.036	−1.612*	−0.198
	(0.739)	(0.879)	(0.905)	(1.077)
High cost*interdependence	1.097		−0.981	
	(0.824)		(1.009)	
Threat*interdependence	0.836		−1.964*	
	(0.879)		(1.077)	
High cost*threat*interdependence	−0.362		2.215	
	(1.205)		(1.476)	
High cost*independence		1.842**		−0.169
		(0.898)		(1.100)
Threat*independence		1.110		−1.230
		(0.908)		(1.112)
High cost*threat*independence		−2.116*		−0.129
		(1.262)		(1.547)
Constant	1.432***	1.573***	0.354	0.389
	(0.437)	(0.482)	(0.535)	(0.590)
N	841	841	841	841
R^2	0.192	0.193	0.110	0.110
Adjusted R^2	0.177	0.178	0.093	0.093

*p < .1; **p < .05; ***p < .01



(a) Cost of action low

(b) Cost of action high

(c) Cost of action low

(d) Cost of action high

Figure 4.7: The marginal effect of strategic context on behavioral predictions across values of entity theory and entitativity based on model estimates from table 4.4.

interdependent self-construal: contrary to hypothesis 9, the effect of strategic context is largest for high values of independence (a 19.42% decrease in perceived expansionism) and insignificant for low values of independence (figure 4.6d).

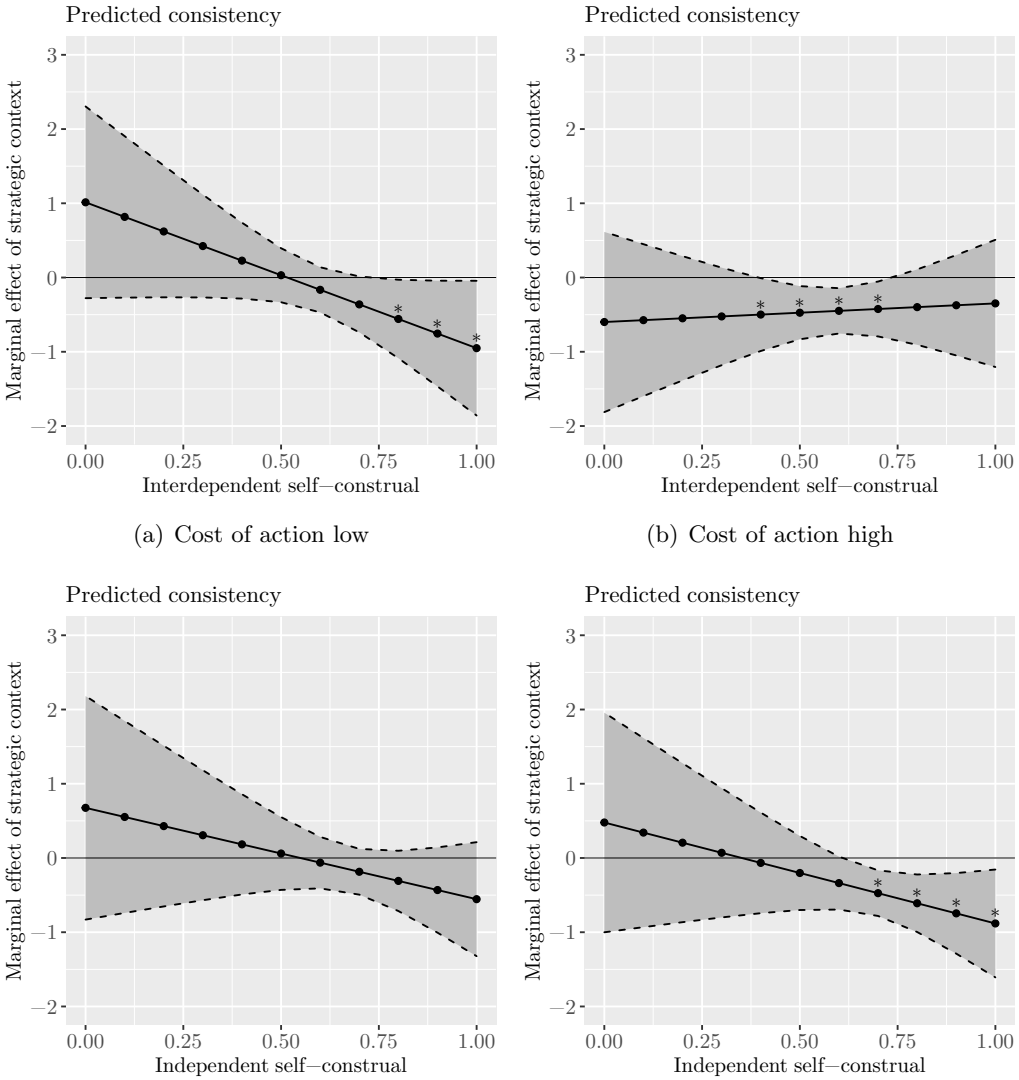
Turning to the linear regression models estimating predictions of Country X's behavioral consistency across strategic contexts, models 3 and 4 from table 4.4 present the results from including three-way interactions between the two treatments and entity theory or entitativity, respectively. Consistent with the findings from the regression model in section 4.4.3 (table 4.3), the presence of a strategic threat significantly decreases perceptions of Country X's expansionism in the high-cost conditions (figures 4.7b and 4.7d),

but not the low-cost conditions (figures 4.7a and 4.7c). In the high-cost conditions, the change in the marginal effect of strategic context across values of entity and entitativity is negligible (figures 4.7b and 4.7d).

By contrast, the results from including a three-way interaction between the two treatments and interdependent self-construal are partly consistent with hypothesis 9 (model 3, table 4.5): in the low-cost conditions, a strategic threat decreased the estimated likelihood of Country X behaving similarly across strategic contexts by 9.3% for recipients with very interdependent self-construals, but not at all for recipients who report less interdependent self-construals. For the highest values of interdependence, a strategic threat reduced estimates by as much as 15.85% (figure 4.8a). In the high-cost conditions, the change in the marginal effect of strategic context across values of interdependence again becomes negligible (figure 4.8b).

Interestingly, the results from the three-way interaction between the two treatments and independent self-construal show a reversed pattern (model 4, table 4.5): in the high-cost conditions, a strategic threat has no effect for recipients with less independent self-construals, but reduces the estimated likelihood of Country X behaving similarly across strategic contexts by 7.89% for recipients with very independent self-construals and by 14.69% at the highest values of independence (figure 4.8d).

Finally, turning to the regression models estimating perceptions of Country X's defensiveness (table 4.6), the results suggest that the effect of strategic context varies depending on all three hypothesized sources of heterogeneity: recipients' entity theory, entitativity and self-construal. However, there are important differences in both, the strength and the direction, of the interactions. For entity theory, the interaction is relatively weak and runs opposite to the hypothesized direction (hypothesis 7, section 4.3.1): the presence of a strategic threat increases perceived defensiveness by 5.52% for recipients who believe personality to be very malleable and by 8.58% for recipients who believe personality to be static (figure 4.9a). The pattern for entitativity is similar, although the interaction is slightly stronger: a strategic threat significantly increases perceived defensiveness by 5.27% for recipients who score relatively low on entitativity, but by 9.79% for recipients who consider countries to be unified and coherent wholes (figure 4.9b). By contrast, the results from the three-way interaction with interdependence provide support for hypothesis



(a) Cost of action low

(b) Cost of action high

(c) Cost of action low

(d) Cost of action high

Figure 4.8: The marginal effect of strategic context on behavioral predictions across values of interdependence and independence based on model estimates from table 4.5.

Table 4.6: Implicit theories and self-construal as moderators

	Perceived defensiveness			
	(1)	(2)	(3)	(4)
Sunk cost: high	0.256***	0.256***	0.261***	0.262***
	(0.084)	(0.084)	(0.084)	(0.084)
Context: threat	0.285	0.200	−0.060	−0.013
	(0.218)	(0.217)	(0.344)	(0.408)
Entity theory	−0.344	−0.237	−0.245	−0.230
	(0.268)	(0.207)	(0.207)	(0.208)
Entitativity	0.680***	0.497*	0.682***	0.661***
	(0.206)	(0.269)	(0.205)	(0.206)
Interdependence	−0.023	0.002	−0.385	−0.007
	(0.339)	(0.339)	(0.426)	(0.339)
Independence	1.408***	1.375***	1.386***	1.077**
	(0.372)	(0.372)	(0.371)	(0.478)
Militant internationalism	0.224	0.227	0.229	0.245
	(0.235)	(0.234)	(0.233)	(0.233)
Political ideology: social	0.283*	0.278*	0.281*	0.280*
	(0.153)	(0.152)	(0.152)	(0.152)
International trust	0.339**	0.335**	0.339**	0.340**
	(0.164)	(0.164)	(0.164)	(0.164)
Right-wing authoritarianism	−0.201	−0.224	−0.190	−0.207
	(0.257)	(0.257)	(0.257)	(0.256)
Need for closure	1.374***	1.376***	1.381***	1.374***
	(0.318)	(0.318)	(0.318)	(0.318)
Need for cognition	0.504*	0.491*	0.523**	0.500*
	(0.259)	(0.259)	(0.259)	(0.259)
Threat*entity	0.229			
	(0.365)			
Threat*entitativity		0.387		
		(0.365)		
Threat*interdependence			0.793	
			(0.560)	
Threat*independence				0.624
				(0.587)
Constant	1.185***	1.253***	1.338***	1.342***
	(0.303)	(0.312)	(0.325)	(0.352)
N	841	841	841	841
R^2	0.159	0.160	0.161	0.160
Adjusted R^2	0.146	0.147	0.148	0.147

*p < .1; **p < .05; ***p < .01

9: whereas strategic context has no effect on recipients who score low on interdependence, a strategic threat increases perceived defensiveness by 12.22% for recipients with very interdependent self-construals (figure 4.9c). Interestingly, the results from the three-way interaction with independence show a similar pattern: whereas strategic context has no effect on recipients who score low on independence, a strategic threat increases perceived defensiveness by 10.18% for recipients with very independent self-construals (figure 4.9d).

In summary, the results from models estimating recipients' predictions of Country X's cross-situational consistency provide no support for the idea that the effect of a sender's strategic context on inferences about the sender's motives vary depending on recipients' implicit theories, and weaker support for the idea that the effect of strategic context varies depending on recipients' self-construal.[101] However, these results must be interpreted with caution: given that the treatments generally had weaker effects on behavioral predictions than on perceptions of expansionism or defensiveness, the three-way interactions are likely to be underpowered. Particularly for interdependent and independent self-construal, where we see a strong interaction slope, but wide confidence intervals, the results might therefore mask substantively important, and large, interaction effects.

By contrast, the results from the regression models estimating perceptions of Country X's expansionism and defensiveness provide strong support for the idea that the effect of costly signals, and responses to a sender's strategic context in particular, vary depending on recipients' self-construal and their implicit theories of personality and countries.[102] Implicit theories and self-construal emerge as important sources of heterogeneity in costly signaling.

At the same time, the direction and magnitude of the estimated interactions are inconsistent across levels of cost and across outcomes. Notably, the effect of strategic context on perceptions of Country X's expansionism is *smaller* for recipients who consider personality to be more static and countries to be more unified, providing support for hypotheses 7 and 8. However, the direction of the interaction reverses for perceptions of Country X's defensiveness. Similarly, the effect of strategic context on behavioral predictions and perceptions of Country X's defensiveness is *larger* for recipients who report very interde-

[101] The results become insignificant if we apply a Bonferroni correction to account for multiple hypothesis testing (i.e. the critical $t-$value is ¿ 2).

[102] The results do not substantially change if we apply a Bonferroni correction, and divide the α of .05 by 3 to account for multiple hypothesis testing.

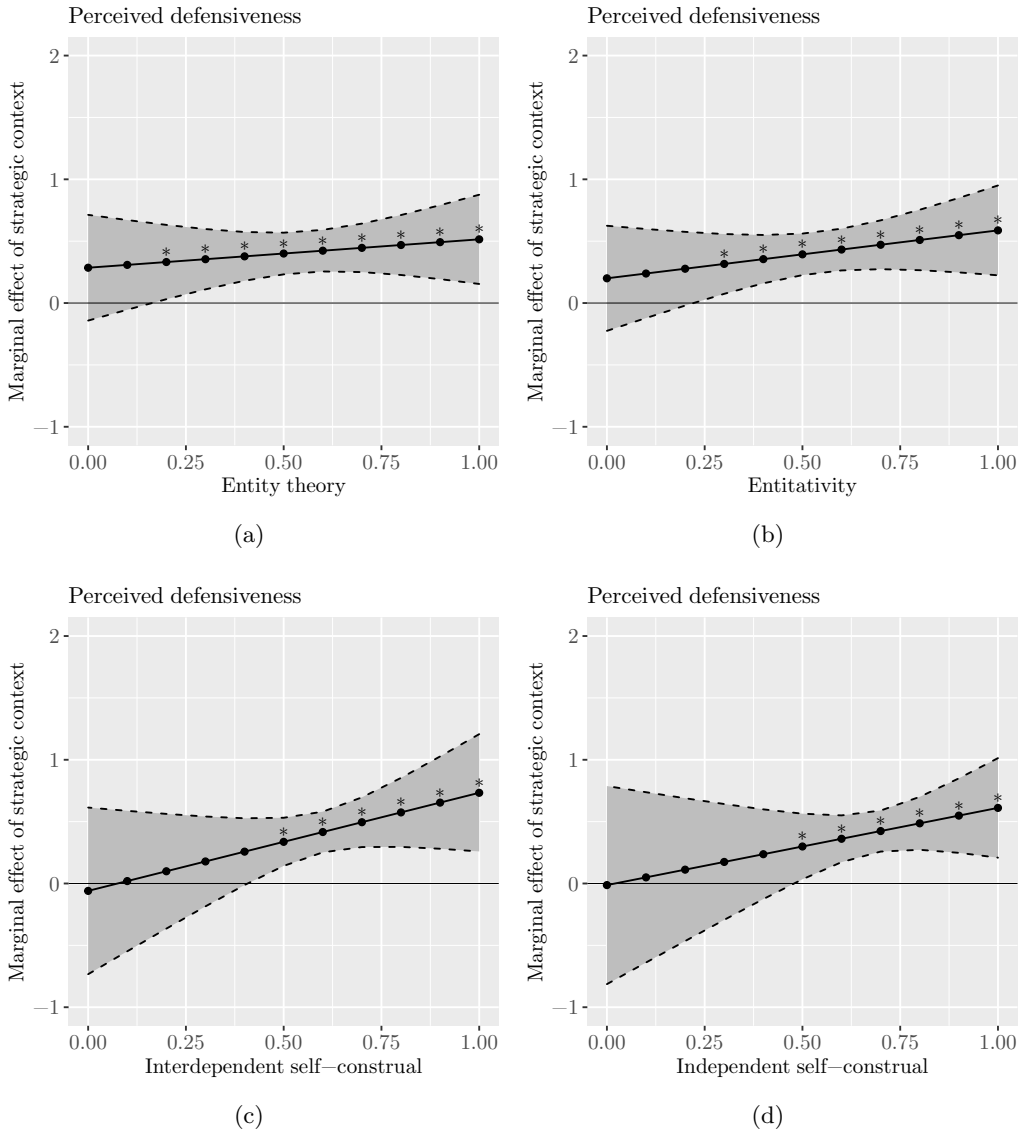


Figure 4.9: The marginal effect of strategic context on perceived defensiveness across values of entity theory, entitativity, interdependence, and independence based on model estimates from table 4.6.

pendent self-construals, providing support for hypothesis 9, yet the pattern reverses for perceptions of Country X's expansionism.

The most plausible explanation of these findings is that the second outcome measure, which assesses perceptions of Country X's defensiveness, indicates a dispositional attribution after all and situational attributions are not equivalent to their dispositional counterpart, i.e. the findings on dispositional versus situational attribution in psychology are not reducible to differences in attention to one disposition versus another.[103] If so, the findings from recipients' perceptions of expansionism and defensiveness would provide consistent support for hypotheses 7 and 8, mirroring the finding from research on implicit theories in psychology that '[w]hen situational information discounted the implied trait (e.g. anxious behavior, but in an anxious situation [here, defensive behavior, but in a threatening situation]), incremental theorists' trait attributions were weakened, while entity theorists' remained firm'.[104] In contrast, this interpretation of the second outcome measure would suggest that the findings consistently disconfirm (rather than partially confirm) hypothesis 9: although the effect of strategic context on perceived expansionism and defensiveness clearly and consistently varies across recipients depending on their self-construal, the interaction runs in the opposite direction, indicating the need for further research to understand the link between conceptions of the self and attribution in international politics.

Not only is this interpretation intuitively plausible in light of the overall findings from this study; it is also more plausible in light of one additional finding: participants' perceptions of Country X's expansionism and perceptions of Country X's defensiveness are positively correlated (Kendall's $\tau = .26$) This runs counter to the idea that the two measures capture the dispositional-situational dichotomy. It also runs counter to the idea that expansionist motives and defensive motives can be treated as opposite ends of a continuum in international relations, instead suggesting that recipients perceive greater complexity in states' dispositions than our conceptual tools tend to capture.

[103] See section 4.3, fn 61.
[104] Plaks, Levy, and Dweck 2009, p. 1073.

4.5 Discussion

In interpreting the results, several questions merit attention. First, beyond their statistical significance, how substantively significant are the treatment effects found in this study? The effect sizes from the primary analysis range from a 4.4% increase in predicted behavioral consistency to a 14.9% decrease in perceived expansionism. These effect sizes are comparable to the effect sizes found in other studies. For example, Kertzer, Rathbun and Rathbun find that signals of reassurance increase an adversary's perceived trustworthiness by 4.33% when they incur a high cost to the sender compared to when they incur a low cost.[105] The largest treatment effect reported in the literature is a 21.35% increase in the perceived credibility of sunk-cost signals when they carry a high cost compared to a low cost.[106] Given the large body of work establishing the importance of relatively static 'images' in international politics, even a small change in recipients' perceptions of the sender in response to a one-off event or signal is a theoretically important finding; a change of 10% or more is a theoretically and substantively important finding.

There is also reason to believe that the experiment underestimates the actual treatment effects: the analysis focuses on the intent-to-treat (ITT) effect, i.e. it includes all participants who were randomly assigned to a treatment condition. However, only between 41.9% (in treatment group 2) and 63.5% (in treatment group 3) of the participants assigned to a treatment condition correctly answered both factual manipulation/attention check questions at the end of the survey (see appendix figures 4.10 and 4.11). While recall is an imperfect measure of compliance,[107] it nonetheless provides a useful approximation of non-compliance. The inclusion of non-compliers in the analysis likely dilutes the treatment effects.[108]

Third, although the aim of this study was to investigate causal processes rather than to generalize to a population, it is worth considering the substantive relevance of the sam-

[105] Kertzer, B. C. Rathbun, and N. S. Rathbun 2020. I cite here the only treatment effect that is both significant *and* comparable to one of the treatments in this study (i.e. levels of cost, not the presence or absence of a signal).
[106] Quek 2021.
[107] Recall could be impacted by factors other than close reading of the assigned vignette, and vignette-induced treatments could impact responses even if participants subsequently recall the relevant detail of the scenario incorrectly.
[108] In line with this claim, the effect sizes are substantially larger if I rerun the Wilcoxon tests with subjects who correctly answered the attention check questions only, particularly for perceptions of expansionism and defensiveness.

ple average treatment effects from this study to signaling and bargaining in international politics. Despite being a convenience sample, the U.S. sample was highly representative. As far as the general public is concerned, we can therefore expect the sample average treatment effect to be very informative. This is highly germane to international crises, where public perceptions of an adversary can be detrimental to efforts to de-escalate, or to international agreements, where public perceptions of the signatories can impact everything from the initial negotiations to the final ratification, particularly in democracies. For example, growing distrust of the Soviet Union among the American public played a vital role in the Soviet brigade crisis of 1979, which fatefully delayed the ratification of the SALT II treaty between the U.S. and Soviet Union.

Should we also expect the findings from this study to bear on policy elites with decision making power whose perceptions of signals can immediately affect diplomatic exchanges and strategic choices? Due to the mutually-reinforcing effects of selection and training, policy elites are considered to be more rational than the general public and, thus, more likely to conform with rationalist predictions.[109] Consequently, we should expect policy elites to be more sensitive to information about a state's strategic environment when they infer the state's motives from its actions. The findings on heterogeneity in attribution presented in this paper further add to this idea: policy elites and, particularly, foreign policy elites with direct exposure to the institutional complexities of foreign policy and strategic decision making are likely to assume states to be less entitative than individuals who observe them from 'afar' and rely primarily on news accounts where state behavior is often attributed either to states as a whole or to prominent leaders. If so, the findings from this study suggest that policy elites will be more responsive to information about a state's strategic environment, attributing behavior to motives in the way costly signaling predicts.

Finally, it is worth speculating about the impact that the choice of the UK as the second location for data collection has on the findings for hypothesis 6. Since the UK is an historically important ally of the U.S., it is possible that participants perceive themselves as being part of the same group or sharing an identity with the U.S. in international politics. Compared to other third-party countries, the effect of group membership dif-

[109] Renshon, Dafoe, and Huth 2018; see B. C. Rathbun 2019 for a countervailing view.

ferences therefore may be smaller than for other countries (for example, Switzerland). However, given that we would expect a level effect, this would not undermine the validity of the findings. To the contrary, the fact that we find a difference, even among allies, is substantively relevant.

4.6 Conclusion

Despite its centrality for understanding conflict and cooperation between states, international relations scholars only recently have begun to systematically and empirically investigate the microfoundations of signaling in international politics. I address a fundamental gap in this growing body of work: the question of whether a sender's strategic context alters the conclusions recipients draw from that sender's actions. Integrating rationalist scholarship on costly signaling with findings from attribution research in psychology, I offer predictions for how a sender's strategic context will impact recipients' conclusions, and which recipients will be most responsive to information about a sender's strategic context. The latter are based on two novel sources of variation in the signaling process – recipients' implicit theories and self-construal – that cut across more familiar, political and ideological fault lines.

To test the theory, I designed a 2x2 factorial experiment that separately varies the two constituent components of a signal's cost: the cost inherent in a sender's actions and a sender's strategic environment. Recipients read an abstract dispute scenario in which the adversary either initiated a challenge to the status quo or responded to a challenge, and in which the adversary's actions inherently carried either a low or a high cost. Since there is no 'view from nowhere' in international politics, and recipients' group membership and role in relation to the scenario likely impacts the conclusions they draw from a signal, I ran the experiment in two locations: the U.S. and the UK. Whereas recipients in the U.S. were party to the hypothetical dispute described in the scenario, recipients in the UK were third-party observers of the dispute. Across both studies I find that a state's strategic environment affects the inferences individuals draw from that state's actions. Specifically, participants attributed a state's sunk-cost signal of resolve less to expansionist motives, and more to defensive motives, when the state's environment was insecure than when it was secure. In the U.S. sample where the cost of a state's actions

affected participants' inferences about motives, the state's strategic context moderated the effect of cost, providing support for the costly signaling hypothesis. At the same time, I find substantively significant variation in the effect that a state's strategic environment has the inferences participants draw from a state's actions depending on participants' implicit theories of individuals and states, as well as their self-construal.

Future research could explore whether the predictions and findings from this study extend to costly signals of reassurance. With the exception of the ultimate attribution error, this study builds on findings from psychology that should bear on the attribution of behavior irrespective of the behavior's valence (i.e. whether it is positive or negative). However, research on the role of group membership in attribution in psychology and research on the role of beliefs in signaling in international relations suggest that recipients may respond differentially to costly signals of reassurance or resolve. By investigating whether this study's findings extend to costly signals of reassurance, future research could shed light on the microfoundations of cooperation between states. Future research could also investigate alternative mechanisms linking self-construal to the assessment of motives in international politics. Although the findings of this study suggest that the effect of a sender's strategic environment on inferences about motives varies depending on recipients' conception of the self, the direction of the interaction runs opposite to that hypothesized in this paper. Finally, future research could employ treatment-by-treatment designs to establish the *causal* role of implicit theories and self-construal in signaling.

This study's findings on the inference of motives in international politics contribute to at least four broader areas of research in international relations. First, the findings offer novel empirical evidence on costly signaling. So far, experimental studies that test the costly signaling hypothesis have manipulated the cost of a signal by varying the sender's actions. However, a key insight of rationalism is that the cost of a signal does not inhere in those actions: it depends on the sender's strategic environment. This paper finds that recipients respond to information about a sender's strategic environment the way costly signaling predicts, providing scholars with new insights into the attributional processes that underpin the costly signaling mechanism on the receiver's end.

Second, the findings shed light on the risk of escalatory dynamics in international conflicts. Rationalist scholarship and psychological research in international relations

commonly emphasize the risk of security dilemmas, where mutual fear drives two security-seekers to engage in a costly and dangerous arms race. However, the findings here suggest that recipients do in fact type-distinguish between states that mobilize in response to a threat and states that mobilize in the absence of a threat: if a state mobilizes in response to a threat, they attribute its arming policy less to underlying motives. This is true even when recipients' own country poses the threat in question. While this does not rule out the possibility of 'rational spirals',[110] it does suggest that the risk of security dilemmas arising from misperceptions may be lower than previously assumed, and highlights the importance of systematic, empirical tests to establish the microfoundational claims invoked within broader explanatory frameworks.

At the same time, this study finds that the extent to which recipients relate a sender's actions to information about its strategic environment varies depending on their implicit theories about individuals and states, as well as their conception of the self. Thereby, it offers novel insights into the question of when individual-level perceptual processes may exacerbate escalatory dynamics in international politics, potentially leading to suboptimal arming policies,[111] and underscores the importance of recipient-driven theories of signaling in international relations. Recipient-driven theories have recently gained traction, with several experimental studies showing that recipients respond differentially to costly signals, depending on their prior beliefs about an adversary or international politics in general.[112] Crucially, this study's findings point toward a need for scholars to consider more complex links between beliefs and information processing: existing research focuses on belief-maintaining mechanisms, like cognitive consistency, confirmation bias, or asymmetric updating, to explain the impact of beliefs on signaling and perception.[113] In contrast, this study suggests that individuals' assumptions about the social world shape, not only the outcome of information processing, but the very process by which external cues like signals are interpreted, giving rise to predictions that cross-cut the effect of belief-maintaining mechanisms on signaling.

Finally, this study's findings also raise conceptual questions about the study of signal-

[110] Glaser 2004; Kydd 1997.
[111] Glaser 2010; Glaser 2004.
[112] Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Yarhi-Milo, Kertzer, and Renshon 2018; Yarhi-Milo 2014.
[113] For example, Holsti 1962; Jervis 1976; Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Rosati 1988; Tetlock 1999; Vertzberger 1990; Yarhi-Milo 2014.

ing in international politics. Notably, while international relations scholars acknowledge that states may have mixed motives and pursue several goals simultaneously, theoretical scholarship on signaling typically assumes that states can be classified clearly according to their 'type'.[114] Though convenient, the findings from this study suggest that this assumption obscures, not only the complexity of senders' motives, but also the extent to which recipients register this complexity. Put differently, the motivational complexity of social agents plays a role at both ends of the signaling mechanism; scholars need only unpack it.

[114] For example, Glaser 2010.

4.7 Appendix

4.7.1 Ethical Approval

Department of Politics & International Relations

University of Oxford, Manor Road, Oxford OX1 3UQ
United Kingdom
Tel: +44 1865 278700
Fax: +44 1865 278725
Email: enquiries@politics.ox.ac.uk
www.politics.ox.ac.uk

Paola Solimena,
University,
Oxford

31 July 2021

Dear Paola,

Research Ethics Approval: Ref No.: SSH_DPIR_C1A_21_029

Theories of mind in International Relations

The above application has been considered on behalf of the DPIR Departmental Research Ethics Committee (DREC) in accordance with the procedures laid down by the University for ethical approval of all research involving human participants.

I am pleased to inform you that, on the basis of the information provided to the DREC, the proposed research has been judged as meeting appropriate ethical standards, and accordingly approval has been granted.

Should there be any subsequent changes to the project, which raise ethical issues not covered in the original application, you should submit details to the DREC for consideration.

Yours sincerely,

Andrew Melling,
Secretary to the DPIR DREC

cc. Dominic Johnson

4.7.2 Experimental Instructions

1. Introduction to the experiment: all respondents

   - On the next screen, you will read about a situation in international politics that could occur in the future. The situation is general and not about a specific country. Please read the description carefully. After you have read the description, we will ask you a few questions. The same description will be shown at the top of each page.

2. Introduction to the scenario: all respondents

   - The year is 2035. [abstraction from present]
   - The U.S. and Country X share access to, and control over, a body of water. [issue; status quo]
   - The body of water is strategically and economically vital for both countries. [stakes]
   - Both countries maintain a strong naval presence and regularly conduct military exercises in the region. [local balance of power]
   - Although the U.S. and Country X regularly face tensions over other issues in international politics, neither country has changed its naval deployment to the body of water in past years. [adversarial relationship; status quo with regard to issue]

3. Randomized conditions: shown on same screen as introduction to the scenario

   (a) Treatment 1

   - Yesterday, Country X decided to make a small increase in its naval deployment to the body of water. [low-cost signal]

   (b) Treatment 2

   - Yesterday, Country X decided to make a large increase in its naval deployment to the body of water. [high-cost signal]

   (c) Treatment 3

- In recent weeks, the U.S. has decided to make a large increase in its naval deployment to the body of water. [threat condition]
- Yesterday, Country X decided to make a small increase in its naval deployment to the body of water. [low-cost signal]

(d) Treatment 4

- In recent weeks, the U.S. has decided to make a large increase in its naval deployment to the body of water. [threat condition]
- Yesterday, Country X decided to make a large increase in its naval deployment to the body of water. [high-cost signal]

4.7.3 CONSORT Participant Flow Diagram

Following the Reporting Guidelines for Experimental Research from the Experimental Research Section Standards Committee, I include a CONSORT participant flow diagrams detailing the number of subjects who began the survey, the number of subjects excluded or otherwise lost prior to survey completion, the number of subjects assigned to each treatment condition, and the number of subjects included in the analysis (A. Gerber et al. 2014).

U.S. Sample

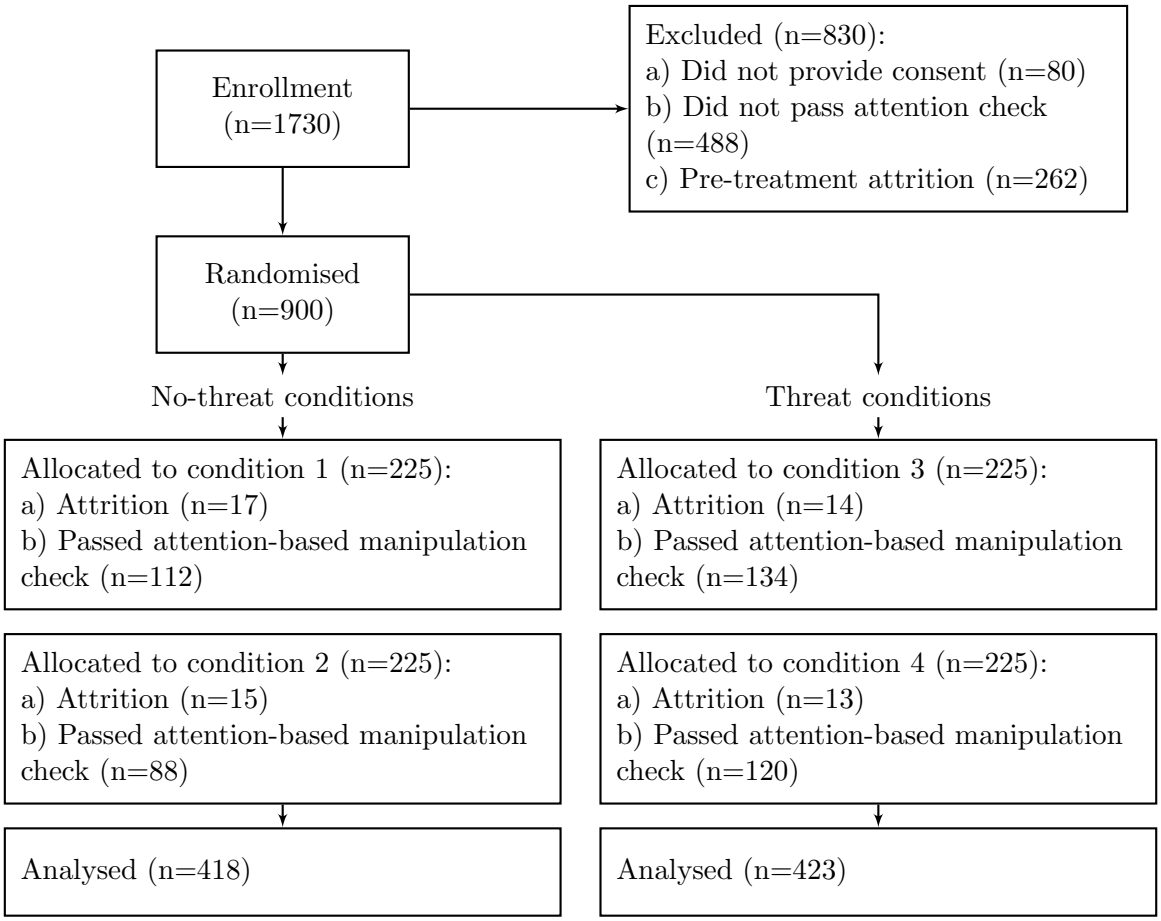


Figure 4.10: Flowchart of subjects' progress through the experiment for U.S. sample, adapted from CONSORT 2010 flow diagram. Note that, for samples recruited via online platforms such as Lucid or MTurk, the total number of subjects assessed for eligibility is unknown. Code adapted from Willert 2010.

UK sample

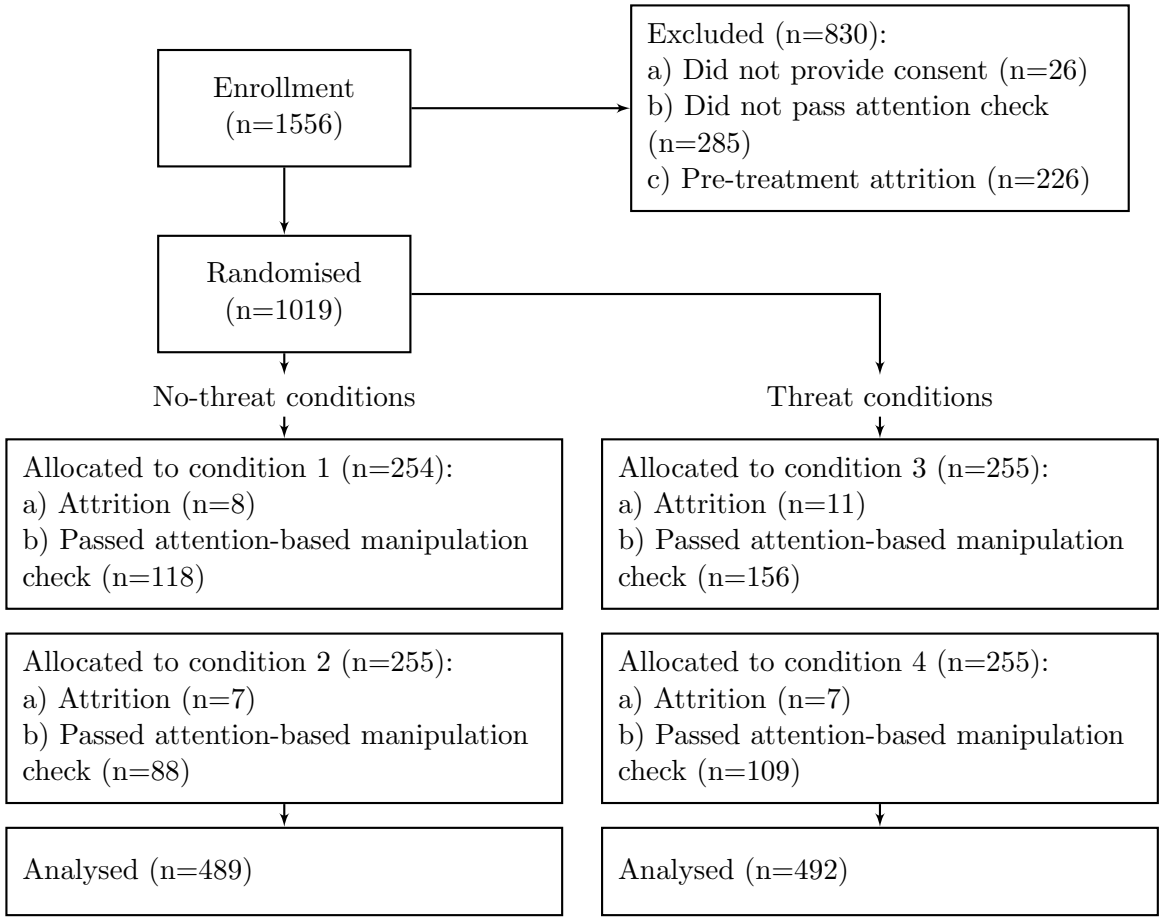


Figure 4.11: Flowchart of subjects' progress through the experiment for the UK sample, adapted from CONSORT 2010 flow diagram.

4.7.4 Supplementary Data Analysis

Attrition

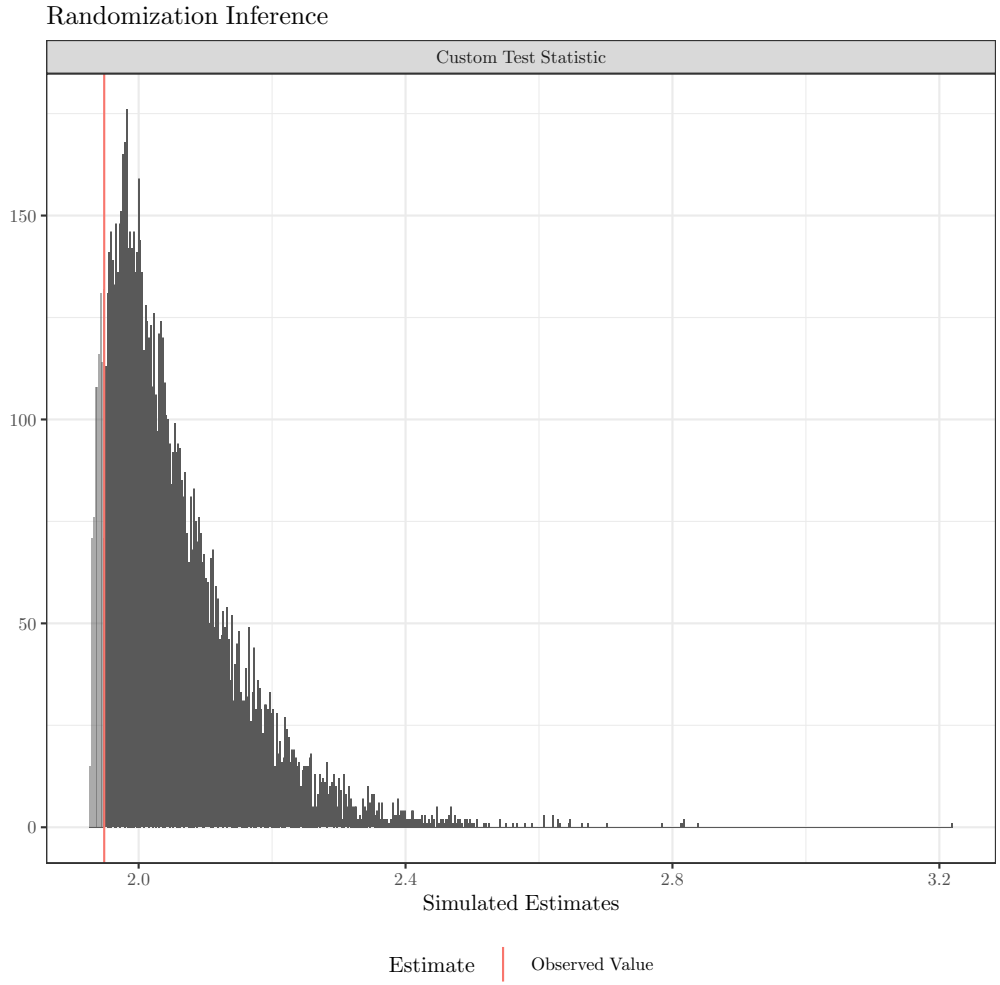


Figure 4.12: U.S. sample: Sampling distribution of F-statistics from linear regression of attrition on treatment assignment and pre-treatment covariates with 10,000 simulations of the randomization procedure. The red line indicates the observed F-statistic, which is non-significant ($p = .92$). The results imply that we cannot reject the null hypothesis of random missingness.

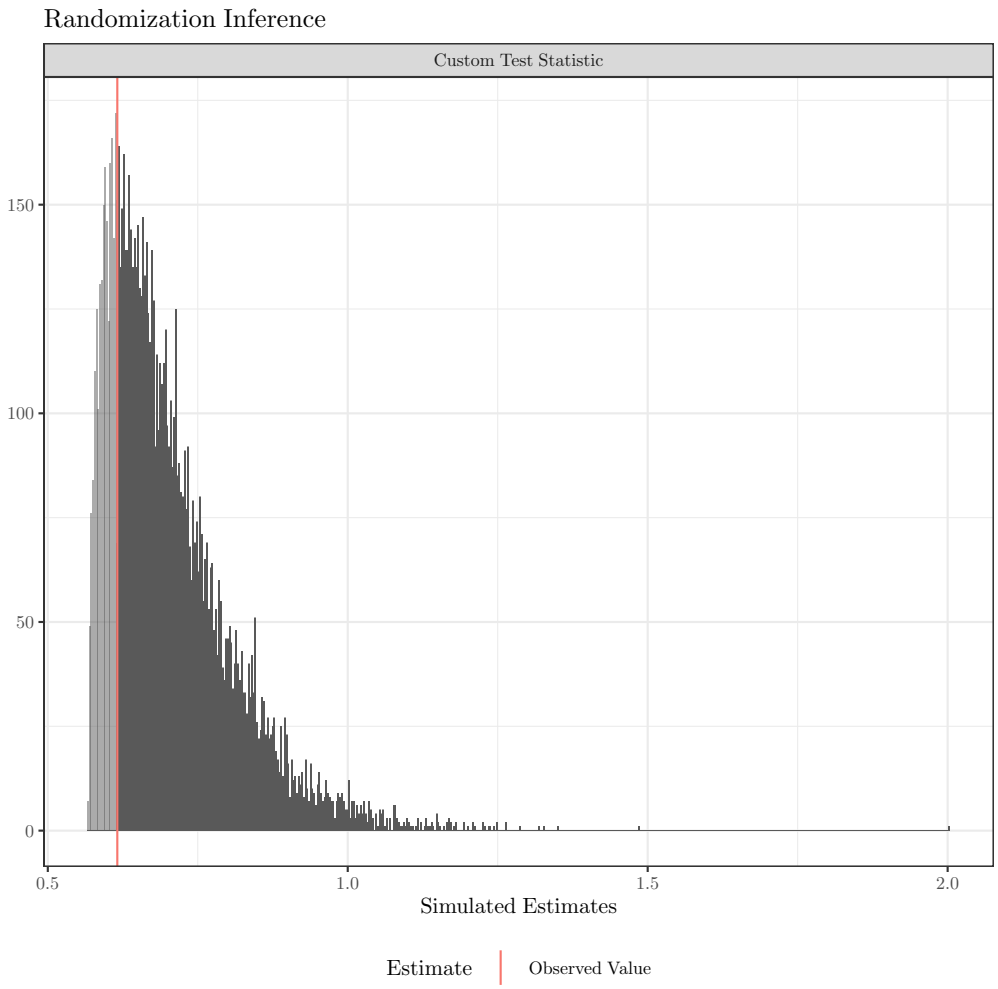


Figure 4.13: UK sample: Sampling distribution of F-statistics from linear regression of attrition on treatment assignment and pre-treatment covariates with 10,000 simulations of the randomization procedure. The red line indicates the observed F-statistic, which is non-significant ($p = .79$). The results imply that we cannot reject the null hypothesis of random missingness.

Covariate Balance

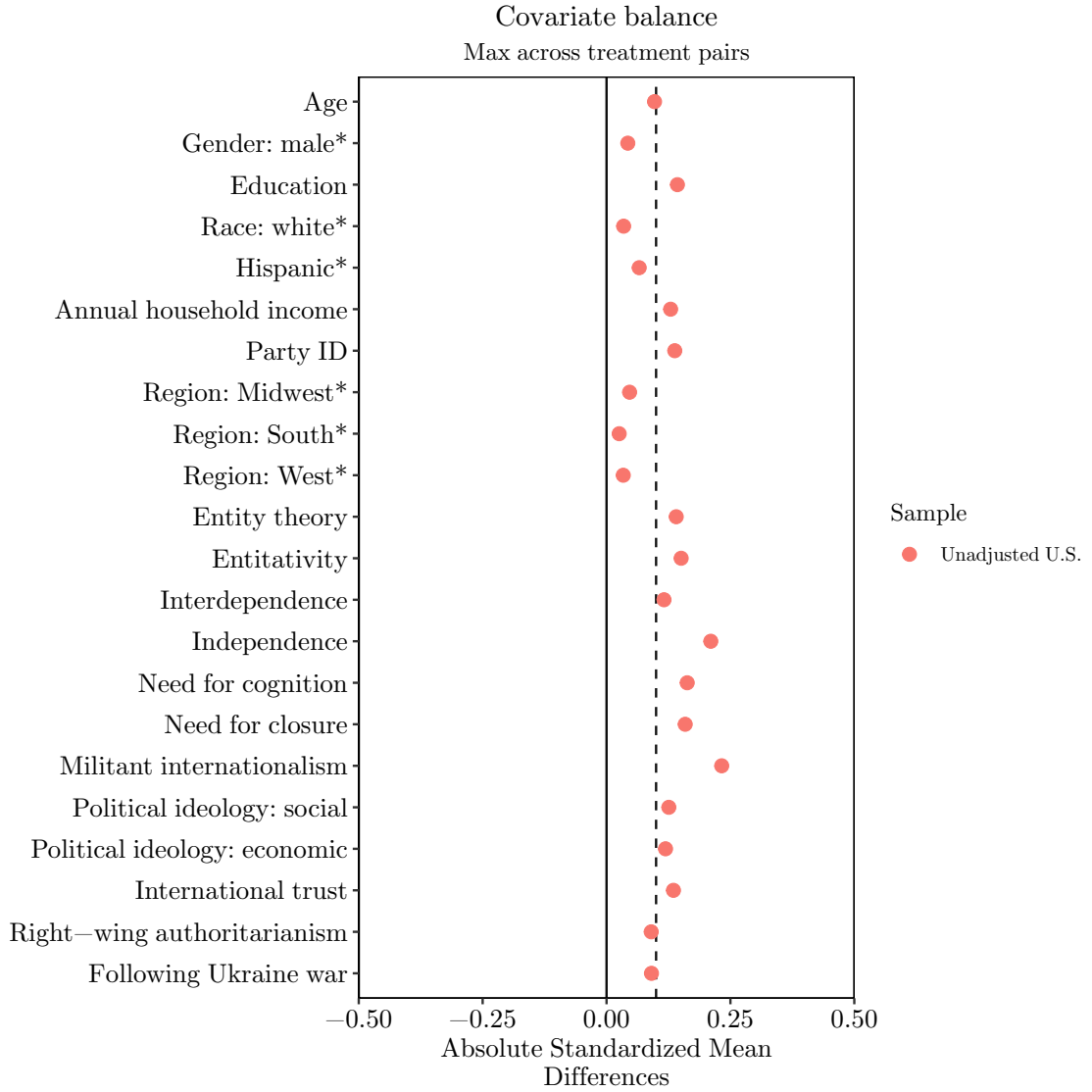


Figure 4.14: Love plot showing maximum difference in standardized means (SDM) across treatment groups for all covariates in the U.S. sample. * indicates the use of raw mean differences rather than standardized mean differences for dummy variables. Dots right of the dashed line indicate differences larger than .1. The largest imbalances occurred for militant internationalism ($SDM = .23$) and independent self-construal ($SDM = .21$).

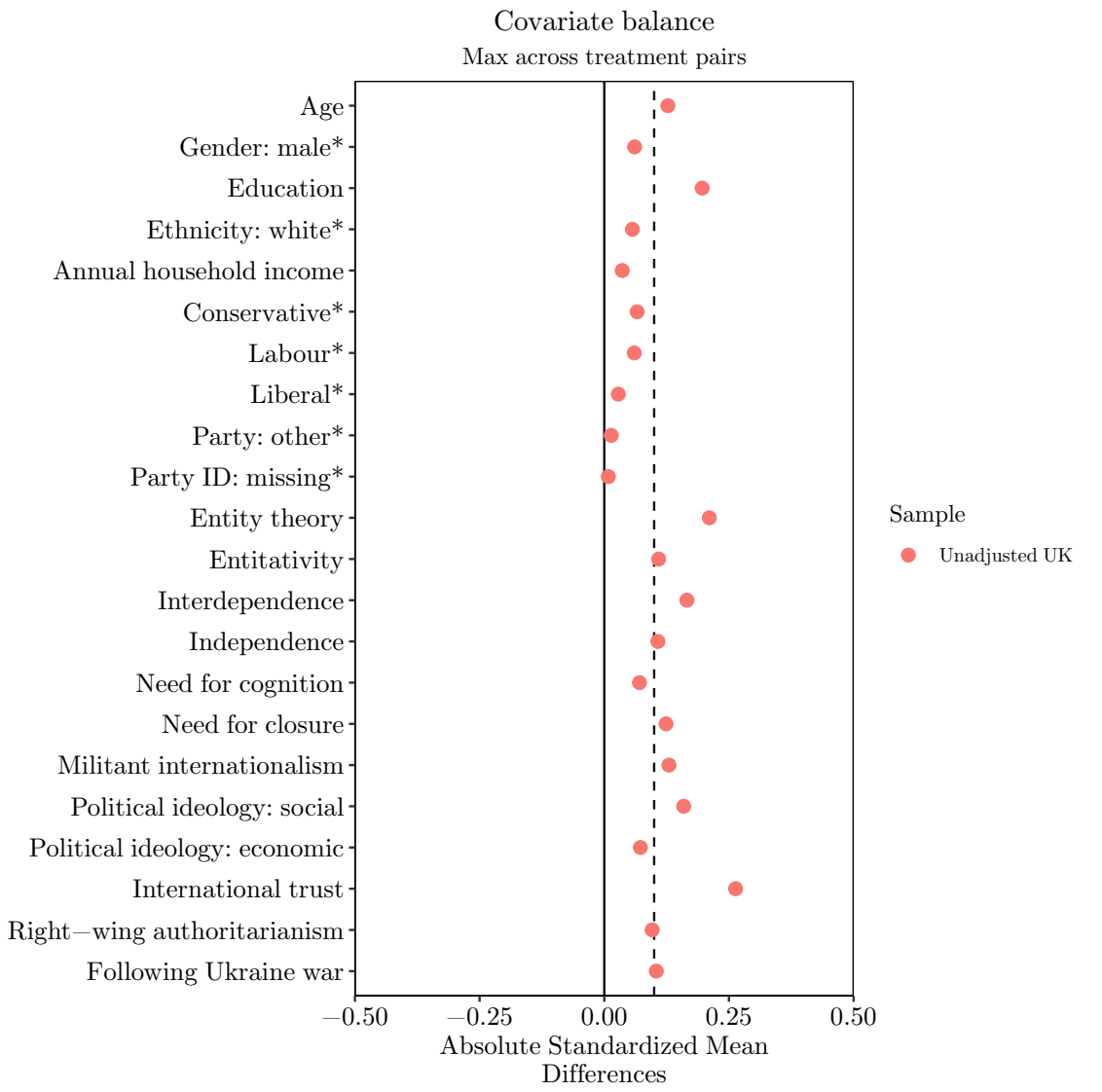


Figure 4.15: Love plot showing maximum difference in standardized means (SDM) across treatment groups for all covariates in the UK sample. * indicates the use of raw mean differences. The largest imbalances occurred for international trust ($SDM = .26$) and entity theory ($SDM = .21$).

Randomization Check

The randomization checks were conducted by regressing the treatment assignment on all covariates, calculating the log-likelihood statistic of the regression, and repeating the procedure for 10,000 simulations of the randomization procedure for the U.S. and UK sample, respectively. Figures 4.16 and 4.17 show the resulting distributions of log-likelihood statistics, which '[represent] the exact sampling distribution under the null hypothesis that no covariates have any effect on the assigned treatment'.[115] The null hypothesis could not be rejected for either sample ($p = 0.0662$ for the U.S. sample, $p = 0.37$ for the UK sample), suggesting that the covariates do not predict treatment assignment better than expected by chance.[116] For the U.S. data, it is worth noting that 93.38% of the log-likelihood statistics are at least as large as the observed log-likelihood, making it, if anything, an unusually low draw.[117]

[115] A. S. Gerber and Green 2012, p. 108.
[116] Lin, Green, and Coppock 2016 recommend a p-value of .01 as the cut-off that should prompt a review of the randomization procedure.
[117] Higher values indicate better model fit.

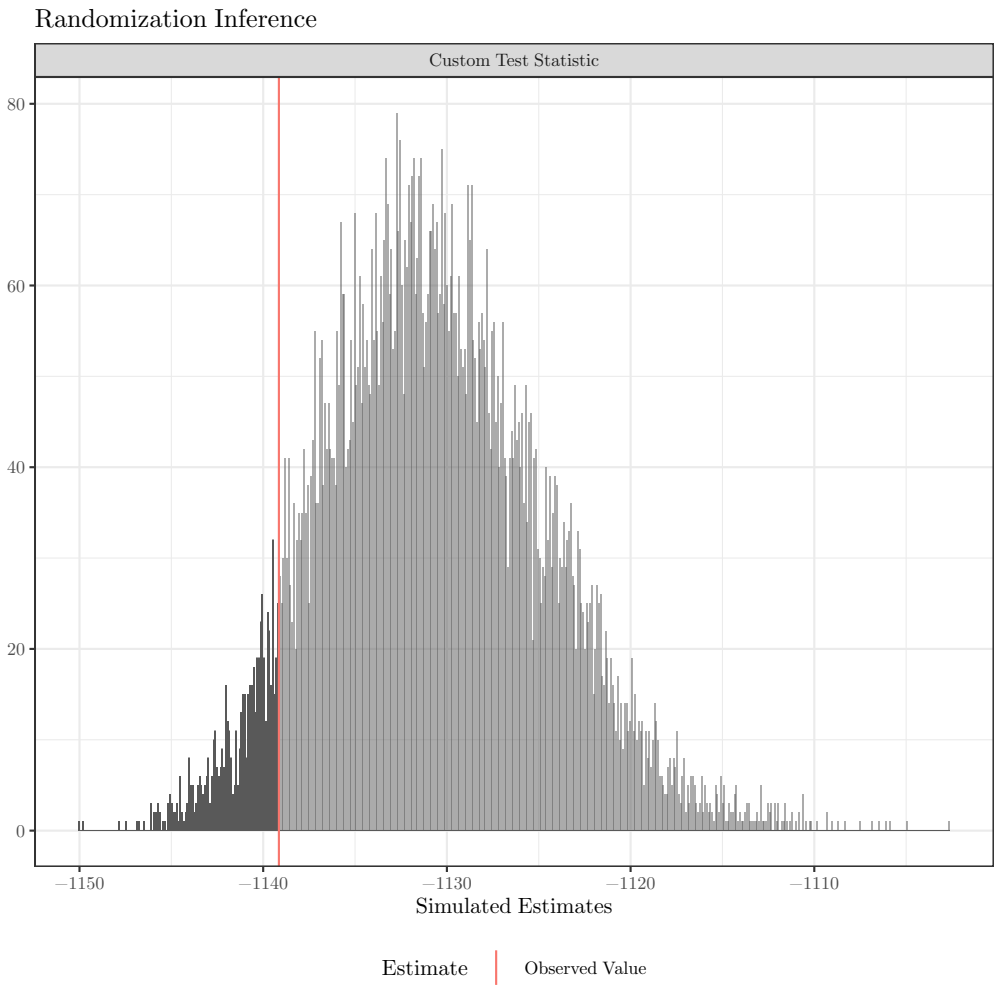


Figure 4.16: Sampling distribution of log-likelihood statistics from 10,000 simulations of the randomization procedure for the U.S. sample. Since there were four treatment groups in this experiment, multinomial logistic regression was used to obtain the log-likelihood statistics. The red line locates the observed log-likelihood statistic.

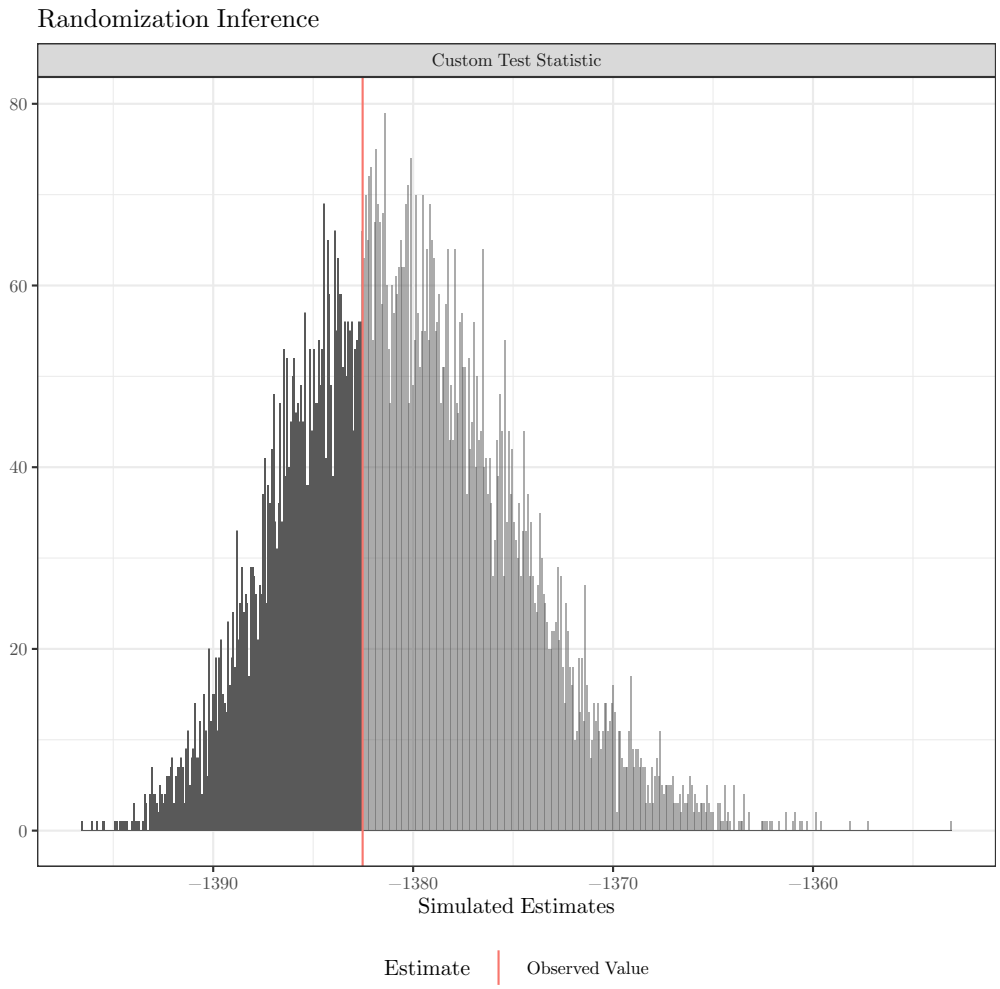


Figure 4.17: Sampling distribution of log-likelihood statistics from 10,000 simulations of the randomization procedure for the UK sample. The red line locates the observed log-likelihood statistic.

U.S. Sample: Correlation of Pre-Treatment Covariates

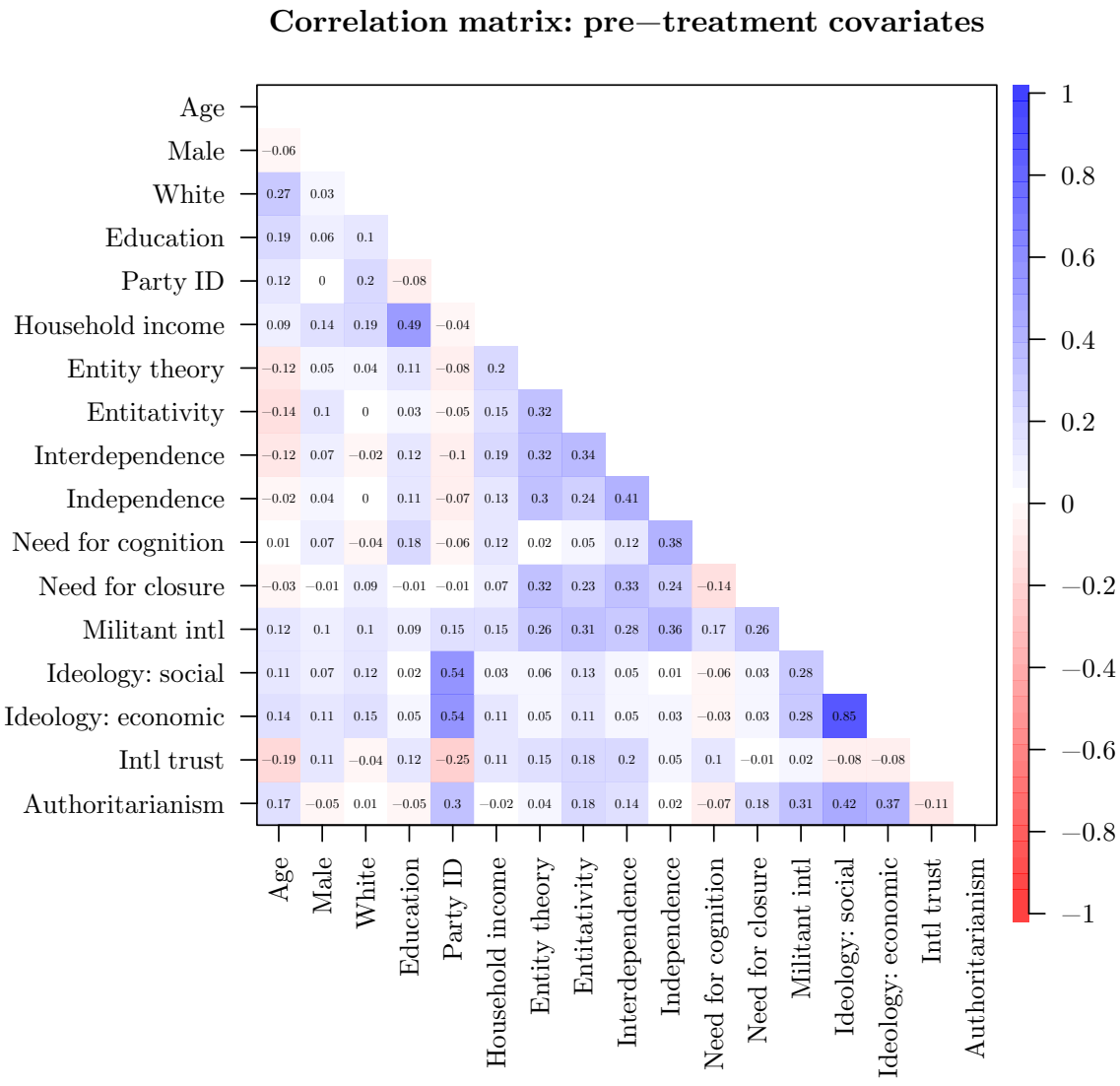
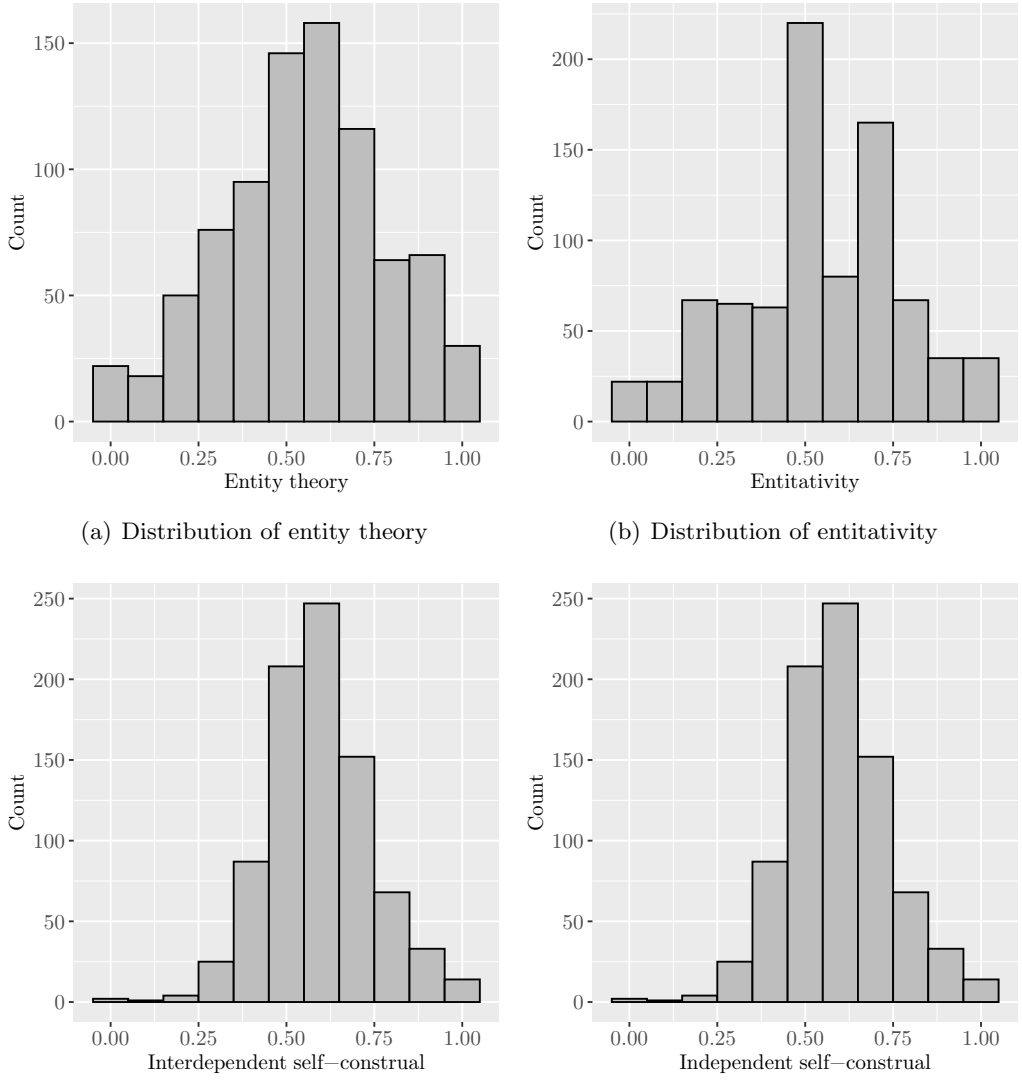


Figure 4.18: A heat map showing the strength of correlation between all pre-treatment covariates used in the analysis. The function psych::corPlot in R automatically calculates Pearson's r (values shown within each cell.) As the covariates are not normally distributed, I also calculated Kendall's τ. The results were similar, with slightly lower values (e.g. $\tau = 0.76$ for the correlation between the two dimensions of political ideology and $\tau = 0.45$ and $\tau = 0.44$ for the correlation between Party ID and the social and economic dimensions of political ideology, respectively.) Note the small to moderate correlations between entity theory, entitativity, interdependent and independent self-construal, cognitive style and militant internationalism.

U.S. Sample: Distribution of Key Covariates



(a) Distribution of entity theory

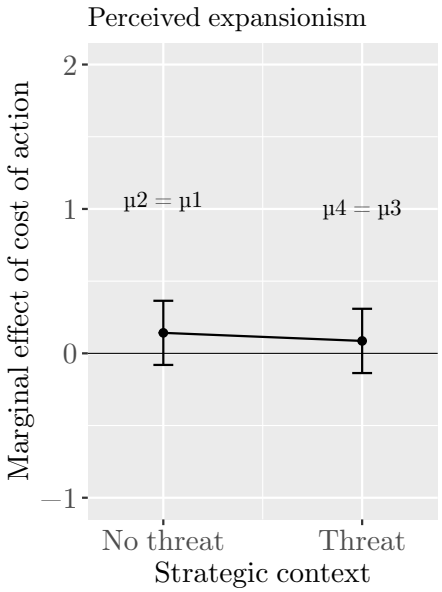
(b) Distribution of entitativity

(c) Distribution of interdependent self-construal

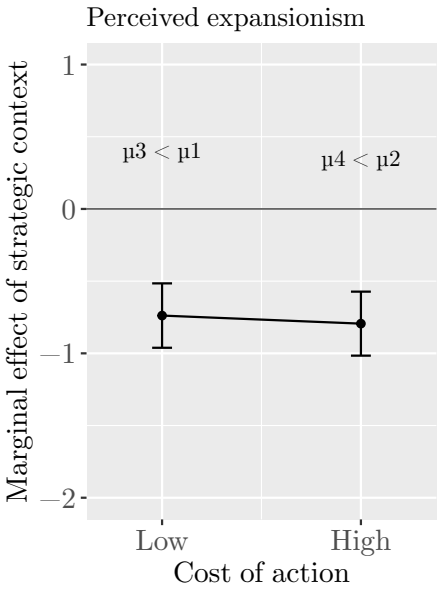
(d) Distribution of independent self-construal

Figure 4.19: Distributions of key covariates in the U.S. sample

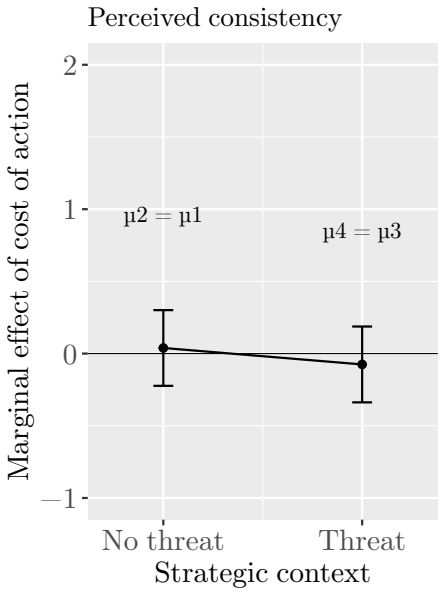
4.7.5 UK Sample: Experimental Findings



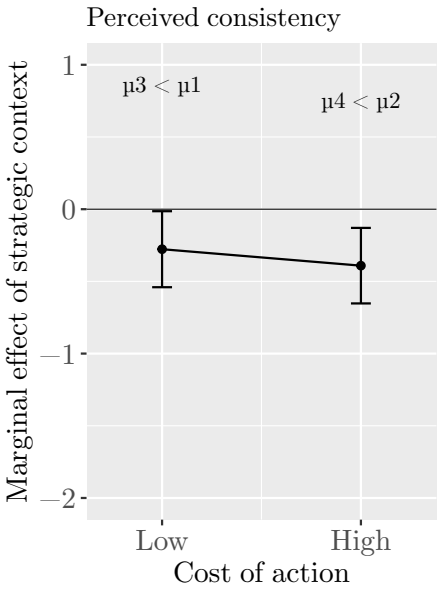
(a) The marginal effect of cost of action on perceived expansionism, by strategic context.



(b) The marginal effect of strategic context on perceived expansionism, by costliness of action.



(c) The marginal effect of cost of action on predicted consistency, by strategic context.



(d) The marginal effect of strategic context on predicted consistency, by costliness of action.

Figure 4.20: The marginal effect of cost of action and strategic context on perceived expansionism and behavioral predictions based on the models 3 and 7 shown in table 4.7.

Table 4.7: Treatment effects with and without covariate adjustment

	Perceived expansionism				Predicted consistency			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sunk cost: high	0.114	0.130	0.142	0.163	−0.018	−0.011	0.039	0.060
	(0.080)	(0.079)	(0.113)	(0.112)	(0.095)	(0.093)	(0.134)	(0.132)
Context: threat	−0.767***	−0.769***	−0.738***	−0.735***	−0.334***	−0.339***	−0.277**	−0.267**
	(0.080)	(0.079)	(0.114)	(0.113)	(0.095)	(0.094)	(0.134)	(0.133)
Militant internationalism		0.352*		0.356*		0.748***		0.756***
		(0.205)		(0.205)		(0.241)		(0.241)
Political ideology: social		0.227		0.229		0.381		0.385
		(0.201)		(0.201)		(0.237)		(0.237)
International trust		0.247		0.248		0.108		0.111
		(0.182)		(0.182)		(0.215)		(0.215)
Right-wing authoritarianism		0.643**		0.641**		0.532*		0.527*
		(0.257)		(0.257)		(0.303)		(0.303)
Need for closure		0.759***		0.756***		0.671**		0.666**
		(0.254)		(0.254)		(0.299)		(0.299)
Need for cognition		0.423*		0.422*		0.262		0.258
		(0.239)		(0.239)		(0.282)		(0.282)
High cost*threat			−0.056	−0.067			−0.114	−0.141
			(0.160)	(0.158)			(0.189)	(0.186)
Constant	3.921***	2.533***	3.907***	2.516***	3.716***	2.307***	3.687***	2.271***
	(0.069)	(0.261)	(0.080)	(0.264)	(0.082)	(0.308)	(0.095)	(0.311)
N	986	986	986	986	986	986	986	986
R^2	0.087	0.121	0.087	0.121	0.013	0.051	0.013	0.052
Adjusted R^2	0.085	0.113	0.084	0.113	0.011	0.043	0.010	0.043

*p < .1; **p < .05; ***p < .01

Table 4.8: Treatment effects with and without covariate adjustment

	Perceived defensiveness		Perceived resolve	
	(1)	(2)	(3)	(4)
Sunk cost: high	−0.064	−0.040	0.046	0.070
	(0.078)	(0.077)	(0.075)	(0.074)
Context: threat	0.639***	0.632***	0.273***	0.270***
	(0.078)	(0.078)	(0.075)	(0.075)
Militant internationalism		−0.098		0.030
		(0.200)		(0.192)
Political ideology: social		0.276		0.122
		(0.197)		(0.189)
International trust		−0.015		0.138
		(0.178)		(0.171)
Right-wing authoritarianism		−0.422*		0.067
		(0.252)		(0.241)
Need for closure		1.041***		0.905***
		(0.248)		(0.238)
Need for cognition		0.949***		1.007***
		(0.234)		(0.224)
Constant	3.680***	2.673***	4.048***	2.787***
	(0.068)	(0.255)	(0.065)	(0.245)
N	986	986	986	986
R^2	0.064	0.096	0.014	0.047
Adjusted R^2	0.062	0.088	0.012	0.039

*p < .1; **p < .05; ***p < .01

4.7.6 Note on Timing of Data Collection

Data collection for this study coincided with the Russian invasion of Ukraine in February 2022. The invasion followed a Russian military build-up along Ukrainian borders since late 2021, and marked the onset of a war that is still ongoing at the time of writing. Given the far-reaching implications of the Ukraine war for international politics, and the extensive media coverage it has received, it was likely to be on the minds of survey-takers. The scenario used in the treatment for this study deliberately abstracts away from current events in international politics, and does not closely resemble the ongoing conflict: the vignette describes a rivalry at sea, not land; the U.S. is directly party to the hypothetical conflict that ensues; and the dispute escalation involves naval deployments – no reference is made to the use of force. If anything, the scenario's semblance to the ongoing conflict decreased when the Russian military build-up escalated into war in February 2022, even as public awareness of the conflict increased.

Nonetheless, it is worth considering whether and how current events may have impacted the study's findings. As for other sample characteristics, random assignment should have ensured that exposure to, and active engagement with, news coverage of the ongoing conflict is similar across treatment groups. Indeed, a balance check confirms that the extent to which survey-takers followed the Ukraine war is well-balanced across treatment groups (see appendix, figure 4.14).[118] Consequently, we should expect the Ukraine war to have a level effect, if any, with no further implications for the internal validity of the study. Moreover, if present, any level effect is likely to lead us to underestimate, not overestimate, the actual treatment effect for two reasons: survey-takers are likely sensitized to the risk of militarized disputes and wider war, resulting in overall higher estimates of adversaries' expansionism; at the same time, the treatment – varying levels of naval deployments to international waters – is likely to pale relative to the use of force, let alone a full-scale invasion into another country.

[118] Note that the relevant item was included in the post-treatment questionnaire, not the pre-treatment questionnaire, in order to avoid any priming effects.

References

Albuyeh, Rod and Mark Paradis (2018). "Thawing Rivalries and Fading Friendships: An Experimental Approach to Rapprochement and Alienation." In: *Political Psychology* 39 (4), pp. 811–827.

American National Election Studies (2021). *The ANES Guide to Public Opinion and Electoral Behavior*. URL: https://electionstudies.org/resources/anes-guide/.

Aronoff, Yael (2014). *The Political Psychology of Israeli Prime Ministers: When Hard-Liners Opt for Peace*. Cambridge University Press.

Baker, Reg et al. (2013). "Summary Report of the AAPOR Task Force on Non-Probability Sampling." In: *Journal of Survey Statistics and Methodology* 1 (2), pp. 90–105.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz (2012). "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." In: *Political Analysis* 20 (3), pp. 351–368.

Bizumic, Boris and John Duckitt (2018). "Investigating Right Wing Authoritarianism With a Very Short Authoritarianism Scale." In: *Journal of Social and Political Psychology* 6 (1), pp. 129–150.

Brambor, Thomas, William Roberts Clark, and Matt Golder (2006). "Understanding Interaction Models: Improving Empirical Analyses." In: *Political Analysis* 14 (1), pp. 63–82.

Brewer, Marilynn B., Ying-yi Hong, and Qiong Li (2004). "Dynamic Entitativity: Perceiving Groups as Actors." In: *The Psychology of Group Perception*. Ed. by Vincent Yzerbyt, Charles M. Judd, and Olivier Corneille. Psychology Press, pp. 19–29.

Brewer, Paul R. et al. (2004). "International Trust and Public Opinion About World Affairs." In: *American Journal of Political Science* 48 (1), pp. 93–109.

Campbell, Donald T. (1958). "Common Fate, Similarity, and Other Indices of the Status of Aggregates of Persons as Social Entities." In: *Behavioral Science* 3 (1), pp. 14–25.

Chiu, Chi-yue, Ying-yi Hong, and Carol S. Dweck (1997). "Lay Dispositionism and Implicit Theories of Personality." In: *Journal of Personality and Social Psychology* 73 (1), pp. 19–30.

Choi, Incheol and Richard E. Nisbett (1998). "Situational Salience and Cultural Differences in the Correspondence Bias and Actor-Observer Bias." In: *Personality & Social Psychology Bulletin* 24 (9), pp. 949–960.

Coppock, Alexander and Oliver A. McClellan (2019). "Validating the Demographic, Political, Psychological, and Experimental Results Obtained From a New Source of Online Survey Respondents." In: *Research & Politics* 6 (1), pp. 1–14.

Dafoe, Allan, Baobao Zhang, and Devin Caughey (2018). "Information Equivalence in Survey Experiments." In: *Political Analysis* 26 (4), pp. 399–416.

D'Agostino, P. R. and R. Fincherkiefer (1992). "Need for Cognition and the Correspondence Bias." In: *Social Cognition* 10 (2), pp. 151–163.

Dickson, Eric S. (2009). "Do Participants and Observers Assess Intentions Differently During Bargaining and Conflict?" In: *American Journal of Political Science* 53 (4), pp. 910–930.

Druckman, James N. and Cindy D. Kam (2011). "Students as Experimental Participants: A Defense of the "Narrow Data Base"." In: *Cambridge Handbook of Experimental Political Science*. Ed. by James N. Druckman et al. Cambridge University Press, pp. 41–57.

Dweck, Carol S., Ying-yi Hong, and Chi-yue Chiu (1993). "Implicit Theories: Individual Differences in the Likelihood and Meaning of Dispositional Inference." In: *Personality & Social Psychology Bulletin* 19 (5), pp. 644–656.

Edward, Jones E. and Richard E. Nisbett (1971). *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior*. General Learning Press.

Faul, Franz et al. (2007). "GPower 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." In: *Behavior Research Methods* 39 (2), pp. 175–191.

Fearon, James D. (1995). "Rationalist Explanations of War." In: *International Organization* 49 (3), pp. 379–414.

Fearon, James D. (1997). "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." In: *Journal of Conflict Resolution* 41 (1), pp. 68–90.

Feldman, Stanley and Christopher Johnston (2014). "Understanding the Determinants of Political Ideology: Implications of Structural Complexity." In: *Political Psychology* 35 (3), pp337–358.

Gerber, Alan S. and Donald P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. W.W. Norton & Company.

Gerber, Alan et al. (2014). "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." In: *Journal of Experimental Political Science* 1 (1), pp. 81–98.

Gilbert, Daniel T. and Patrick S. Malone (1995). "The Correspondence Bias." In: *Psychological Bulletin* 117 (1), pp. 21–38.

Glaser, Charles L. (2004). "When Are Arms Races Dangerous? Rational versus Suboptimal Arming." In: *International Security* 28 (4), pp. 44–84.

Glaser, Charles L. (2010). *Rational Theory of International Politics: The Logic of Competition and Cooperation*. Princeton University Press.

Hall, Todd and Keren Yarhi-Milo (2012). "The Personal Touch: Leaders' Impressions, Costly Signaling, and Assessments of Sincerity in International Affairs." In: *International Studies Quarterly* 56 (3), pp. 560–573.

Heider, Fritz (1958). *The Psychology of Interpersonal Relations*. Wiley.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010). "The Weirdest People in the World?" In: *Behavioral and Brain Sciences* 33 (2-3), pp. 61–83.

Hewstone, Miles (1989). *Causal Attribution : From Cognitive Processes to Collective Beliefs*. Basil Blackwell.

Hewstone, Miles (1990). "The 'Ultimate Attribution Error'? A Review of the Literature on Intergroup Causal Attribution." In: *European Journal of Social Psychology* 20 (4), pp. 311–335.

Holanda Coelho, Gabriel Lins de, Paul H. P. Hanel, and Lukas J. Wolf (2020). "The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version." In: *Assessment* 27 (8), pp. 1870–1885.

Holsti, Ole R. (1962). "The Belief System and National Images: A Case Study." In: *Journal of Conflict Resolution* 6 (3), pp. 244–252.

Jervis, Robert (1976). *Perception and Misperception in International Politics*. Princeton University Press.

Jervis, Robert (1978). "Cooperation under the Security Dilemma." In: *World Politics* 30 (2), pp. 167–214.

Jones, Edward E. and Victor A. Harris (1967). "The Attribution of Attitudes." In: *Journal of Experimental Social Psychology* 3 (1), pp. 1–24.

Kam, Cindy D. and Marc J. Trussler (2017). "At the Nexus of Observational and Experimental Research: Theory, Specification, and Analysis of Experiments with Heterogeneous Treatment Effects." In: *Political Behavior* 39 (4), pp. 789–815.

Kertzer, Joshua D. (2016). *Resolve in International Politics*. Princeton University Press.

Kertzer, Joshua D., Brian C. Rathbun, and Nina Srinivasan Rathbun (2020). "The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations." In: *International Organization* 74 (1), pp. 95–118.

Kertzer, Joshua D., Jonathan Renshon, and Keren Yarhi-Milo (2019). "How Do Observers Assess Resolve?" In: *British Journal of Political Science* 51 (1), pp. 308–330.

Kuehnen, Ulrich et al. (2013). "How Self-construal Affects Dispositionalism in Attributions." In: *Social Cognition* 31 (2), pp. 237–259.

Kurizaki, Shuhei (2007). "Efficient Secrecy: Public versus Private Threats in Crisis Diplomacy." In: *American Political Science Review* 101 (3), pp. 543–558.

Kydd, Andrew H. (1997). "Game Theory and the Spiral Model." In: *World Politics* 49 (3), pp. 371–400.

Kydd, Andrew H. (2005). *Trust and Mistrust in International Relations*. Princeton University Press.

Larson, Deborah Welch (1997a). *Anatomy of Mistrust : U.S.-Soviet Relations During the Cold War*. Cornell University Press.

Larson, Deborah Welch (1997b). "Trust and Missed Opportunities in International Relations." In: *Political Psychology* 18 (3), pp. 701–734.

Levy, Sheri R., Chi-yue Chiu, and Ying-yi Hong (2006). "Lay Theories and Intergroup Relations." In: *Group Processes & Intergroup Relations* 9 (1), pp. 5–24.

Levy, Sheri R., Steven J. Stroessner, and Carol S. Dweck (1998). "Stereotype Formation and Endorsement: The Role of Implicit Theories." In: *Journal of Personality and Social Psychology* 74 (6), pp. 1421–1436.

Lin, Winston, Donald P. Green, and Alexander Coppock (2016). "Standard Operating Procedures for Don Green's Lab at Columbia." URL: https://alexandercoppock.com/Green-Lab-SOP/Green_Lab_SOP.pdf.

Markus, Hazel Rose and Shinobu Kitayama (1991). "Culture and the Self: Implications for Cognition, Emotion, and Motivation." In: *Psychological Review* 98 (2), p. 224.

Masuda, Takahiko and Shinobu Kitayama (2004). "Perceiver-Induced Constraint and Attitude Attribution in Japan and the US: A Case for the Cultural Dependence of the Correspondence Bias." In: *Journal of Experimental Social Psychology* 40 (3), pp. 409–416.

Mattes, Michaela and Jessica L. P. Weeks (2018). "Hawks, Doves, and Peace: An Experimental Approach." In: *American Journal of Political Science* 63 (1), pp. 53–66.

Miller, Joan G (1984). "Culture and the Development of Everyday Social Explanation." In: *Journal of Personality and Social Psychology* 46 (5), pp. 961–978.

Miyamoto, Yuri and Shinobu Kitayama (2002). "Cultural Variation in Correspondence Bias: The Critical Role of Attitude Diagnosticity of Socially Constrained Behavior." In: *Journal of Personality and Social Psychology* 83 (5), pp. 1239–1248.

Molden, Daniel C. and Carol S. Dweck (2006). "Finding "Meaning" in Psychology: A Lay Theories Approach to Self-Regulation, Social Perception, and Social Development." In: *American Psychologist* 61 (3), pp. 192–203.

Molden, Daniel C., Jason E. Plaks, and Carol S. Dweck (2006). ""Meaningful" Social Inferences: Effects of Implicit Theories on Inferential Processes." In: *Journal of Experimental Social Psychology* 42 (6), pp. 738–752.

Morris, Michael W., Richard E. Nisbett, and Kaiping Peng (1996). "Causal Attribution across Domains and Cultures." In: *Causal Cognition: A Multidisciplinary Debate*. Ed. by Dan Sperber, David Premack, and Ann James Premack. Oxford University Press.

Morris, Michael W. and Kaiping Peng (1994). "Culture and Cause: American and Chinese Attributions for Social and Physical Events." In: *Journal of Personality and Social Psychology* 67 (6), pp. 949–971.

Morrow, James D. (1999). "The Strategic Setting of Choices: Signaling, Commitment, and Negotiation in International Politics." In: *Strategic Choice and International Relations*. Ed. by David A. Lake and Robert Powell. Princeton University Press, pp. 77–114.

Mutz, Diana C. and Robin Pemantle (2015). "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods." In: *Journal of Experimental Political Science* 2 (2), pp. 192–215.

Na, Jinkyung et al. (2010). "Cultural Differences Are Not Always Reducible to Individual Differences." In: *Proceedings of the National Academy of Sciences (PNAS)* 107 (14), pp. 6192–6197.

Nisbett, Richard E. (2003). *The Geography of Thought: How Asians and Westerners Think Differently - And Why*. Nicholas Brealey.

Norenzayan, Ara, Incheol Choi, and Richard E. Nisbett (2002). "Cultural Similarities and Differences in Social Inference: Evidence from Behavioral Predictions and Lay Theories of Behavior." In: *Personality & Social Psychology Bulletin* 28 (1), pp. 109–120.

Pettigrew, Thomas F. (1979). "The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice." In: *Personality & Social Psychology Bulletin* 5 (4), pp. 461–476.

Peyton, Kyle, Gregory A. Huber, and Alexander Coppock (2021). "The Generalizability of Online Experiments Conducted During the COVID-19 Pandemic." In: *Journal of Experimental Political Science*, pp. 1–16.

Plaks, Jason E., Sheri R. Levy, and Carol S. Dweck (2009). "Lay Theories of Personality: Cornerstones of Meaning in Social Cognition." In: *Social and Personality Psychology Compass* 3 (6), pp. 1069–1081.

Plaks, Jason E., Sheri R. Levy, Carol S. Dweck, and Steven J. Stroessner (2004). "In the Eye of the Beholder: Lay Theories and the Perception of Group Entitativity, Variability, and Essence." In: *The Psychology of Group Perception*. Ed. by Vincent Yzerbyt, Charles M. Judd, and Olivier Corneille. Psychology Press, pp. 96–109.

Pocock, Stuart J. et al. (2002). "Subgroup Analysis, Covariate Adjustment and Baseline Comparisons in Clinical Trial Reporting: Current Practice and Problems." In: *Statistics in Medicine* 21 (19), pp. 2917–2930.

Quek, Kai (2016). "Are Costly Signals More Credible? Evidence of Sender-Receiver Gaps." In: *The Journal of Politics* 78 (3), pp. 925–940.

Quek, Kai (2021). "Four Costly Signaling Mechanisms." In: *American Political Science Review* 115 (2), pp. 537–549.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. URL: http://www.R-project.org.

Rathbun, Brian C. (2007). "Uncertain about Uncertainty: Understanding the Multiple Meanings of a Crucial Concept in International Relations Theory." In: *International Studies Quarterly* 51 (3), pp. 533–557.

Rathbun, Brian C. (2019). *Reasoning of State: Realists, Romantics and Rationality in International Relations*. Cambridge University Press.

Rathbun, Brian C., Joshua D. Kertzer, and Mark Paradis (2017). "Homo Diplomaticus: Mixed-Method Evidence of Variation in Strategic Rationality." In: *International Organization* 71 (S1), S33–S60.

Renshon, Jonathan, Allan Dafoe, and Paul Huth (2018). "Leader Influence and Reputation Formation in World Politics." In: *American Journal of Political Science* 62 (2), pp. 325–339.

Roets, Arne and Alain Van Hiel (2010). "Item Selection and Validation of a Brief, 15-item Version of the Need for Closure Scale." In: *Personality and Individual Differences* 50 (1), pp. 90–94.

Rosati, Jerel A. (1988). "Continuity and Change in the Foreign Policy Beliefs of Political Leaders: Addressing the Controversy over the Carter Administration." In: *Political Psychology* 9 (3), pp. 471–505.

Rosenberg, Shawn W. and Gary Wolfsfeld (1977). "International Conflict and the Problem of Attribution." In: *Journal of Conflict Resolution* 21 (1), pp. 75–103.

Ross, Lee (1977). "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process." In: *Advances in Experimental Social Psychology* 10, p. 174.

Rydell, Robert J. et al. (2007). "Implicit Theories About Groups and Stereotyping." In: *Personality & Social Psychology Bulletin* 33 (4), pp. 549–558.

Sabini, John, Michael Siepmann, and Julia Stein (2001). "The Really Fundamental Attribution Error in Social Psychological Research." In: *Psychological Inquiry* 12 (1), pp. 1–15.

Shamon, Hawal and Carl Berning (2020). "Attention Check Items and Instructions in Online Surveys with Incentivized and Non-Incentivized Samples: Boon or Bane for Data Quality?" In: *Survey Research Methods* 14 (1), pp. 55–77.

Singelis, Theodore M. (1994). "The Measurement of Independent and Interdependent Self-Construals." In: *Personality & Social Psychology Bulletin* 20 (5), pp. 580–591.

Slantchev, Branislav L. (2005). "Military Coercion in Interstate Crises." In: *American Political Science Review* 99 (4), pp. 533–547.

Talhelm, Thomas (2018). "Hong Kong Liberals Are WEIRD: Analytic Thought Increases Support for Liberal Policies." In: *Personality & Social Psychology Bulletin* 44 (5), pp. 717–728.

Talhelm, Thomas et al. (2015). "Liberals Think More Analytically (More "WEIRD") Than Conservatives." In: *Personality & Social Psychology Bulletin* 41 (2), pp. 250–267.

Tang, Shiping (2009). "The Security Dilemma: A Conceptual Analysis." In: *Security Studies* 18 (3), pp. 587–623.

Taylor, Donald M. and Vaishna Jaggi (1974). "Ethnocentrism and Causal Attribution in a South Indian Context." In: *Journal of Cross-Cultural Psychology* 5 (2), pp. 162–171.

Ternovski, John and Lilla Orr (2022). "A Note on Increases in Inattentive Online Survey-Takers Since 2020." In: *Journal of Quantitative Description: Digital Media* 2, pp. 1–35. URL: https://journalqd.org/article/view/2985.

Tetlock, Philip (1999). "Theory-Driven Reasoning about Possible Pasts and Probable Futures in World Politics: Are We Prisoners of our Preconceptions?" In: *American Journal of Political Science* 43 (2), p. 335.

Tingley, Dustin H. and Barbara F. Walter (2011). "Can Cheap Talk Deter? An Experimental Analysis." In: *Journal of Conflict Resolution* 55 (6), pp. 996–1020.

Trager, Robert F. (2010). "Diplomatic Calculus in Anarchy: How Communication Matters." In: *American Political Science Review* 104 (2), pp. 347–368.

Trager, Robert F. (2017). *Diplomacy: Communication and the Origins of International Order*. Cambridge University Press.

Vertzberger, Yaacov (1990). *The World in Their Minds: Information Processing, Cognition and Perception in Foreign Policy Decisionmaking*. Stanford University Press.

Webster, Donna M. and Arie W. Kruglanski (1994). "Individual Differences in Need for Cognitive Closure." In: *Journal of Personality and Social Psychology* 67 (6), pp. 1049–1062.

Willert, Morten Vejs (2010). *A CONSORT-Style Flowchart of a Randomized Controlled Trial*. URL: https://texample.net/tikz/examples/consort-flowchart/.

Yarhi-Milo, Keren (2014). *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations*. Princeton University Press.

Yarhi-Milo, Keren, Joshua D. Kertzer, and Jonathan Renshon (2018). "Tying Hands, Sinking Costs, and Leader Attributes." In: *Journal of Conflict Resolution* 62 (10), pp. 2150–2179.

Yzerbyt, Vincent Y., Anouk Rogier, and Susan T. Fiske (1998). "Group Entitativity and Social Attribution: On Translating Situational Constraints into Stereotypes." In: *Personality & Social Psychology Bulletin* 24 (10), pp. 1089–1103.

5 Who Responds to Costly Signals? Implicit Theories, Attribution, and the End of Detente

Abstract

Signaling about motives is central in identifying opportunities for cooperation and anticipating points of escalation among states in international politics. However, foundational questions remain, including the question of how decision makers draw inferences about a state's motives and, crucially, why decision makers systematically vary in the timing and substance of belief updating about motives. I offer a domain-specific theory of attribution that unifies rationalist, psychological, and constructivist accounts of belief updating about motives in international politics. Integrating scholarship on costly signaling with findings from attribution research in psychology, I theorize that individuals' responses to signals will vary depending on a signal's informational structure and the individual's implicit theory of the sender. I derive a typology of implicit theories from two underlying assumptions about the sender: whether the sender is a unified actor in international politics, and whether actors generally are considered to be mutable. To test the theory, I examine the responses of three key decision makers to Soviet signals during the Carter Administration. The findings suggest that, while decision makers' inferences about motives are subjective, they are neither static nor idiosyncratic: decision makers do respond to signals, yet how they update their beliefs about motives is shaped by their implicit theory – with significant implications for their policy preferences and, ultimately, prospects for cooperation between states.

5.1 Introduction

How do decision makers infer the motives of actors in international politics? Hardly any event since the end of the Cold War has done more to highlight the theoretical and practical centrality of this question in international relations than the Russian military build-up along Ukrainian borders in 2021 and Russia's invasion of Ukraine in early 2022. Within months, the international community witnessed a dramatic shift from pre-war speculation about the likelihood of an invasion to ongoing assessments of the extent of Russian territorial ambitions and military resolve. Already, updated beliefs about Russia's motives have prompted a remarkable restructuring of Europe's security order, including, most recently, Finland's and Sweden's accession to NATO. Moreover, as the U.S. and European powers debate the nature and extent of their military support to Ukraine, assessments of Russian motives, as well as NATO's ability to communicate its own motives to Russia clearly and credibly, have become a focal point of efforts to prevent an escalation into wider war.

Given the importance of inferences about motives, it is unsurprising that the question of how states assess them has figured centrally in theoretical scholarship on signaling and bargaining in international relations. By contrast, the systematic, empirical study of how states assess motives is relatively recent. This growing body of empirical work challenges a key assumption of theoretical scholarship – the assumption that states rely on costly signals to convey their own motives credibly and to infer another state's motives accurately. Costly signals are actions that incur a significant cost to the sender and, therefore, are less likely to be taken by bluffers than sincere states, making them a reliable indicator of a state's motives. However, contrary to theoretical expectations, empirical work finds that individuals vary in the conclusions they draw from costly signals, both across individuals and within individuals over time, raising the fundamental question: how do individuals, including decision makers, infer a state's motives from its signals? Why do their responses to costly signals vary?

Existing research broadly offers two approaches to answering this question and explaining individual-level variation in belief updating about motives. Consistency theories argue that individuals process novel information, including signals, in ways that serve to

uphold their prior beliefs about a particular state, international politics in general, or both. On this view, the most important distinction is, not between costly and costless signals, but between belief-consistent and belief-inconsistent signals. While consistency theories vary in the exact mechanism they describe to connect prior to posterior beliefs, they agree that decision makers will respond to belief-consistent signals, while minimizing or dismissing belief-inconsistent signals. Consequently, we should expect decision makers' beliefs about a state's motives to remain relatively stable over time, and we should not expect their beliefs to converge in response to costly signals.

By contrast, personalized theories seek to explain variation in responses to costly signals while recognizing the possibility of change in individuals' beliefs about motives. They do so by focusing on individual differences in cognitive style or in beliefs about what constitutes an informative indicator of a state's motives. By emphasizing variation in information processing across individuals, personalized theories are able to explain belief change and variation in individuals' propensity toward belief change. However, they do not provide an account of why individuals' inferences from costly signals vary dynamically, giving rise to systematic variation in the timing of belief change, not merely individuals' propensity toward belief change.

In this article, I develop a third approach to explaining variation in the conclusions individuals draw from costly signals and the way individuals update their beliefs about a state's motives. Integrating rationalist scholarship on signaling with attribution research in psychology, the theory posits that the inferences individuals draw from costly signals will depend on the interaction between their tacit assumptions about states – their 'implicit theories' of states – and the informational structure of a signal. It theorizes that individuals' implicit theories shape the extent to which they attend to a sender's strategic context, and the extent to which they attribute motives to states as a whole or to individual leaders. As a result, individuals' beliefs about motives will dynamically vary, sometimes diverging and other times converging in response to costly signals, depending on the signal's informational structure.

As an initial test of the theory, I conduct a qualitative analysis of elite decision makers' evolving perceptions of Soviet motives during the Carter Administration. I focus on three decision makers' in particular: President Jimmy Carter, National Security Advisor Zbig-

niew Brzezinski, and Secretary of State Cyrus Vance. Combining comparative analysis with process tracing, I analyze how each decision maker updated their beliefs about Soviet motives in response to three instances of sunk-cost signaling: the Soviet intervention in the Ogaden War, the stationing of a Soviet brigade in Cuba, and the Soviet invasion of Afghanistan. To do so, I rely on verbal and behavioral indicators of decision makers' beliefs and process of belief updating, and draw on a rich secondary literature on the Carter Administration, as well as primary documents like records from policy discussions, diaries, memoirs, and interviews. The findings support the domain-specific theory of attribution I advance, showing that decision makers' implicit theories of the Soviet Union can explain why their inferences from sunk-cost signals varied dynamically, leading to differences in the timing and substance of belief updating about Soviet motives.

Although motives are central to international relations theory, as a subject matter and an unit of analysis, they are not defined clearly within the field: whereas some scholars distinguish 'motives' from 'intentions' or 'resolve',[1] other scholars deliberately conflate these terms to varying degrees.[2] It is therefore worth clarifying that, for the purpose of studying attribution in international politics, this paper defines motives as a state's goals in international politics, thereby distinguishing it from the specific foreign-policy preferences a state formulates or policies and strategies a state adopts to achieve those goals.[3]

The article proceeds as follows. First, I review existing approaches to explaining how decision makers form and update beliefs about states's motives. Second, building on insights from psychology and integrating them with rationalist insights into signaling in international politics, I lay out a new approach, a domain-specific theory of attribution. Third, I outline the research design for evaluating the theory based on three cases and present the findings from the case studies. Finally, I conclude by discussing the implications of my findings for scholarship on signaling and the role of meaning in psychological approaches to international relations.

[1] For example, Glaser 2010; Kertzer 2016; Rosato 2021.
[2] For example, Jervis 1976; Yarhi-Milo 2014.
[3] For a discussion, see chapter 2, section 2.2.

5.2 Costly Signaling and the Puzzle of Attribution

The question of how actors in international politics infer a state's motives is central to international relations theory: motives cannot be observed directly, and international politics is ripe with incentives for states to purposely conceal or misrepresent them.[4] Some scholars therefore conclude that motives are fundamentally unknowable, and actors are caught in a state of perpetual conflict where, '[f]rom the perspective of any one great power, all other great powers are potential enemies.'[5] On this view, international relations theory is a grammar of the inevitable: it seeks to uncover the laws that govern a complex, yet fundamentally immutable, system. By contrast, many scholars in international relations agree only with the premise of this argument, not its conclusion: they consider motives to be a key source of uncertainty in international politics, but not an insurmountable obstacle to sustained cooperation.[6] On this view, international relations theory can generate insights into the scope conditions and mechanisms of cooperation and, thereby, help facilitate it.[7]

According to theoretical scholarship on signaling and bargaining, a chief mechanism that states rely on to reduce uncertainty about motives and cooperate is costly signaling. Costly signals are actions that incur a significant cost to the sender and, therefore, are less amenable to bluffing than actions that incur a low cost or no cost at all.[8] Whether a costly signal indicates malign or benign motives, scholars generally distinguish between two types of costly signals based on the type of cost they incur. Sunk-cost signals incur ex ante costs; these signals are 'costly for the state to take in the first place but do not affect the relative value of fighting versus acquiescing in a challenge'.[9] A common example is the decision to mobilize troops.[10] By contrast, tying-hands signals incur ex post costs; these signals 'increas[e] the costs of backing down if the would-be challenger actually challenges but otherwise entail[ ] no costs if no challenge materializes'.[11] Examples include publicly

[4] Fearon 1994
[5] Mearsheimer 2014, p. 34; Rosato 2021; Rosato 2015; Waltz 1979
[6] Glaser 1995; Keohane 1984; Kydd 2005; Snidal 1991.
[7] Although note Gartzke 1999.
[8] cite
[9] Fearon 1997, p. 70.
[10] Strictly speaking, decisions to mobilize troops are not pure instances of sunk-cost signaling, since they also change the ex post costs of fighting versus backing down (Fearon 1997; Slantchev 2005).
[11] Fearon 1997, p. 70.

stated threats or alliance commitments.[12] Because costly signals are inherently more credible than costless signals, theoretical scholarship leads us to expect that actors in international politics will respond to them by updating their beliefs about the sender's motives and that actors' beliefs will converge over time.[13]

Despite the centrality of motives in international relations theory, scholars have begun to investigate the empirical foundations of the costly signaling hypothesis systematically only recently.[14] Combining experimental and qualitative methods, they find mixed evidence of costly signaling. In line with theoretical expectations, several studies show that individuals update their beliefs about a sender's motives in response to costly signals.[15] However, scholars also find considerable heterogeneity in the extent to which individuals do.[16] For example, in an experimental study of the Iranian nuclear deal, individuals varied in the extent to which they updated their beliefs about Iranian motives in response to costly signals of reassurance, with some responding strongly and others not responding at all.[17] This suggests that, contrary to costly signaling arguments, the process of belief formation and updating about motives is not uniform across actors in international politics; it varies.[18] What explains this variation? How do individuals, including decision makers, infer a state's motives from costly signals?

Existing research broadly offers two approaches to answering these questions and explaining systematic variation in the inferences individuals draw from costly signals: 'consistency theories' and what I call 'personalized theories'. Consistency theories take their cue from 'cognitive consistency theories' in psychology, which originated in the 1950s.[19] They posit that individuals experience 'the existence of nonfitting relations among cognitions' – that is, 'any knowledge, opinion, or belief about the environment, about oneself, or about one's behavior' – as uncomfortable.[20] According to Leon Festinger, an important

[12] More recently, Quek 2021 has proposed two additional types of costs: installment costs and reducible costs.
[13] Some scholars advance the costly signaling argument as part of a purely normative theory of international relations, in which case we would not necessarily expect the logic of costly signaling to be borne out empirically (for example, Glaser 2010).
[14] For example, Brutger and Kertzer 2018; Tingley and Walter 2011; Trachtenberg 2012; Trager 2017; Yoder and Haynes 2021.
[15] Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Quek 2021.
[16] B. C. Rathbun, Kertzer, and Paradis 2017; Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Yarhi-Milo, Kertzer, and Renshon 2018.
[17] Kertzer, B. C. Rathbun, and N. S. Rathbun 2020.
[18] For a seminal study exploring this variation, see Yarhi-Milo 2014.
[19] Festinger 1957.
[20] Festinger 1957, p. 3.

proponent of cognitive consistency theories, individuals are thus motivated to reduce or altogether avoid such 'nonfitting relations' or states of 'cognitive dissonance'.[21] Drawing on this body of work in psychology, consistency theories in international relations argue that an individual's prior beliefs shape the way he or she perceives signals and behaves in international politics.[22] While proponents of consistency theories have proposed a range of concepts to characterize the prior beliefs in question,[23] as well as distinct mechanisms to link those beliefs to an individual's perceptions and behaviors, they share the same basic prediction: individuals will attend to information if it is consistent with their prior beliefs, but discredit, minimize, or disregard it otherwise.[24]

While this prediction, and consistency theories generally, are not necessarily incompatible with rationalist assumptions about information processing,[25] they do depart from rationalist arguments about costly signaling: individuals, they predict, will not respond to costly signals uniformly; their responses are subjective and will vary depending on their prior beliefs. For example, as Keren Yarhi-Milo shows, differences in elite decision makers' images of the Soviet Union explain why they differed in their responses to Soviet signals during the Carter and Reagan Administrations.[26] Similarly, in their experimental study of the Iranian nuclear deal, Kertzer, Rathbun and Rathbun show that the conclusions individuals draw from costly signals vary depending on individuals' foreign-policy attitudes: individuals with more cooperative views of international politics update their

[21] Festinger 1957.

[22] For example, Jervis 1976; Rosati 1988; Tetlock 1999; Vertzberger 1990.

[23] These include the 'operational code', which focuses on beliefs about 'the fundamental nature of politics, the nature of political conflict, the role of the individual in history' as well as 'beliefs about ends-means relationships in the context of political action' (George 1969, p. 199; for the original concept, see Leites 1951; for applications of operational code analysis to specific decision makers, see O. Holsti 1970; Walker 1977; Walker and Schafer 2000; Renshon 2008; Kai and Huiyun 2013); research on foreign-policy attitudes focuses on beliefs about the nature of international politics and the relative efficacy of different foreign policy tools (Chittick, Billingsley, and Travis 1995; O. R. Holsti and Rosenau 1990; O. R. Holsti and Rosenau 1986; Hurwitz and Peffley 1990; Wittkopf 1990); research on 'images' focuses more narrowly on beliefs about a specific country and, famously, includes work on 'enemy images' (Boulding 1959; Bronfenbrenner 1961; Herrmann et al. 1997; O. R. Holsti 1962; Hurwitz and Peffley 1990; Shimko 1991; Silverstein and Flamenbaum 1989; White 1965); 'schemas' may vary in their level of abstraction and generality (Herrmann et al. 1997; Larson 1994).

[24] For example, Jervis 1976; Lebow and J. G. Stein 2019; Herrmann et al. 1997; Renshon 2008.

[25] For example, early work on consistency theories emphasized the compatibility of its arguments with Herbert Simon's notion of 'bounded rationality', which is now widely accepted within rationalist scholarship (for examples, see George 1969; Jervis 1976; Renshon 2008; for the original concept, see Simon 2017; for rationalist scholarship that adopts the concept, see Keohane 1984). More recently, Kertzer, B. C. Rathbun, and N. S. Rathbun 2020 have introduced the concept of 'asymmetric updating' into scholarship on signaling. Asymmetric updating incorporates consistency theories' central prediction with Bayesian models of belief updating (also see Taber and Lodge 2006).

[26] Yarhi-Milo 2014; Yarhi-Milo 2013.

beliefs about Iranian motives more in response to costly signals of reassurance than individuals with less cooperative views, reinforcing their prior beliefs about international politics.[27] As a result, an individual's beliefs about motives will remain relatively stable over time and costly signals will not necessarily result in belief convergence across individuals with distinct prior beliefs.

The key strength of consistency theories lies in their ability to explain why responses to costly signals systematically vary across individuals. However, their ability to do so comes at a cost: they emphasize belief stability; belief change, and the mechanisms that drive it, remain undertheorized.[28] Where consistency theories do explain instances of belief change, they typically resort to external shocks, i.e. dramatic events that cannot be reconciled with prior beliefs and, therefore, lead to sudden shifts in an individual's beliefs. This results in relatively ad-hoc, descriptive accounts of belief change with limited predictive value.[29] More problematically yet, these accounts raise the question of why systematic variation in beliefs about motives may persist once an external shock has occurred. Why do external shocks not necessarily lead to belief convergence? For example, why did elite decision makers' beliefs about Soviet motives only partly converge in response to the Soviet invasion of Afghanistan in late 1979?[30]

Personalized theories offer an answer to this question. Whereas consistency theories assume that individuals uniformly tend toward belief stability, personalized theories posit variation in the extent to which they do and, more broadly, the way individuals update their beliefs about motives. Specifically, building on findings from psychology,[31] international relations scholars have argued that individuals with distinct cognitive styles will differ in the extent to which they are receptive to new information and, thus, susceptible to belief change. For example, Yael Aronoff explains variation in belief change among Israeli prime ministers with similar prior beliefs by examining individual differences in their perceptions of time, risk propensity, cognitive flexibility, and emotional intelligence.[32] In a similar vein, Brian Rathbun distinguishes between two types of deci-

[27] Kertzer, B. C. Rathbun, and N. S. Rathbun 2020. Another recent example is Yarhi-Milo, Kertzer, and Renshon 2018.
[28] Renshon 2008.
[29] For example, Lebow and J. G. Stein 2019, p. 96.
[30] For example, Rosati 1988.
[31] For example, Kruglanski and Webster 1996; Rokeach 1960.
[32] Y. Aronoff 2014.

sion makers in international politics based on their cognitive style: 'realists' display high levels of procedural rationality; 'romantics' feature low levels of procedural rationality. Rathbun conceptualizes procedural rationality as 'epistemic motivation', a concept psychologists use to describe 'a commitment to rational thought',[33] and operationalizes it as an individual's 'need for closure' and 'need for cognition'.[34] According to Rathbun, these differences in cognitive style give rise to differences in belief updating between realists and romantics: whereas romantics tend toward belief stability, realists are highly responsive to new information.[35] Alternatively, Keren Yarhi-Milo has argued that belief updating about motives varies across individuals depending on what they subjectively perceive to be a credible indicator of a state's motives.[36]

Whether scholars focus on variation in cognitive style or individuals' subjective beliefs about what constitutes a credible signal of motives, personalized theories add considerable explanatory power to recipient-driven accounts of signaling: by highlighting variation in individuals' propensity toward belief stability and the process of belief updating about motives, they can explain instances of belief change and, importantly, why belief change occurs unevenly across individuals. An important drawback, however, is that they are highly individualistic. Moreover, although personalized theories provide scholars with an account of belief change, they do not explain why individuals' responses to costly signals vary dynamically, that is, why the timing of belief change, not merely the propensity toward belief change, systematically varies across individuals.[37]

5.3 Attribution in International Politics

Integrating rationalist scholarship on signaling with attribution research in psychology, I offer a third approach to explaining variation in individuals' responses to costly signals, a domain-specific theory of attribution. This theory confers the advantages of both

[33] B. C. Rathbun 2019, p. 22.
[34] Kruglanski and Webster 1996; Cacioppo and Petty 1982.
[35] B. C. Rathbun 2019.
[36] Yarhi-Milo 2013; Yarhi-Milo 2014. Alongside belief consistency, this forms one of several mechanisms that underpin Yarhi-Milo's 'subjective credibility hypothesis'.
[37] An important exception to this, and to the distinction I make between consistency theories and personalized theories, is Yarhi-Milo, Kertzer, and Renshon 2018 who link foreign policy attitudes to differences in the *type* of costly signal individuals respond to, showing that foreign policy 'hawks' are more responsive to sunk-cost signals and foreign policy 'doves' are more responsive to tying-hands signals.

consistency and personalized theories: like consistency theories, it focuses on socially-constructed meaning as a source of variation in belief updating about motives and is thus able to explain systematic variation in individuals' responses to costly signals; like personalized theories, it posits variation in the *process* of belief updating and can explain, not only why individuals' responses to costly signals vary, but also why beliefs about motives can change. However, unlike either approach, the theory argues that socially-constructed meaning is not a mere input into uniform processes, acting as a 'filter' that allows belief-consistent information in and keeps belief-inconsistent information out, but that socially-constructed meaning fundamentally shapes the way information is processed and beliefs are formed. As a result, the theory is able to explain why individuals' responses to costly signals vary *dynamically*, giving rise to systematic variation in the timing of belief change across individuals.

A central premise of the theory is the idea that the problem of motives in international politics – the question of how actors infer states' unobservable motives from their behavior – entails the problem of attribution. A key insight of rationalist theories is that interactions in international politics are characterized by strategic interdependence, i.e. 'the best course of action for each participant [state] depends on what he expects the other participants [other states] to do'.[38] Consequently, a state is never simply acting on its goals: it is always acting on its goals *in light of the incentive structures created by its strategic environment*, including other states' likely responses to its actions. In his seminal contribution to international relations theory, Charles Glaser highlights this feature of international politics by distinguishing between 'motives' and 'intentions': whereas motives refer to a state's fundamental goals and '[...] are inherent features of states', intentions refer to a state's behavior, i.e. 'what a state intends to do', and '[...] result from the interaction of a state with its international environment'.[39] Strategic interdependence generates incentives for states with distinct motives to behave similarly under certain circumstances, giving rise to the problem of motives.[40]

According to rationalist theories, costly signals help solve this problem: in virtue of being prohibitively costly for one type of state but not another, they are credible indicators

[38] Schelling 1960, p. 9; also see Morrow 1999.
[39] Glaser 2010, p. 38; also see Jervis 1976, chapter 2.
[40] Fearon 1995; Glaser 2010; Schelling 1960; Morrow 1999.

of a state's motives.[41] However, the rationalist solution to the problem of motives does not escape the logic of strategic interdependence: how costly a signal is depends on the state's strategic environment and is relative to the state's alternative courses of action in light of that environment. This is strikingly apparent in the case of tying-hands signals, like verbal threats: a verbal threat is not itself costly to make; the cost, and thus credibility, of the signal arises from its strategic implications for the sender.[42] In the case of sunk-cost signals, the sender's action itself incurs a cost, however, the action's credibility as an indicator of motives similarly hinges on the sender's strategic context. For example, a state's decision to mobilize its troops is less costly when the state faces an incursion into its territory than when the state faces no external threat. The latter decision is more likely to be taken by a state with malign motives than a state with benign motives and is thus informative; the former is likely to be taken by any state under attack and is thus far less informative. Put simply, costly signals solve the problem of motives only insofar as they reflect the underlying logic of strategic interdependence.

The upshot of this is that costly signals do not solve the deeper problem of attribution that strategic interdependence entails: to infer a state's motives accurately, actors must distinguish the '[...] internal from situational components' of a state's behavior, i.e. disentangle a state's motives from the incentive structures of its environment.[43] Because the cost of a signal does not inhere in a state's actions, but emerges from its strategic environment, the costly signaling mechanism is premised on, rather than explains, actors' ability to do so: actors must distinguish the '[...] internal from situational components' of a state's behavior and relate the state's behavior to those situational components in order to gauge a signal's cost in the first place. In other words, costly signaling presumes a solution to, rather than solves, the problem of attribution that underlies the problem of motives in international politics.

Recognizing the importance of attribution in signaling, several international relations scholars have turned to psychology to understand the problem of motives.[44] Research on attribution in psychology dates back to the 1930s and 1940s, when European scholars,

[41] Fearon 1997; Glaser 2010; Morrow 1999.
[42] Fearon 1997.
[43] Jervis 1976, p. 48.
[44] Dickson 2009; Jervis 1976; Johnson 2020; Larson 1997a; Lebow and J. G. Stein 2019; Mercer 1996; Rosenberg and Wolfsfeld 1977.

notably Fritz Heider, began to study the 'common sense' or 'naive' psychology individuals employ to explain human behavior and navigate social interactions.[45] It gained traction in the 1960s when two American scholars, Harold Kelley and Edward Jones, popularized attribution research in American social psychology.[46] Since then, attribution theory has been studied and applied across a vast array of research areas within psychology. While attribution research does not offer a single, unifying analytical framework, a central concept that forms the starting point for most studies is the distinction between 'dispositional' and 'situational' causes of behavior to investigate attributional patterns.[47] Dispositional causes, like motives or traits, are 'internal' to the actor and characterized by cross-situational and temporal consistency; situational causes, like peer pressure, hardship, or political coercion, are 'external' to the actor and lead to behavior that is highly contextual. Only dispositional causes afford behavioral predictions *across situations*.[48] Attribution research investigates when, and why, individuals attribute another's behavior to dispositions ('dispositional attribution') versus situations ('situational attribution'), and explores the implications of attributional patterns for phenomena ranging from learning outcomes to stereotype formation and intergroup conflict.[49]

In drawing on attribution research, international relations scholars have focused on three canonical findings from psychology – the 'fundamental attribution error', 'actor-observer bias', and the 'intergroup bias'.[50] The fundamental attribution error describes individuals' tendency to attribute behavior to dispositions even when situational factors are highly salient. For example, in a classic experiment, participants tend to attribute the political views expressed in an article to the article's author, even when the author was coerced into expressing those views.[51] The actor-observer bias describes individuals' tendency to make dispositional attributions for another's behavior, but situational attri-

[45] Heider 1958.
[46] Weiner 2008; Hewstone 1989.
[47] Hewstone 1989. Scholars have also distinguished causes based on how stable and how controllable they are, see Pettigrew 1979; Weiner 1985.
[48] The distinction between disposition and situation is as controversial as it is important in social psychology (for example, see Sabini, Siepmann, and J. Stein 2001).
[49] Molden and Dweck 2006; Pettigrew 1979; Weiner 1985.
[50] The intergroup bias is also known as the 'ultimate attribution error', a term originally coined by Pettigrew 1979. However, as the reknown social psychologist Miles Hewstone has pointed out, the ultimate attribution error is a misnomer, since the inferences are not necessarily erroneous (Hewstone 1990). For studies in international relations that cite these findings see Dickson 2009; Gries 2014; Jervis 1976; Johnson 2020; Larson 1997a; Lebow and J. G. Stein 2019; Mercer 1996; Rosenberg and Wolfsfeld 1977.
[51] Ross 1977; Jones and Harris 1967; Gilbert and Malone 1995

butions for their own behavior.[52] Finally, intergroup bias broadly describes individuals' tendency to make more dispositional attributions for positive behaviors, and more situational attributions for negative behaviors, by ingroup members than outgroup members.[53]

Applying these findings to signaling, international relations scholars have argued that the psychology of attribution exacerbates the problem of motives: first, actors will attribute a state's *competitive policies* to its dispositions and underestimate the extent to which the state is responding to their own policies; second, actors will attribute a state's *cooperative policies* to situational factors and overestimate the extent to which the state is responding to their own policies; third, actors will overestimate a state's awareness of the situational factors driving their own policies.[54] Consequently, international relations scholars predict that actors' existing, benign or malign, beliefs about a state's motives will become entrenched, increasing the risk of arms races and rendering conflicts more intractable.[55] More broadly, they predict that states will be prone to miscalculation – for example, overestimating their deterrence success – which in turn increases the overall risk of bargaining failure.[56]

However, in making these predictions, international relations scholars have tacitly made two assumptions about the findings from psychology that more recent research calls into question. First, scholars have assumed that these canonical findings uniformly apply to distinct targets of attribution and, more broadly, that information processing is similar for individual and group targets. For example, the fundamental attribution error and the self-serving bias are findings from attribution research focusing on individuals, however, studies citing either bias in international relations apply them to states.[57] Similarly, while the intergroup bias recognizes the role of group membership and intergroup relations (i.e. ally or adversary) in attribution, studies investigating the bias have focused on individual group members, not groups as a whole, as the targets of attribution.[58] However,

[52] Edward and Nisbett 1971.
[53] This bias is also known as the 'ultimate attribution error'. In its original formulation – the version most commonly cited by international relations scholars – the bias also predicts the inverse pattern: individuals will tend to make more situational attributions for negative behaviors, and more dispositional attributions for positive behaviors, by outgroup than ingroup members (see D. M. Taylor and Jaggi 1974; Pettigrew 1979; Rosenberg and Wolfsfeld 1977). However, in a review analysis, Hewstone 1990 finds consistent support only for the former, not the latter, prediction.
[54] Dickson 2009; Jervis 1976; Larson 1997a; Mercer 1996; Rosenberg and Wolfsfeld 1977.
[55] Dickson 2009; Larson 1997a; Rosenberg and Wolfsfeld 1977.
[56] Jervis 1976; Rosenberg and Wolfsfeld 1977.
[57] For example, Larson 1997a; Mercer 1996.
[58] An important exception to this is Rosenberg and Wolfsfeld 1977.

research on group perception in psychology finds important differences in attribution and impression formation for individual and group targets.[59]

Second, international relations scholars have assumed that the canonical findings from attribution research uniformly apply to recipients of signals in international politics. However, a large literature in psychology shows that attributional patterns systematically vary across individuals. Notably, studies in cultural psychology find that participants from non-Western societies are less susceptible to the attributional tendencies described above than participants from Western societies.[60] Scholars have sought to explain these differences based on culturally-induced differences in thinking style[61] and individuals' conceptions of the 'self',[62] among others. However, attributional patterns vary, not only across societies, but also within societies.[63] Notably, studies in social psychology show that individuals' attributional patterns and the extent to which individuals display the attributional tendencies described above depend on their 'implicit theories', i.e. their tacit assumptions about the social world.[64]

Beyond these two assumptions, the link between particular attributional patterns and signaling arguments in international relations, especially the costly signaling mechanism, remains undertheorized in previous work. For example, Deborah Larson argues that the fundamental attribution error compounds the problem of motives in international politics,[65] yet also maintains that costly signals can help states convey their motives credibly and overcome obstacles to cooperation.[66] These contradictory claims raise the question, what are the implications of attribution for (costly) signaling arguments?

I advance a domain-specific theory of attribution that addresses these gaps and sheds light on the problem of attribution in international politics, providing scholars with a framework for understanding how individuals infer states' motives and, in particular, why their inferences from costly signals vary. Integrating findings from research on the attribution of individual *and* group behavior with insights from rationalist scholarship on

[59] Hamilton and Sherman 1996; Hamilton, Sherman, and Castelli 2002; Susskind et al. 1999.
[60] Choi and Nisbett 1998; Masuda and Kitayama 2004; Miller 1984; Miyamoto and Kitayama 2002; Morris and Peng 1994; Norenzayan, Choi, and Nisbett 2002.
[61] Nisbett 2003.
[62] Markus and Kitayama 1991; Kuehnen et al. 2013.
[63] Molden and Dweck 2006, p. 200.
[64] Implicit theories are also known as 'lay theories' in psychology. For an overview of this research, see Levy, Chiu, and Hong 2006; Morris, Nisbett, and Peng 1996; Plaks, Levy, and Dweck 2009.
[65] Larson 1997a; Larson 1997b.
[66] Larson 1997a; Larson 1997b.

signaling, the theory argues that individuals' inferences from costly signals will depend on the interaction between an individuals' tacit assumptions about the target of attribution in international politics – their 'implicit theories' of states – and a signal's informational structure.

Specifically, building on psychological insights into information processing for individual and group targets, as well as the role of implicit theories in attribution, I propose a typology of implicit theories to explain variation in individuals' attributional patterns in international politics. The typology is derived from two underlying dimensions, 'entitativity' and 'mutability'. The first dimension, entitativity, captures the extent to which individuals perceive states as unified and coherent agents in international politics. The term 'entitativity' was coined by the psychologist Donald Campbell and refers to '[t]he degree of having the nature of an entity, of having real existence'.[67] While the basis for entitativity is an area of active research within psychology, scholars have shown that perceptions of entitativity can be both target-driven, varying depending on features of the group in question,[68] and recipient-driven, varying across recipients depending on their assumptions about the social world.[69] Scholars have argued that entitativity is central to explaining differences in attribution and impression formation for individual versus group targets,[70] showing that information processing for group targets is more similar to information processing for individual targets when the group in question is perceived as more entitative.[71] Notably, Vincent Yzerbyt and colleagues have found that individuals make more dispositional attributions and are less responsive to situational information, both for individual group members and the group as a whole, when they perceive a group as entitative.[72]

The second dimension, mutability, captures the extent to which individuals perceive agents in international politics – states or individuals – to be mutable. This dimension builds on research in social psychology investigating the role of implicit theories in attribution. Focusing on implicit theories of personality, Carol Dweck and colleagues

[67] D. T. Campbell 1958, p. 17.
[68] For example, Lickel, Hamilton, and Sherman 2001; Lickel, Hamilton, Uhles, et al. 2000; Hamilton, Sherman, and Rodgers 2004.
[69] Plaks, Levy, Dweck, and Stroessner 2004.
[70] Hamilton and Sherman 1996.
[71] McConnell, Sherman, and Hamilton 1997; Susskind et al. 1999; Welbourne 1999; Yzerbyt, Rogier, and Fiske 1998. For a review, also see Hamilton and Sherman 1996; Hamilton, Sherman, and Castelli 2002.
[72] Yzerbyt, Rogier, and Fiske 1998.

distinguish between two types of implicit theory: 'entity theory', which assumes an individual's traits to be fixed, and 'incremental theory', which assumes an individual's traits to be changeable.[73] They find that attributional processes vary depending on individuals' implicit theories: in attributing an individual's behavior, entity theorists make stronger dispositional inferences and are less responsive to situational factors than incremental theorists.[74] While research on implicit theories has predominantly focused on assumptions about individuals, some scholars have extended the concept to assumptions about groups.[75] Notably, Marilynn Brewer and colleagues have argued that information processing about groups depends, not only on perceptions of entitativity, but also on assumptions about the basis of a group's entitativitiy.[76] They distinguish between an essentialist view of groups, which assumes group entitativity to be grounded in common group traits, and an agentic view of groups, which assumes group entitativity to be founded on group goals (and values).[77] Whereas the former regards groups as fixed, the latter considers groups to be more malleable.[78]

The notion of mutability adopted here is both broader and more specific than Brewer et al.'s distinction. It is broader because it makes fewer assumptions about the substantive basis of mutability: a focus on common group values, for example, may imply a conception of groups as being more or less mutable, depending on an individual's concrete assumptions about those values and where they come from. At the same time, it is more specific because mutability is clearly distinguished from entitativity: an individual may hold an essentialist view of groups, while simultaneously '[...] responding to the group as if it were a single individual, but on a larger scale'.[79]

---

[73] Dweck, Hong, and Chiu 1993; Plaks, Levy, and Dweck 2009.

[74] Chiu, Hong, and Dweck 1997; Dweck, Hong, and Chiu 1993; Plaks, Levy, and Dweck 2009; Molden and Dweck 2006.

[75] Brewer, Hong, and Li 2004; Rydell et al. 2007.

[76] Brewer, Hong, and Li 2004.

[77] Brewer, Hong, and Li 2004, pp. 21-23. Also see Welbourne 1999.

[78] Brewer, Hong, and Li 2004.

[79] Brewer, Hong, and Li 2004, p. 23. According to Brewer, Hong, and Li 2004, '[...] essence and agency reflect two different meanings of the entitativity concept, one based on perceived common attributes, the other based on common goals and intentions' (Brewer, Hong, and Li 2004, p. 22). At the same time, they argue that only an agentic view of groups '[...] leads to responding to the group as if it were a single individual, but on a larger scale' (Brewer, Hong, and Li 2004, p. 23). By contrast, they maintain that '[...] essentialist theories of groupness have implications primarily for the perception [...] of group members as individuals and of the group prototype' (Brewer, Hong, and Li 2004, p. 23). This blurs the line between the concepts of entitativity and social category, raising ontological questions about the meaning of 'groupness' for essentialist theories of groups. In the proposed typology, I conceptualize mutability and entitativity as independent dimensions, and a more static view of groups does not itself imply a decrease in perceptions of the group's 'groupness'.

Building on this research, and the twofold distinction between an individual's assumptions about a state's entitativity and agents' mutability, I propose a typology of implicit theories that predicts variation in individuals' attributional patterns in international politics (see table 5.1). First, I predict that individuals who consider agents to be more mutable will be more responsive to situational information about the agent's behavior than individuals who consider agents to be less mutable. Second, I predict that individuals who consider a state to be less entitative will attribute dispositions more to individual decision makers than to the state as a whole. Integrating these predictions with rationalist arguments about signaling, the theory hypothesizes that individuals who assume states to be both entitative and mutable will update their beliefs about a state's motives in line with the costly signaling hypothesis. In contrast, individuals who assume states to be entitative, but less mutable, will be less responsive to the sender's strategic context, responding to costly behavior regardless of the context in which it was taken. The theory makes a similar prediction for individuals who assume states to be both less mutable and less entitative, however, hypothesizes that they will attribute costly behavior more to the dispositions of individual decision makers than to states as a whole. Finally, the theory hypothesizes that individuals who assume states to be less entitative, but agents to be mutable, are responsive to a sender's strategic context the way rationalist arguments predict, but that they ascribe dispositions more to individual decision makers than to states as a whole.

	Low mutability	High mutability
Low entitativity	• Dispositional attribution to leader(s) • Less responsive to strategic environment	• Dispositional attribution to leader(s) • More responsive to strategic environment
High entitativity	• Dispositional attribution to state • Less responsive to strategic environment	• Dispositional attribution to state • More responsive to strategic environment

Table 5.1: Implicit theories and attribution in international politics: theoretical predictions.

5.4 Research Design

As an initial test of the theory, this paper offers a plausibility probe. Plausibility probes 'are preliminary studies on relatively untested theories and hypotheses to determine

whether more intensive and laborious testing is warranted'.[80] They are particularly useful in developing novel psychological theories, which require the in-depth and fine-grained analysis of qualitative documents.[81] Moreover, plausibility probes may '[aim] at clearer specification of a theory and its variables'.[82] Although the variables of the domain-specific theory of attribution proposed here are clear, the theory's explanatory power hinges on the typology of implicit theories and the typology's applicability to real-world cases. A plausibility probe is useful as a first investigation and demonstration of how the abstract typology applies to real-world cases.

For the plausibility probe, I focus on elite decision makers' evolving beliefs about Soviet motives during the Carter Administration. The Carter Administration is especially well-suited to probing the theory for several reasons. First, elite decision makers within the administration varied on a key explanatory variable of the theory, implicit theories of the Soviet Union. The case thus lends itself to Chaim Kaufmann's competitive test strategy for investigating psychological theories in international relations qualitatively.[83] This strategy relies on 'patterns of variation among members of decision making groups' to establish the explanatory leverage of psychological theories vis-à-vis rationalist theories.[84] Specifically, by focusing on top-level decision makers within the same administration, I minimize variation on a key rationalist variable, namely information, as well as other, competing explanations of individual-level variation in decision makers' responses to costly signals, like partisanship, political ideology, or foreign-policy interests. I focus on three decision makers in particular: President Jimmy Carter, National Security Advisor Zbigniew Brzezinski, and Secretary of State Cyrus Vance.

Second, Jimmy Carter and his advisors entered into office at a pivotal moment in U.S.-Soviet relations: the stability of detente had already begun to wane, but optimism about the possibilities for increasing superpower cooperation were still running high, making the case a most-likely case for theories of belief updating in international relations.[85] Specifically, as U.S.-Soviet relations deteriorated and the structural pressures of superpower

[80] George and Bennett 2005, p. 75.
[81] Hall and Yarhi-Milo 2012, p. 564.
[82] George and Bennett 2005, p. 108.
[83] For studies employing this strategy, see Johnson and Tierney 2011; B. C. Rathbun 2019; Yarhi-Milo 2014.
[84] Kaufmann 1994, p. 561.
[85] Bennett and Elman 2008, p. 505.

rivalry intensified, rationalist theories lead us to expect belief convergence among elite decision makers. By contrast, given differences in decision makers' beliefs about Soviet motives from the outset, consistency theories lead us to expect belief divergence, even in response to costly signals of resolve. Finally, personalized theories lead us to expect that belief updating was uneven across decision makers, depending on their cognitive style and thus receptiveness to new information.

The analysis here focuses on three instances of competitive Soviet behavior that are essential to understanding the deterioration in U.S.-Soviet relations in the late 1970s in particular:[86] the Soviet intervention in the Ogaden War in 1977/1978, the Soviet stationing of a 'combat' brigade in Cuba, and the Soviet invasion of Afghanistan in late 1979. In all three cases, Soviet actions incurred a cost to the Soviet Union. However, each case differs with respect to the second explanatory variable of the theory, the salience of situational factors. Consequently, these cases lend themselves to testing the theory's key hypotheses and to probing its ability to explain both convergence *and* divergence in decision makers' beliefs about Soviet motives during the Carter Administration.

Finally, several psychological and beliefs-based studies have previously examined the Carter Administration and key decision makers' evolving beliefs of Soviet motives.[87] This facilitates the close comparison of the proposed theory's explanatory leverage with alternative psychological theories in international relations, notably consistency theories.[88]

The case studies test the theory in two ways. First, I trace the three decision makers' beliefs about Soviet motives over time, examining whether or not they updated their beliefs about Soviet motives, notably the extent of its expansionism, in response to Soviet behavior and how well the timing of belief change tracks the theory's predictions. Second, wherever possible, I analyze each decision maker's reasoning about the driving forces of Soviet behavior. This allows me to trace and validate the causal mechanism that underpins the theory's predictions: decision makers' attributional patterns. In particular, I examine the extent to which decision makers make dispositional and situational attri-

[86] Rosati 1988; Yarhi-Milo 2014.
[87] Notably, Y. S. Aronoff 2006; Lebow and J. G. Stein 2019; Rosati 1988; Yarhi-Milo 2014.
[88] Y. S. Aronoff 2006 provides an example of a personalized theory as applied to the case of President Carter. However, since she focuses primarily on explaining belief change about Soviet motives for Carter, not his advisers, it is more difficult to compare its explanatory leverage to the domain-specific theory of attribution I propose.

butions of Soviet behavior, respectively.[89] To do so, I rely on both verbal and behavioral indicators, and draw on secondary literature on the Carter Administration as well as primary documents, including policy documents, policy discussions, diary entries, memoirs, and interviews during and after the presidency.

5.5 Attribution and the End of Detente

5.5.1 Implicit Theories of the Soviet Union

Previous work on decision makers' beliefs about Soviet motives in the Carter Administration has emphasized decision makers' prior beliefs about Soviet motives – their images of the Soviet Union – as a key source of variation in the conclusions they drew from, and policies they proposed in response to, Soviet behavior. In particular, previous work has argued that, from the outset, Brzezinski's image of the Soviet Union was more hawkish than Vance's or Carter's images, respectively, and that Brzezinski was more responsive to hostile Soviet behavior as a result.[90]

	Low mutability	High mutability
Low entitativity	–	Jimmy Carter
High entitativity	Zbigniew Brzezinski	Cyrus Vance

Table 5.2: Implicit theories in the Carter Administration.

By contrast, this study focuses on variation in the three decision makers' implicit theories of the Soviet Union and resulting differences in their attributional processes. Specifically, I argue that Brzezinski and Vance both considered the Soviet Union to be a unified agent, however, each advisor held distinct views about the basis for that unity, giving rise to divergent assumptions about the Soviet Union's mutability: whereas Brzezinski considered the Soviet Union to relatively fixed, Vance considered it to be more malleable. Although President Carter focused more on individuals, particularly his counterpart in Moscow, Brezhnev, than the state as a unified agent, he shared with Vance the assumption that agents are fundamentally mutable.

[89] Note that dispositional and situational attributions are a matter of degree and not necessarily mutually exclusive.

[90] Rosati 1988, p. 486; Yarhi-Milo 2014, pp. 115-116. Although note that the secondary literature is divided on whether Brzezinski indeed had a (more) hawkish image of the Soviet Union from the outset, see S. J. Campbell 2003.

President Carter

Reflecting on the Soviet Union and the SALT negotiations in his memoirs, Carter observed that:

> '[...] it is much easier to negotiate a SALT treaty in a totalitarian society with a relatively consistent policy [...] the Politburo, in which military leaders play an important role, can make a decision, and the nation's propaganda apparatus can then proclaim the 'unanimously agreed upon' new policy essential to self-defense.'[91]

We might take this short excerpt to suggest that, thanks to its totalitarian political system, Carter regarded the Soviet Union as highly entitative. However, as scholars have noted elsewhere, Carter's lack of foreign policy experience prior to his presidency meant that his views of international politics, and the Soviet Union in particular, were less coherent and developed when he entered into office than either of his advisors' views were.[92] And, as Brzezinski later observed, 'Carter would be the first to admit that he came to the White House without a detailed plan for managing U.S.-Soviet relations'.[93] Therefore, in forming impressions of other countries in international politics, Carter relied, to a considerable extent, on his interactions with their leaders, some of whom he developed a strong personal relationship with.[94] This focus on foreign leaders qua individuals suggests an implicit theory of countries, including the Soviet Union, that is not very entitative.

Moreover, Carter's view of individuals was characterized by a deep personal belief in the possibility of transformation. This belief not only shaped his personal life, through his commitment to missionary work, for example,[95] but also shaped his approach to foreign policy: as Brzezinski recalls, Carter considered personal diplomacy to be an important vehicle for change in international politics.[96] For example, upon entering into office, Carter hoped to improve U.S.-Soviet relations by establishing a good rapport with the Soviet leader, whom he sent 'quite personal' letters, '[...] some of them [even written] in long-

[91] Carter 1982, p. 213.
[92] Y. S. Aronoff 2006, p. 432; Rosati 1988.
[93] Brzezinski 1983, p. 147.
[94] Yarhi-Milo 2014; Brzezinski 1983, p. 24.
[95] Y. S. Aronoff 2006.
[96] Brzezinski 1983, p. 151.

hand'.[97] The hopes and importance Carter attached to these efforts is apparent in the deep personal sense of disappointment he felt when they remained unreciprocated,[98] and suggests that, far from essentialist in his understanding of foreign leaders, he approached them openly and with genuine faith in the possibility of change through personal exchange.

Cyrus Vance

In contrast to Carter, Vance and Brzezinski had both acquired extensive foreign policy experience throughout their careers and, working as an academic, Brzezinski had published several influential studies on the Soviet Union.[99] Consequently, Vance's and Brzezinski's views of the Soviet Union were very well-developed when they entered into office.[100] And, while Brzezinski in particular noted the impact different leadership styles had had on Soviet politics, as well as the potentially destabilizing role domestic factions might play in the future,[101] both advisors regarded the Soviet ruling elite of the 1970s as a relatively entitative, i.e. unified and coherent, group.[102] Where they differed was in their implicit understanding of the basis for this unity.

Vance's understanding was predominantly interest-based: he tended to emphasize that Soviet foreign policy was driven, first and foremost, by national interest and warned against the 'preoccupation with ideology and confrontation' in U.S.-Soviet relations that came to characterize Reagan's presidential campaign.[103] Writing about the principles that guided his foreign policy approach toward the Soviet Union as secretary of state, he explains:

> 'I believed that we faced in the Soviet Union a powerful potential adversary with growing global interests and a compelling stake in avoiding military conflict with the United States. In my view, it was doubtful that there was a Soviet master plan for world domination, but rather an unceasing probing for advantage in furthering its national interests'.[104]

[97] Carter 1982, p. 223.
[98] For example, Carter 1982, p. 223; Brzezinski 1983, p. 155.
[99] S. J. Campbell 2003.
[100] Y. S. Aronoff 2006, p. 432.
[101] Brzezinski 1970; Carter 1982, p. 223.
[102] For example, Brzezinski 1970.
[103] Vance 1983, p. 21.
[104] Vance 1983, p. 28.

This focus on national interest suggests a view of the Soviet Union that emphasizes mutable goals rather than deep-seated values, attitudes, or unchangeable traits as the key factor underpinning Soviet group agency in international politics. While these goals might be informed by ideology, of course,[105] Vance did not consider one to entail the other. Consequently, although he did not share Carter's emphasis on *personal* diplomacy in particular, he did place great faith in the tools of diplomacy – the possibility of establishing common ground through sustained negotiation – to shape the Soviet leadership's goals and to increase superpower cooperation.[106]

Zbigniew Brzezinski

Compared to Vance, Brzezinski's understanding of the Soviet Union, and the Soviet ruling elite in particular, was more essentialist and less mutable: he emphasized ideological adherence and, to a lesser extent, formative political experiences as core commonalities of Soviet leaders and as key unifying themes underpinning Soviet foreign policy.[107] For example, in his book, *Between Two Ages*, which laid out his view of international politics and changes therein in the late 1960s and 1970s,[108] Brzezinski writes:

> '[...] the Soviet conception of the broad framework of contemporary reality, as articulated by top leaders [...] remains fundamentally dogmatic. The basic premise continues to be the Manichaean notion of the antagonistic dichotomy between the socialist and the capitalist worlds [...]. Though [...] the destructiveness of nuclear weapons dictates the necessity of peaceful coexistence...the underlying reality of our age is still said to be the competition between the two systems [...] this theme runs like a thread through all major speeches, foreign-policy analyses, or scholarly commentaries on world affairs [...] its importance lies in the influence of the ideological framework on more immediate, and otherwise quite well-informed, policy judgments.'[109]

Moreover, Brzezinski maintained that this ideology had, under the successive influences

[105] For example, Laucella 2004, p. 245 quotes Vance as saying 'I do believe [...] that they are going to take advantage of opportunities, wherever they may be found, which will advance their basic ideology or theology, whatever you want to call it'.
[106] Y. S. Aronoff 2006; Laucella 2004; Yarhi-Milo 2018.
[107] Y. S. Aronoff 2006.
[108] S. J. Campbell 2003; Vaisse 2018.
[109] Brzezinski 1970, p. 59.

of Stalinism and post-Stalinist leadership, taken on a very specific, context-dependent form: under Stalin, Soviet ideology had paired with an older, more conservative, Russian nationalism, thereby acquiring a degree of historical continuity with pre-revolutionary Russia;[110] under Stalin's successors, it had become increasingly divorced from society, dogmatic, bureaucratized and, consequently, static. In Brzezinski's own words:

> 'Soviet ideology is no longer shaped through the creative interaction of theoretical thought and practice – as was the case until Stalin's power became supreme – but is the product of a bureaucratic process of definition, a process entirely monopolized by career party officials. Ideology emanates from the offices of the Central Committee [...] before being submitted to the Politburo for approval [...]. The result is a condition of arrested ideological development, of ideological petrification rather than erosion [...]'[111]

This analysis suggests that, in addition to his emphasis on ideology, Brzezinski also considered Soviet ideology in the 1970s to be particularly immutable, rendering his view of the Soviet Union as an actor in international politics essentialist. It is therefore unsurprising that he considered change in the Soviet Union to hinge on '[...] changes in the internal character and outlook of the professional, ruling party bureaucracy.'[112] In his eyes, such changes were unlikely to occur before the early 1980s, when '[...] the first fully post-Stalin political leadership will enter the political arena.'[113]

5.5.2 Intervention in the Horn

Background

When Carter assumed office in 1977, several developments were already underway in the Horn of Africa. In Ethiopia, what began as a military revolt in early 1974 escalated into a revolution when teachers, students, and taxi drivers joined the protests. Against the background of wider challenges to his rule – notably, economic problems, a widespread famine, and a rebellion in Eritrea – the Emperor Haile Selassie was unable to appease the protesters and restore his power. By August 1974, a committee of army officers, the

[110] Brzezinski 1970, p. 56.
[111] Brzezinski 1970, p. 62.
[112] Brzezinski 1970, p. 62.
[113] Brzezinski 1970, p. 69.

'Dergue', had de facto taken over control of the government and, by mid-, deposed the Emperor. However, once in power, divisions within the Dergue – now the 'Provisional Military Administrative Council' (PMAC) – were quick to emerge. When tensions came to a head in late November 1974, a group of younger, more radical officers killed the PMAC's leader, Lt.-General Aman Andom. His successor, Mengistu Haile Mariam, was committed to the establishment of a socialist state and, within months, embarked on a radical land reform to rid Ethiopia of its feudal structures.[114]

Early on, Mengistu reached out to Moscow for aid. However, throughout 1974 and 1975, Moscow remained cautious toward the new Ethiopian regime and supported it only minimally. In part, this was due to the country's longstanding ties to the U.S.: though Mengistu's takeover and his radical policies severed relations with Washington, the Ford Administration continued to provide aid to the Ethiopian government. Regional considerations were also important: in 1974, Moscow and Somalia had signed the Treaty of Friendship and Co-operation, making Somalia Moscow's most important ally in the Horn. Somalia's claims to the Ogaden region in the East of Ethiopia were a source of conflict between the two neighboring states. Finally, caution about the PMAC's brand of socialism seems to have played a role, too: not only were Maoist tendencies apparent in some of Mengistu's initial reforms; the PRC also advised the new leadership directly.

It was not until early 1977 that loyalties in the Horn started to shift. In the fall and winter of 1976, the left-wing civilian opposition in Addis Ababa instigated an urban guerrilla against the Dergue. Mengistu responded with a violent campaign directed against the 'counter-revolutionaries'.[115] The massacre that ensued severed relations with the U.S., who found it increasingly difficult to justify the provision of aid to Ethiopia. By contrast, Mengistu's resolve in dealing with domestic opposition to the revolution impressed Moscow, whose confidence in the Ethiopian revolution and its leader subsequently grew. This paved the way for a secret arms deal between Moscow and the PMAC in late 1976.

Initially, the Soviet Union hoped that socialism could provide the basis for a 'Pax Sovietica' in the Horn of Africa, and for reconciliation between Ethiopia and Somalia. However, its hopes would soon be thwarted: at a meeting between Mengistu and the

[114] Patman 2009, pp. 150-203; Westad 2007, pp. 250-287.
[115] Patman 2009, p. 195.

Somalian President Siad Barre, facilitated by Fidel Castro in March 1977, President Siad resolutely rejected the Soviet proposal for a solution of the Ogaden conflict. Moscow subsequently stepped up its military and economic support to Ethiopia. Meanwhile, U.S.-Ethiopian relations had continued to deteriorate: in light of the PMAC's human rights abuses, the Carter Administration decided to reduce its military aid to Addis in February 1977.[116] Around the same time, President Siad started to seek closer contact with Washington. However, it was not until April 1977, when Mengistu broke ties with the U.S. and expelled all remaining American officials from Ethiopia, that the Carter Administration more seriously considered cooperating with Somalia.[117]

These considerations came to a sudden halt in July 1977, when Siad Barre decided to invade the Ogaden region. The Soviet response to this development was twofold. For one, Moscow further increased its support to Ethiopia. For another, it reduced aid to Somalia and tried to exercise what little influence it still had over the Somali President – to little avail: in mid-November, Siad Barre altogether terminated relations with Moscow. His decision coincided with an important development in Ethiopia: on November 11, Mengistu executed his most important rival within the PMAC, Vice Chairman Atnafu Abate. In conjunction, these two developments '[...] had the effect of removing any remaining restraint on the Soviet part', the historian Robert Patman contends.[118] Already in September, Soviet and Cuban military advisors had begun to arrive in Ethiopia. Now, the Soviet Union launched a massive air- and sealift to supply the Ethiopian army with military equipment, advisors and Cuban troops, who were brought over from Angola and Cuba.

Brzezinski later claimed that 'SALT lies buried in the sands of the Ogaden'.[119] Whether or not we agree with his conclusion, the Soviet intervention in the Horn certainly complicated efforts to improve U.S.-Soviet relations during the Carter Administration: not only did it undermine the Administration's efforts to build congressional support for SALT; it also exposed tensions within the executive that would damage the Administration's public image at home and generate confusion abroad. It therefore marks an important development in understanding the deterioration of U.S.-Soviet relations over the next two

[116] Woodroofe 2013.
[117] Woodroofe 2013.
[118] Patman 2009, p. 223.
[119] Brzezinski 1983, p. 189.

years.

Theoretical Predictions

The Soviet intervention in the Horn was mostly indirect: the Soviet Union supplied the Ethiopian army with military equipment and advisors, and facilitated the arrival of Cuban troops in Ethiopia, however, it did not send in troops of its own. Nonetheless, Moscow's massive air- and sealift carried a clear and substantial cost. At the same time, the Soviet intervention came in response to a set of offsetting strategic pressures that were not of Moscow's own making. First, Somalia invaded the Ogaden, thereby violating the OAU's principle on the inviolability of established borders.[120] Consequently, the international community widely considered Ethiopia to be waging a defensive war.[121] In other words, the Soviet Union was intervening on behalf of an ally under attack. Second, the Soviet Union intervened at the request of the Ethiopian government; it did not impose itself forcefully. Finally, Moscow only intervened after Somalia cut off ties to the Soviet Union and alternative courses of action to resolve the conflict became unviable. Situational factors, then, were clearly salient in the Soviet decision to intervene.

Given this salience of situational factors, the theory leads us to expect that Vance and Carter did not update their beliefs in response to Soviet actions in the Horn; that they strongly attributed the intervention to situational, alongside dispositional, factors; and that they considered the predictive value of Soviet behavior in the Horn to be low. By contrast, the theory leads us to expect that Brzezinski did update his beliefs about the Soviet Union; that he attributed the intervention primarily to dispositional factors; and that he considered the predictive value of Soviet behavior to be high.

Responses to the Intervention

Consistent with the theory's predictions, Vance's overall assessment of Soviet motives did not change in response to Soviet actions in the Horn, and he clearly attributed the intervention to situational, alongside dispositional, factors. In Vance's own words, the Soviet intervention presented a '[...] textbook case of Soviet exploitation of a local conflict'.[122]

[120] Patman 2009.
[121] Of course, this did not preclude the possibility that Ethiopia would cross into Somali territory once it had regained the Ogaden region. Indeed, this concern figured in security council meetings on the Horn.
[122] Vance 1983, p. 74.

Crucially, he did not dismiss or minimize the importance of Soviet actions in the Horn, as consistency theory might predict, but he did regard situational pressures as key to understanding why they had occurred. The following quote from his memoirs summarizes this position well:

> 'It was not that Soviet actions were unimportant, but I felt realism required us to deal with those problems in the local context in which they had their roots.'[123]

Commenting on an internal review of the Administration's Africa policy that was conducted in response to events in the Horn, Vance elaborates this position:

> 'This exhaustive governmentwide study led most participants to reject the 'grand design' interpretation of Soviet behavior. It concluded that Soviet policy in Africa could best be described as an exploitation of opportunity, an attempt to take advantage of African conflicts to increase Soviet influence. This view did not, of course, make Soviet behavior any more tolerable, but it did suggest that we should deal with each problem in its own context and not as a local battle in a global East-West geopolitical struggle.'[124]

Vance's behavioral predictions mirrored this emphasis on situational, alongside dispositional, factors: he did not consider the intervention to be indicative of broader trends in Soviet foreign policy nor did he predict strong cross-situational consistency in Soviet behavior. To the contrary, he assumed that situational factors could shape Soviet behavior in the future. His policy preferences in discussions on the Horn reflect this: on the one hand, Vance advocated policies to strengthen American political, economic and diplomatic ties in Africa, and the Horn in particular, and thereby help resolve conflicts that otherwise might be exploited by the Soviets;[125] on the other, Vance opposed proposals that more immediately and directly targeted Moscow or Soviet presence in the Horn, and risked severing Washington's ties with its adversary. Foremost among these were Brzezinski's proposal to send a U.S. carrier task force to the Red Sea in a show of force, the proposal to supply Somalia with arms, either directly or indirectly through Saudi Arabia, and the

[123] Vance 1983, p. 84.
[124] Vance 1983, p. 91.
[125] For example, Vance 1983, p. 91.

idea to link Soviet actions in the Horn to the SALT negotiations (the 'linkage issue'). As Vance declared at a security council meeting in early March, 'I do not think there is much leverage anyway on this issue'.[126] At best, he felt that these measures would '[...] underline our impotence on the Horn';[127] at worst, they would irrevocably undermine the Administration's efforts to increase cooperation with its adversary. Linking Soviet actions in the Horn to SALT, for example, '[...] will toughen the Russian's position,' Vance argued.[128] Similarly, when the idea of a joint statement of condemnation (of Soviet actions) with China was floated at the security council meeting, Vance objected:

> 'That would get their [Soviet] attention but we are at the point where we are on the brink of ending up with a real souring of relations between ourselves and the Soviet Union and it may take a helluva long while to change and may not be changed for years [...] we should examine it carefully before we go down that road.'[129]

Vance's reasoning presumes that U.S. foreign policy choices could shape, not only Soviet behavior, but Soviet motives, and that competitive policies were likely to change them for the worse. In other words, Vance's behavioral prediction mirrored his attribution of Soviet actions in the Horn.

By contrast, Brzezinski attributed Soviet actions in the Horn to underlying dispositions and, as a result, his overall assessment of Soviet motives changed for the worse. In a security council meeting in early March, for example, he stated:

> '[...] the Soviets are demonstrating a predisposition to exploit a local conflict for larger purposes. They are frightening more countries in the region and they are creating a precedent for more involvement elsewhere.'[130]

Clearly, Brzezinski regarded the intervention less as a response to a particular situation and more as part of a broader trend within Soviet foreign policy, driven by Soviet motives. In his own words, the Soviet Union was

[126] "Special Coordination Committee Meeting on the Horn of Africa, March 2, 1978"
[127] "Special Coordination Committee Meeting on the Horn of Africa, March 2, 1978"
[128] "Special Coordination Committee Meeting on the Horn of Africa, March 2, 1978"
[129] "Special Coordination Committee Meeting on the Horn of Africa, March 2, 1978"
[130] "Special Coordination Committee Meeting on the Horn of Africa, March 2, 1978"

> '[...] pursuing deliberately a policy on which it embarked some fifteen years ago: to structure a relationship of stability with the United States in those areas that are congenial or convenient to the USSR, while pursuing assertively every opportunity for the promotion of Soviet influence.'[131]

Moreover, in accordance with the theory, Brzezinski not only emphasized dispositional factors; he also dismissed the role of situational factors outright. In his memoirs, for example, he recalls:

> 'For Vance, the African matter was largely a local issue [...] I argued that the newly discovered Soviet-Cuban passion for the integrity of frontiers could hardly be analyzed in such narrow terms. Moreover, even if one allowed what seemed to me to be a preposterous notion, namely that the Soviets were acting out of some sort of strange territorial legalism, their presence so close to Saudi Arabia was bound to have strategic consequences [...] the most likely conclusion is that the Soviets at this time believed they could probably have their cake and eat it, too. They were thus pursuing a strategy of indirect expansionism [...]'.[132]

Consistent with this attribution of events, Brzezinski considered Soviet actions in the Horn to be predictive of Soviet behavior in the future. For example, responding to Vance's concerns about making a joint statement of condemnation with China, Brzezinski argued:

> 'If we allow the Soviets to send expeditionary forces to resolve territorial conflicts in ways that are beneficial to them, then we are going to have more and more problems [...] On the business of souring relations with the Soviets, the real question is why are they being soured? Do the Soviets want to sour these relations? If they can do what they want in the Horn without getting evidence of concern from us, we are going to have major problems with them in the south. We should communicate to the Soviets that they do not have a free hand and that what they do entails risks.'[133]

[131] Yarhi-Milo 2014, p. 307.
[132] Brzezinski 1983, pp. 186-187.
[133] "Special Coordination Committee Meeting on the Horn of Africa, March 2, 1978"

This line of reasoning presupposes that the U.S. had little to no agency in shaping Soviet motives. It could only seek to 'regulate' Soviet behavior by demonstrating resolve and increasing the cost of Soviet actions – now and in future. As a means of achieving both, Brzezinski advocated the deployment of a U.S. carrier task force to the Red Sea and favored indirect arms transfers to Somalia, among others.

President Carter's attribution of Soviet actions in the Horn is more difficult to establish than either Vance's or Brzezinski's: he did not participate in key NSC and SCC meetings, where Vance, Brzezinski, Harold Brown, and Walter Mondale, debated possible U.S. responses. Nor does Carter explicitly discuss the Soviet intervention in his memoirs or diary.[134] Prompted to speak about his views on Soviet actions in the Horn and the controversy it ignited among his advisors in an interview in 1987, Carter commented plainly: '[...] that was an altercation that I didn't particularly deplore because it involved conflicting and contrary opinions [within the Administration] with me being the final arbiter'.[135] However, Carter's beliefs about Soviet motives in 1977 and the first half of 1978, as well as the foreign policy decisions he made in his role as 'the final arbiter' of internal disagreements on the Horn, offer a window into his attribution of Soviet actions.

Specifically, whereas Carter's public speeches throughout early 1978 reflected the dual influences of Vance and Brzezinski, simultaneously emphasizing the importance of detente and areas of cooperation with the Soviet Union and critiquing Soviet actions in the Horn,[136] he ultimately decided against Brzezinski's proposal to deploy a carrier force to the Red Sea: '[i]n the end,' Brzezinski concedes, 'I did not carry the day. The President did not approve at this time the deployment of U.S. aircraft into the area [...]'.[137] And, although Carter raised Soviet actions in the Horn in diplomatic negotiations, he was not willing to stake progress on SALT on events in the Horn.[138] This reflected his broader optimism about cooperation with the Soviet Union and, particularly, the SALT negotiations. For example, in a diary entry from November 17, 1977, Carter wrote: '[w]e're making good progress with the Soviets. They've yielded on some major points, and we have been quite

[134] Unfortunately, due to library closures during the global pandemic in 2020 and 2021, I was unable to consult additional primary material at the Carter Presidential Library.
[135] "War and Peace in the Nuclear Age; Missile Experimental; Interview with Jimmy Carter, 1987" 1987.
[136] Glad 2009, pp. 79-84.
[137] Brzezinski 1983, p. 183.
[138] Yarhi-Milo 2014, pp. 134-135.

inflexible up till now'.[139] Similarly, on November 18, 1977, he noted that '[t]he letter from Brezhnev was positive [...] in the past there have always been some hookers in the Soviets' proposal [...] [l]ately, though, they've shown some more forthcoming attitudes'.[140] In line with the theory's predictions, these assessments, and Carter's policy decisions in early 1978, suggest that Carter did not make strong dispositional attributions of the Soviet intervention nor did he seem to regard the intervention as predictive of Soviet behavior in the future. To the contrary, his predictions of Soviet behavior and, particularly, Soviet willingness to cooperate with the U.S. within the framework of SALT remained distinctly positive.

5.5.3 Soviet Brigade Crisis

Background

The events that would culminate in the Soviet brigade crisis of August 1979 began in March, when Brzezinski and William Odom, one of his aides, requested an intelligence investigation into the Soviet military presence in Cuba.[141] Previously, the intelligence community had informed Brzezinski of the possible presence of Soviet ground forces in Cuba. However, their size, purpose, and conformity with agreements reached after the Cuban Missile Crisis in 1962 remained unclear.[142] In anticipation of the Senate hearings on SALT II, which were due to take place later that year, it made sense for Brzezinski to seek clarity: the deployment of Cuban troops in the Horn of Africa and the detection of Soviet-flown MIG-23 aircrafts in Cuba in 1978 had fuelled concerns about Cuban ties to the Soviet Union. Thus, Cuba would likely come into focus in debates about the treaty ratification.[143]

Responding to Brzezinski and Odom's inquiry, the intelligence community released a preliminary report on July 12, confirming the presence of Soviet ground forces as well as 'instances of small-scale Soviet tactical training that took place over a three-year period beginning in July 1976'.[144] It was not until August, however, that the issue snowballed

[139] Carter 2010, p. 137.
[140] Carter 2010, pp. 138-139.
[141] Caldwell 1991, p. 156.
[142] Glad 2009, p. 188.
[143] Strong 2000, p. 210.
[144] Caldwell 1991, p. 156.

into a political and diplomatic crisis. In early August, the CIA observed a field exercise undertaken by the brigade independently of any Cuban forces.[145] Shortly thereafter, the intelligence community '[...] characterized the unit in Cuba, for the first time, as a 'combat brigade",[146] thereby distinguishing it from the advisory missions that were known to be stationed in Cuba and served training purposes only. The CIA adopted this wording in its subsequent report, which was widely circulated within the government on August 22.[147] Within a few days, the contents had been leaked to a reporter for *Aviation Week and Space Technology*, and the issue quickly became one of public concern.[148]

Though the estimated size of the brigade was small, ranging between 2000 and 3000 men, its presence nonetheless alarmed the policy elite in Washington and the wider public: 'intelligence had never reported a Soviet combat unit in Cuba,' former CIA director Stansfield Turner recalls, 'people assumed that the brigade just arrived [...] it looked like a threatening move made by the Soviets'.[149] Indeed, throughout July and August the intelligence community maintained that the Soviet forces had been stationed in Cuba only a few years earlier, in the mid-1970s.[150]

By early September, however, the NSA and CIA started to revise this assessment, and to date both, the unit's presence as well as its characterization as a 'brigade', further back. On September 3, the NSA provided the President with information suggesting that the unit had been in Cuba at least since 1968.[151] An updated interagency intelligence report, published on September 18, conceded that 'the brigade or its precursors may date as far back as 1962'.[152] As it turned out, the unit had been stationed in Cuba since 1962 and the Kennedy Administration had been aware of its presence. Depending on the source one consults, the Kennedy Administration either acquiesced to, or permitted, its continued presence.[153] In either case, it was not the result of recent Soviet actions and, from early September at least, members of the executive were aware of this: in a dairy entry from September 4, 1979, Carter writes 'chances are they'd had approximately this level of

[145] Turner 1986.
[146] Glad 2009, p. 188.
[147] Glad 2009, p. 188.
[148] Glad 2009, p. 188; Strong 2000, p. 215.
[149] Turner 1986.
[150] Caldwell 1991, p. 156; Vance 1983, p. 360
[151] Glad 2009, p. 190.
[152] "Updated Report on Soviet Ground Forces Brigade in Cuba: Interagency Intelligence Memorandum, September 18, 1979" 1979.
[153] Turner 1986.

troops for the last 15 or 20 years'.[154] Similarly, in a briefing with Carter on September 5, Brzezinski noted that '[probably] it [the unit] has been in Cuba for a number of years.'[155] 'By late September,' Vance confirms, 'it was evident that the unit in question had almost certainly been in Cuba continuously since 1962'.[156]

The Soviet brigade crisis had far-reaching consequences for U.S.-Soviet relations: it led to a fateful delay in the Senate vote on the ratification of SALT II. In December, the Soviet Union invaded Afghanistan and any chance of the treaty being ratified evaporated.

Theoretical Predictions

The Soviet brigade crisis presents a rare test of the theory's predictions: because it ensued from an informational blunder of the U.S. intelligence community, we can observe both, decision makers' initial responses to the Soviet brigade's presence in Cuba when the salience of situational factors appeared to be low and decision makers' subsequent responses when the salience of situational factors turned out to be high. Initially, when the stationing of the brigade in Cuba appeared to be relatively recent, it could plausibly be construed as a costly signal: although the brigade itself was small and did not formally violate agreements reached between the U.S. and Soviet Union during the Cuban missile crisis,[157] its strategic and symbolic significance were substantial. Moreover, there were no immediate strategic threats to the Soviet Union that might have offset this cost. At the same time, the brigade did not pose an 'immediate strategic threat' to the U.S., making it relatively less costly than our third case, the Soviet intervention in Afghanistan.[158] By contrast, once it transpired that the brigade had been in Cuba since the 1960s, it altogether lost its informational value: clearly, the crisis resulted, not from decisions made it Moscow, but from an informational blunder in Washington.

The theory leads us to expect that, initially, when the salience of situational factors appeared to be low, all three decision makers were similarly responsive to the brigade's presence in Cuba; that they attributed Soviet 'actions' to Soviet dispositions; and that they considered them to hold predictive value. By contrast, the theory leads us to expect

[154] Carter 1982, p. 263.
[155] Brzezinski 1983, p. 348.
[156] Vance 1983, p. 362.
[157] The agreements prohibited the installation of offensive weapon systems, but were silent on the question of ground forces (Vance 1983, p. 359).
[158] Brzezinski 1983, p. 348.

that, from early September, when the intelligence community updated its conclusions on the brigade and it became apparent that situational factors drove the crisis, the three decision makers' responses diverged: whereas Vance and Carter revised their assessment of the brigade, correctly attributing the crisis to an informational blunder and assigning no predictive value to the brigade's presence, Brzezinski continued to attribute the crisis to Soviet dispositions and make behavioral predictions based on Soviet actions.

Responses to the Soviet Brigade Crisis

Consistency theory would predict that Vance dismissed or minimized information about the Soviet brigade in Cuba: it was inconsistent with Vance's optimism about the prospects for cooperation with the Soviet Union, but not costly or disruptive enough to induce a wholesale shift in Vance's assessment of Soviet motives. However, contrary to consistency theory, Vance did neither. Publicly, Vance responded to the preliminary intelligence report in July with caution: in his correspondence with Senator Stone, who had obtained information about the Soviet brigade from an unofficial source,[159] Vance emphasized the inconclusiveness of available data. There was 'no evidence of any substantial increase of the Soviet military presence in Cuba over the past several years,' he affirmed.[160] Yet, behind the scenes, Vance responded to the preliminary intelligence report with concern: 'on July 27 [...],' he recalls, 'I had Marshall Shulman meet with a senior Soviet embassy official to warn them that the presence of a combat unit in Cuba would inflame U.S.-Soviet relations at a critical moment in the Senate debate on the SALT II Treaty.'[161] In late August, when the intelligence community confirmed its July report, Vance's response was swift and assertive: on August 29, the State Department proposed to send a démarche to the Soviet Union. The proposal met with heavy resistance from the National Security staff.[162] Nonetheless, a telegram with instructions for a démarche was sent to the American embassy in Moscow later that day. The telegram insisted that:

> 'Our concern is not theoretical. We know that there are Soviet combat forces in Cuba. We regard this as a serious matter which cannot help but burden our

[159] "Letter From Senator Richard Stone to President Carter, July 24, 1979"
[160] Correspondence dated July 27, quoted in Vance 1983, p. 360. Vance was essentially repeating an earlier assessment made by Harold Brown before the Senate Foreign Relations Committee.
[161] Vance 1983, p. 360.
[162] "Summary of Conclusions of a Mini-Special Coordination Committee Meeting, August 29, 1979"; Brzezinski 1983, p. 347

> relations. The Secretary sees the meeting with the Foreign Minister in New York as an early opportunity to have comprehensive discussion on the trends in our relationship of which this deployment is one significant element.'[163]

While the content of the démarche had been drafted jointly by the State Department and National Security staff, the decision to send it to Moscow was Vance's. Brzezinski, who strongly opposed the measure, was still on vacation and not included in the final decision.[164] In addition to the démarche, Vance's Under Secretary, David Newsom, contacted the Soviet charge d'affairs to press Moscow for a clarification of the brigade's mission in Cuba.[165]

Importantly, these measures were taken before the issue became known to the wider public. In part, they certainly reflected Vance's concern that news of the brigade would thwart the Administration's efforts to get SALT II ratified by the Senate, and that any attempt to resolve the issue with Moscow diplomatically, through concessions, would be undermined once it became public.[166] At the same time, the very fact that Vance called on Moscow to clarify and resolve the issue and that, prior to any public escalation, the démarche explicitly linked the presence of the brigade to wider trends in U.S.-Soviet relations suggests a dispositional attribution of the brigade's 'discovery'. Vance's public statements are consistent with this conclusion, and underscore the seriousness of Vance's response. For example, on September 5, shortly after Senator Church brought the Soviet brigade to the public's attention, Vance held a press conference where he stated his position clearly: 'I will not be satisfied with the maintenance of the status quo,' he asserted.[167] This echoed his position, taken at a security council meeting on September 4, that 'we must have withdrawal'.[168]

However, in September, Vance's reading of the situation seemed to evolve. Although '[p]ublicly he assumed an extremely tough stance', Brzezinski contends that '[...] privately he moved gradually toward a softer position'.[169] By mid-September, Vance was evidently more concerned about resolving the issue publicly, and saving SALT, than about any

[163] "Telegram From the Department of State to the Embassy in the Soviet Union, August 29, 1979"
[164] Brzezinski 1983, p. 347.
[165] Glad 2009, p. 189.
[166] Vance 1983, p. 360.
[167] Quoted in Vance 1983, p. 362.
[168] Brzezinski 1983, p. 348.
[169] Brzezinski 1983, p. 349.

strategic implications of the brigade itself. The following excerpt from a meeting with the Soviet Ambassador Anatoly Dobrynin illustrates this:

> Vance: 'Anatoly, were those troops there during the Kennedy administration? During the Johnson administration? During the Nixon administration? During the Ford administration?'
>
> Dobrynin: 'Yes. Exactly.'
>
> Vance: 'Then what is this all about?'
>
> Dobrynin: 'I should ask you what it is all about.'
>
> Vance: 'Anatoly, can't you get them to move some ships around – to move some troops a little bit – so that we could say that it was now acceptable?'[170]

As this exchange suggests, Vance was responsive to information indicating that, far from being a signal, let alone a costly signal, the 'discovery' of the Soviet brigade and the crisis that ensued were the result of factors entirely external to Moscow. When the Soviet Union refused to withdraw the brigade or alter its composition, Vance supported several unilateral measures to reassure the American public and pave the way for the upcoming Senate hearings on SALT II.

By contrast, Brzezinski's interpretation of events remained strikingly consistent throughout the crisis. Like Vance, he responded to the preliminary reports in July with concern, and requested Stansfield Turner to step up the CIA's intelligence collection efforts. A diary entry from August 14 captures the growing sense of alarm with which he received subsequent updates on the Soviet unit:

> 'I briefed the President in the morning on the fact that according to intelligence reports there is now an actual Soviet brigade in Cuba, with headquarters and regular organization and that in fact it is scheduled to hold firing exercises

[170] Quoted in Glad 2009, p. 191.

> within a week. I told him that this is an extremely serious development which could most adversely affect SALT.'[171]

He attributed this 'extremely serious development' to Soviet dispositions, arguing that

> '[...] we [the Administration] should stress Cuban activism worldwide on behalf of Soviet interests as the main problem, that we should hint to the Soviets that we will intensify our relationship with the Chinese if the Soviets are not cooperative [...]'.[172]

Brzezinski was less concerned about the brigade itself, which posed no 'immediate strategic threat', than about what the brigade signaled – underlying Soviet dispositions – and the implications this had for U.S. foreign policy.[173]

It is worth noting that, even initially, when Vance and Brzezinski's both attributed the brigade's 'discovery' to Soviet action and dispositions, their policy preferences differed in ways that were consistent with their implicit theories: consistent with his belief in the mutability of Soviet dispositions, Vance emphasized diplomatic channels as primary means to resolve the crisis; consistent with his belief in the immutability of Soviet dispositions, Brzezinski had little faith in the possibility of Soviet willingness to cooperate on the issue, instead emphasizing the need to demonstrate resolve in order to deter the Soviet Union from future expansion.

These differences deepened once the intelligence community began to revise its initial assessment of the Soviet unit: unlike Vance, Brzezinski did not update his interpretation of the situation in response to new information indicating that the 'discovery' of the brigade was no discovery at all, but an informational blunder. For example, at a press meeting held on September 14, a reporter asked Brzezinski whether available data indicated that the brigade had been in Cuba since 1962. Brzezinski responded:

> 'No. I don't think the evidence now indicates that this unit has been there. I think there are some grounds, really firm grounds, for believing military personnel has been in Cuba throughout those years. Organized formation?

[171] Brzezinski 1983, p. 347.
[172] Brzezinski 1983, p. 348.
[173] Brzezinski 1983: 348

No [...]'[174]

Asked even more explicitly whether his views had been altered by intelligence that had surfaced since the issue became known to the public, Brzezinski affirmed his previous answer with a simple 'no'.[175] Consequently, he continued to attribute the brigade's 'discovery' to Soviet actions and dispositions:

> '[...] what has come to us as a new development and a disturbing one is the existence of a combat unit with a combat capability engaging in combat exercises. This raises the question as to what is its mission [...] especially given this somewhat more active role of Cuba [...] it is not a crisis, but it is a serious complication which if it isn't constructively resolved is likely to have some adverse impact on the U.S.-Soviet relationship'.[176]

Responding to a reporter's question why the U.S. could not simply accept the presence of the brigade, given that it did not constitute a threat, Brzezinski explained:

> 'What concerns us in this particular case is the fact that there is a combat unit by the Soviet Union in a country very close to us, stationed in circumstances unbeknownst to us, and perhaps even with a degree of deliberate camouflage [...]'.[177]

Rather than recognize the role that situational factors – an informational blunder – had played in precipitating the crisis, he maintained that 'the reported Soviet combat brigade in Cuba stems from a Soviet 'pattern of disregard' for American interests'.[178] Accordingly, while he did not '[...] want to make it the litmus test of the overall relationship', he did consider the ball to be in Moscow's corner: 'we would be escaping from reality not to recognize the fact that in some fashion the relationship will be affected by whether this issue is positively resolved or if it is not.'[179]

These statements were not mere public bravado. In late September and early October, Vance's and Carter's attempts to conclude the issue publicly led to considerable tension

[174] "Background Briefing, September 14, 1979." The statements were officially attributed to an 'Administration official'.
[175] "Background Briefing, September 14, 1979."
[176] "Background Briefing, September 14, 1979."
[177] "Background Briefing, September 14, 1979."
[178] Quoted in Gwertzman 1979.
[179] "Background Briefing, September 14, 1979."

between the President and his National Security Advisor. As Brzezinski recalls in his memoir:

> 'After the President delivered his speech [a televised address on the brigade issue, delivered on October 1], a small champagne party was held to celebrate the occasion and also his birthday. Neither Donovan nor I attended, for we both felt that his speech was a significant, but unfortunate, turning point in our political fortunes. Indeed, this was the only time that I ever thought seriously of the possibility of resigning. I felt so deeply about this that on October 4 I made what probably were the most disagreeable comments I ever made to the President [...] namely, that for the first time since World War II the United States told the Russians on several different occasions that we take great exception to what they are doing [...] I said that the way we handled this matter could be dangerous for the future because the Russians could miscalculate [...]. The President looked quite furious, and told me that he had no intention of going to war over the Soviet brigade in Cuba.'[180]

Clearly, Brzezinski considered it to be incumbent on the Administration to demonstrate resolve lest the Soviet Union would act similarly in future, persevering in what he considered to be its quest for global expansion.

Carter's initial response to the Soviet brigade issue is less well-documented than either Vance's or Brzezinski's. However, the available records suggest that his response to the issue followed a similar trajectory as Vance's. As Brzezinski recounts in his memoirs:

> 'On August 14, 'I briefed the President [...] on the fact that according to intelligence reports there is now an actual Soviet brigade in Cuba, with headquarters and regular organization and that in fact it is scheduled to hold firing exercises within a week. I told him that this is an extremely serious development which could most adversely affect SALT. The President looked quite concerned.'[181]

On September 7, 1979, Carter expressed his concern about the brigade publicly, declaring that 'we consider the presence of a Soviet combat brigade in Cuba to be a very serious

[180] Brzezinski 1983, p. 351.
[181] Brzezinski 1983, p. 347.

matter and that this status quo is not acceptable'.[182] As he noted in his diary on September 5, '[...] the Soviet troop presence in Cuba [...] is obviously not a threat to our country or a violation of any Soviet commitment, but politically devastating to SALT'.[183] However, like Vance, he was responsive to new information suggesting that the crisis had resulted from an informational blunder rather than any Soviet action. In his memoirs, he explained:

> 'Further investigations confirmed that the presence of these troops was not a new development; the brigade was only a remnant of a much larger force that had been there since the early 1960's. It represented no offensive threat to the United States, and the Soviet Union promised that would remain the case. [...] it was an embarassment [...] that for many years our government had not been aware of the brigade's existence.'[184]

In line with the theory's predictions, this excerpt shows that Carter acknowledged his government's part in generating the crisis rather than attributing it to Soviet dispositions. Accordingly, he did not pursue U.S. demands of withdrawal further, but instead accepted Soviet assurances that they would continue to use the brigade for training purposes only and announced several unilateral measures to allay public concerns, conclude the crisis, and pave the way for the ratification of SALT II.[185]

5.5.4 The Invasion of Afghanistan

Background

The Horn of Africa and Latin America were not the only areas that came into focus in the late 1970s. From 1979, developments in the Middle East drew much of the Administration's attention and resources. In January, the Iranian Revolution forced Shah Pahlavi, the U.S.' closest ally in the region, into exile. Amid the political turmoil that ensued, a group of students occupied the American embassy in Teheran and took its 52 American employees hostage. Whether the student group originally planned to hold the employees hostage for more than a few days is unclear. However, the Shah's ties to the U.S. had

[182] Carter 1982, pp. 263-264.
[183] Carter 2010, p. 354.
[184] Carter 1982, p. 264.
[185] Carter 2010, p. 355, Carter 1982, p. 264.

been an important rallying point for the protests, uniting leftist and religious forces within the country. Unsurprisingly therefore, the hostage-taking met with public approval and Ayatollah Khomeini, by then the leading figure of the revolution, officially endorsed it. It would take the Carter Administration 444 days to negotiate the release of the hostages.

The Iranian Revolution raised the stakes for both superpowers in the Middle East: for the U.S., the revolution resulted in the loss of its most important ally in the region; for the Soviet Union, Khomeini's rise to power and the strengthening of Islamic forces in the region were no less worrisome.

By the time the Shah fled Iran, Afghanistan, too, was engulfed in a political crisis. In April 1978, the People's Democratic Party of Afghanistan (PDPA), Afghanistan's Communist party, had staged a coup against President Daoud. The new government, led by Nur Mohammad Taraki, faced serious internal and external threats to its rule from the outset. Internally, factional infighting undermined the PDPA's effectiveness as a ruling party: the party was divided between the more radical Khalqis, who dominated the new government, and the more moderate Parchamis, who advocated alliances with older Afghan elites. Once in power, the conflict between the two factions quickly escalated and Babrak Karmal, the leader of the Parcham, went into exile in Czechoslovakia. In July, the government was purged of all Parchamis.[186]

Externally, a growing number of rebellions challenged the Khalqi regime. Resistance was particularly strong in the provinces, where the party lacked local support. Early on, Taraki and Hafizullah Amin, Taraki's influential rival within the Khalq, sought support from Moscow to bolster their position within the PDPA and to counter the rebellions. However, the Soviet Ambassador to Kabul, Aleksandr Puzanov, and leadership in Moscow remained sceptical of the Khalq's ability to conclude the revolution successfully. It was not until late 1978, when the Iranian Revolution picked up speed, that the Politburo agreed to provide Taraki with limited economic and military aid.[187] Only a few months later, in March 1979, the PDPA faced the most violent uprising yet: in Herat. Alarmed at these developments, the Politburo increased its commitment to Taraki further. At the same time, Moscow urged the the leader to forge broader alliances within the country and to improve relations with Afghanistan's neighbors.

[186] Westad 1997c.
[187] Westad 1997c, pp. 123-124.

It was in connection with the revolt in Herat that members of the Politburo considered the possibility of sending troops into Afghanistan for the first time.[188] As of yet, however, they excluded it as a policy option. It was not until October 1979 that an intervention became more likely. The preceding months had seen the political situation in Kabul deteriorate further: over the summer, the internal conflict between Taraki and Amin had escalated, culminating in the assassination of Taraki and takeover by Amin in early October. This development gave Moscow cause for concern: known to be a hardliner, Amin was uncompromising in his treatment of Parchamis or of anyone else who did not share his ultra-left views. The prospect of a more stable, Communist Afghanistan with broad-based support from the population thus became ever more elusive. Added to this were Soviet suspicions that Amin might have ties to the U.S., or even the CIA: Amin had studied at Columbia University, and tensions with Moscow led him to seek closer contact with the U.S.

The Soviet decision to intervene directly in the Afghan civil war must be understood against this background. Contrary to the 'grand design' interpretation of Soviet behavior, we know, with the benefit of hindsight, that the Soviet leadership only reluctantly agreed to the KGB's plan for invasion; and it did so only in early December, once Moscow seemed to be running out of alternatives courses of action.[189] On December 25, 1979, the Soviet Union started to airlift soldiers into Kabul. By December 27, the Afghan President Hafizullah Amin had been assassinated and replaced by Babrak Karmal, the leader of the Parcham. The decision to invade not only sealed the fate of detente, but also embroiled the Soviet Union in a costly war that would last for almost a decade and, ultimately, contribute to the collapse of the Soviet Union in 1991.[190]

Theoretical Predictions

Among the three instances of Soviet sunk-cost signaling, or alleged signaling, examined in this paper, the Soviet invasion of Afghanistan was the costliest: it was neither indirect nor small, nor were the situational factors that led to the decision as salient as in the Horn of Africa. To be sure, Amin's suspected ties to the U.S., the prospect of strengthening Islamic

[188] Westad 1997c, p. 132.
[189] Byrne and Savranskaya 2012.
[190] Westad 1997c, p. 118.

forces along Soviet borders, and deteriorating superpower relations all posed potential threats. Yet, in the months prior to the invasion, Moscow's relationship with Kabul had visibly deteriorated and, Soviet assurances to the contrary, its troops had not been sent in at President Amin's request (Amin was assassinated within hours of the invasion). Nor was Moscow responding to an external threat to Afghan security, as was the case in Ethiopia.

Given that situational pressures were less salient, and that the cost incurred by Soviet actions themselves was high, the theory leads us to expect that all three decision makers made dispositional attributions of Soviet actions in Afghanistan and considered the invasion to be predictive of Soviet behavior in the future. For Vance and Carter, who previously held less malign views of Soviet motives than Brzezinski, we would expect the invasion to result in significant belief change. However, despite this broad convergence, the theory also predicts differences in each decision makers' inferences. Notably, to the extent that situational factors were nonetheless present, we would expect Vance and Carter, but not Brzezinski, to have been perceptive to those factors (or the absence thereof). Moreover, the theory leads us to expect that Carter more strongly attributed the invasion to elite decision makers personally than either of his advisors.

Responses to the Invasion

Scholars typically characterize Vance's response to the Soviet invasion of Afghanistan as a textbook example of belief stability: unlike Carter, they contend, Vance did not update his beliefs about Soviet motives – even in response to this dramatic show of force.[191] This interpretation echoes a common theme in literature on the Carter Administration: divisions between Brzezinski and Vance.[192] At best, however, it provides a partial account of Vance's response to the invasion. To be sure, Vance's differed from Brzezinski's. Most notably, Vance was perceptive to situational factors in the Soviet decision to intervene:

> 'My view was that Moscow had acted as it did for a number of reasons. Its immediate aim was to protect Soviet political interests in Afghanistan which they saw endangered. Some Soviet leaders [...] feared that the regime would be replaced by a fundamentalist Islamic government and that this would, in turn,

[191] For example, Y. S. Aronoff 2006; Yarhi-Milo 2013, p. 25.
[192] For example, Glad 2009; Rosati 1988.

> be followed by a spread of 'Khomeini fever' to other nations along Russia's southern border [...] Moreover, the downward spiral in U.S.-Soviet relations had released the brakes on Soviet international behavior [...] it probably concluded that there was little reason to show restraint in dealing with a dangerous problem on its border [...] I am not suggesting that the Soviets would not have acted had relations with us been better. But it is possible that had there been more to lose in its relationship with the United States, the Soviet Union would have been more cautious. Finally, the Soviet leadership badly miscalculated [...] In sum, there were both local and global pressures pushing the Soviets in the same direction.'[193]

At the same time, and in line with the theory's predictions, Vance did consider the Soviet invasion to be a qualitative departure from past behavior: '[...] the question of why Moscow had elected to intervene massively with its own combat forces rather than increase its advisers and military assistance' was central to any U.S. response, Vance explains in his memoir.[194] Moreover, while he believed some Soviets to be motivated primarily by national security concerns, he recognized that '[o]ther Soviets believed that they should seize this opportunity to position themselves more favourably with respect to China and Pakistan'.[195] In his testimony before the Senate Foreign Relations Committee in March 1980, Vance argued:

> 'We confront a serious and sustained Soviet challenge, which is both military and political. Their military buildup continues unabated. The Soviet Union has shown a greater willingness to employ that power directly, and through others. In that sense, Afghanistan is a manifestation of a larger problem [...]'.[196]

Thus, although Vance did not share Brzezinski's view that Soviet action in Afghanistan was indicative of greater regional ambitions to expand, or part of a 'grand design', he clearly attributed the invasion to Soviet dispositions, and updated his beliefs about Moscow's motives and willingness to engage in military action accordingly.

[193] Vance 1983, p. 388.
[194] Vance 1983, pp. 387-388.
[195] Vance 1983, p. 388.
[196] Vance 1983, p. 396.

Vance's policy preferences are consistent with this conclusion: he strongly advocated punitive measures, like the grain embargo, to increase the costs of the invasion and signal Washington's condemnation of Soviet actions.[197] Indeed, in his memoir, Carter recalls that '[...] on these punitive measures, the State Department advocated stronger action than the National Security Council staff, a reversal of their usual attitudes'.[198]

At the same time, Vance sought to strike a balance between these punitive measures on the one hand, and an overall approach that signaled Washington's continued willingness to engage the Soviet Union constructively on the other. In particular, he agreed with his advisor, Marshall Shulman, that the basic framework of East-West relations should be kept in place. As Shulman wrote in a memo to Vance:

> 'The Soviets will not move in the direction of acceptance of greater restraint if they perceive that, despite our words, we are committed to a relationship of maximum confrontation on all fronts.'[199]

Moreover, to avoid miscalculation and to contain the conflict, Vance advocated the maintenance of high-level diplomatic contact with the Soviet Union throughout the crisis – a measure that found more support among European leaders than it did within the White House.[200] Clearly, Vance was perceptive to both, situational and dispositional factors.

Brzezinski attributed the invasion to underlying Soviet dispositions, too. However, unlike Vance, we find no mention of situational factors whatsoever. Unsurprisingly therefore, he took the invasion as further confirmation of his 'grand design' interpretation of Soviet motives and considered the invasion to be highly predictive of Soviet behavior in the future. In a memorandum for Carter written a day after the invasion, he noted: '[i]f the Soviets succeed in Afghanistan [...] the age-long dream of Moscow to have direct access to the Indian Ocean will have been fulfilled.'[201]

Consistent with this reading of events, Brzezinski supported, not only the grain embargo, but also the projection of U.S. power in the Middle East. In particular, he proposed

[197] Glad 2009, p. 200.
[198] Carter 1982, p. 476.
[199] "Letter from Marshall Shulman to Cyrus Vance on US-Soviet relations after Afghanistan, February 15, 1980"
[200] Glad 2009, p. 202.
[201] "Memorandum From the President's Assistant for National Security Affairs (Brzezinski) to President Carter, December 26, 1979"

a 'Carter Doctrine', which would provide assurances to Pakistan and the wider region.[202] He also advocated arms transfers to the Islamic insurgents in Afghanistan via Pakistan with the aim of ensuring that Afghanistan would become the Soviet Union's 'Vietnam', as well as the acceleration of efforts to improve and expand relations with China.[203] At the same time, in line with Brzezinski's conception of the Soviet Union as immutable, he opposed continued diplomatic contact with Moscow throughout the crisis, for it would '[...] send the wrong message to Moscow': a willingness to accommodate.[204]

Finally, Carter, like Vance and Brzezinski, attributed the invasion to Soviet dispositions. In a phone call with his Chief of Staff, Hamilton Jordan, shortly after the invasion, the President summarized his views on Soviet behavior in Afghanistan:

> '[...] this is deliberate aggression that calls into question detente and the way we have been doing business with the Soviets for the past decade. It raises grave questions about Soviet intentions [...]'[205]

Similarly, in a diary entry from December 28, the President described the invasion as '[...] a radical departure from the reticence the Soviets have shown since they overthrew the government in Czechoslovakia'.[206] Elsewhere, he even noted the absence of offsetting situational grounds for the Soviet course of action explicitly:

> '[...] the Soviets' actual invasion of a *previously nonaligned country, an independent, freedom-loving country, a deeply religious country*, with their own massive troops is a radical departure from the policy or actions that the Soviets have pursued since the Second World War'.[207]

In line with this attribution, Carter's beliefs about Soviet motives changed in response to the invasion: in his own words, 'the action of the Soviets had made a more dramatic change in my opinion of what the Soviets' ultimate goals are than anything they've done in the previous time that I've been in office'.[208]

[202] Glad 2009.
[203] "Memorandum From the President's Assistant for National Security Affairs (Brzezinski) to President Carter, December 26, 1979"
[204] Glad 2009, pp. 202-204; Brzezinski 1983, pp. 435-436.
[205] Quoted in Glad 2009, p. 197.
[206] Carter 2010, p. 382.
[207] Emphasis added. Quoted in Yarhi-Milo 2014, p. 140.
[208] The quote is from an ABC interview with Carter on Jan 19, 1980, cited in Y. S. Aronoff 2006, p. 444.

Mirroring this change, Carter's behavioral predictions were dispositional, too. In a diary entry from January 3, 1980, he anticipates that, '[...] unless the Soviets recognize that it [their action] has been counterproductive for them, we will face additional invasions or subversion in the future'.[209] In his memoirs, he expands on these concerns:

> '[...] the situation created a threat to both Iran and Pakistan which had not existed previously. If the Soviets could consolidate their hold on Afghanistan, the balance of power in the region would be drastically modified in their favor, and they might be tempted toward further aggression'.[210]

Clearly, in contrast to the Soviet intervention in the Ogaden War and the Soviet brigade crisis, Carter took the Soviet invasion of Afghanistan to be predictive of Soviet behavior in the future. At the same time, in making dispositional attributions of the invasion, Carter's response differed from Brzezinski's and Vance's in one important respect: as scholars have noted elsewhere, Carter was unusually angry and emotional and '[...] felt [a] sense of betrayal by Leonid Brezhnev'.[211] As he recorded in his diary, on December 28: 'I [...] sent on the Hot Line the sharpest message I have ever sent to Brezhnev, telling him that the invasion of Afghanistan would seriously and adversely affect the relationship between our two countries.'[212] In line with the theory's predictions, this suggests that Carter made more dispositional attributions to the Soviet leader personally than Vance or Brzezinski. This pattern of attribution matches psychologists finding that 'when situational information augmented the implied trait (e.g. anxious behavior in mundane situation), incremental theorists' trait attributions became more extreme than those of entity theorists'.[213]

The President's preferred policy responses to the invasion reflected his attribution of the invasion and his predictions of Soviet behavior in the future. Notably, Carter stood by the proposal to impose a grain embargo on the Soviet Union even though several members of his Administration, including Vice President Mondale, raised objections to the punitive measure.[214] Carter also increased the U.S.'s direct and indirect aid to the

[209] Carter 2010, p. 388.
[210] Carter 1982, p. 473.
[211] Y. S. Aronoff 2006, p. 435; Lebow and J. G. Stein 2019, p. 112; Yarhi-Milo 2014, p. 141.
[212] Carter 2010, p. 382.
[213] Plaks, Levy, and Dweck 2009, p. 1073.
[214] Objections to the grain embargo centered on two issues: the domestic repercussions of the grain embargo, which entailed a financial loss for farmers in the U.S., and the effectiveness of the embargo as a foreign policy tool to increase the cost of the invasion for the Soviet Union.

Islamist groups that were resisting the Khalq regime in Afghanistan's countryside. Finally, in what is sometimes referred to as the 'Carter Doctrine', the President publicly declared the Persian Gulf part of the U.S.'s areas of vital interest in his State of the Union Address in 1980.[215]

[215] "Address by President Carter on the State of the Union Before a Joint Session of Congress, January 23, 1980." It is worth noting that policy proposals to this effect were made by Brzezinski and his team of advisors throughout 1979 and predated the invasion (Brzezinski 1983, pp. 443-445).

Decision maker	Sunk-cost signal	Salience of situational factors	Attribution	Belief change
Brzezinski	Intervention in Ethiopia	High	Dispositional	Yes
	Brigade in Cuba I	Low	Dispositional	No
	Brigade in Cuba II	High	Dipositional	No
	Invasion of Afghanistan	Low	Dispositional	No
Vance	Intervention in Ethiopia	High	Situational	No
	Brigade in Cuba I	Low	Dispositional	No
	Brigade in Cuba II	High	Situational	No
	Invasion of Afghanistan	Low	Dispositional	Yes
Carter	Intervention in Ethiopia	High	Situational$\star$	No
	Brigade in Cuba I	Low	Dispositional$\star$	No
	Brigade in Cuba II	High	Situational	No
	Invasion of Afghanistan	Low	Dispositional	Yes

Table 5.3: Belief updating about Soviet motives during the Carter Administration: key findings. $\star$ tentative evidence only.

Please do not circulate or cite.

5.6 Conclusion

Recent empirical work on signaling and bargaining challenges a central assumption of theoretical scholarship in international relations: the assumption that actors rely on costly signals to infer states' motives in strategic interactions. Notably, it finds that the inferences individuals draw from costly signals systematically vary both across individuals and within individuals. Existing scholarship offers two approaches to explaining this variation: consistency theories and personalized theories. While these approaches shed light on the important role that prior beliefs, particularly prior beliefs about a state's motives, and cognitive processes play in individuals' perceptions of costly signals, neither provides an account of why individuals' inferences from costly signals vary both systematically *and* dynamically.

I advance a domain-specific theory of attribution to address this gap and provide international relations scholars with an explanation of why individuals' beliefs about a state's motives vary dynamically, sometimes converging and other times diverging in response to a state's costly signals. Integrating findings from attribution research with rationalist insights into the role of strategic interdependence in interstate communication, the theory argues that individuals' inferences from costly signals will depend on the interaction between their implicit theories of states and a signal's informational structure. Specifically, I theorize that an individual's implicit theory will shape that individual's attributional processes, making her more or less responsive to a sender's strategic context. Since a sender's strategic context is key to gauging the cost, and thus credibility, of its signals, individuals' responses to costly signals will vary as a result.

As a first test of the theory, I qualitatively examined three decision makers' evolving beliefs of Soviet motives during the Carter Administration. Each decision maker held a distinct implicit theory of the Soviet Union, allowing me to compare the inferences they drew from Soviet behavior while holding other variables, notably information, constant. I focused on three instances of Soviet signaling in particular: the Soviet intervention in the Ogaden War, the (supposed) stationing of a Soviet brigade in Cuba, and the Soviet invasion of Afghanistan. Exploiting variation in the Soviet Union's strategic context across all three cases, this allowed me to test the theory's ability to explain variation in the attribu-

tion of Soviet behavior, not only across decision makers, but also within decision makers, depending on a signal's informational structure. The findings support the domain-specific theory of attribution I advance, showing that decision makers' implicit theories of the Soviet Union, in interaction with a signal's properties, can explain individual differences in the timing and substance of belief updating about Soviet motives. In doing so, they explain important anomalies in previous work on belief updating and foreign policy debates within the Carter Administration. Notably, the theory developed in this paper can reconcile change in Carter's and Vance's beliefs about Soviet motives, and greater belief convergence within the Administration, in response to the Soviet invasion of Afghanistan with continued substantive differences in their interpretation of events and preferred policy responses. While the trajectory of responses to the Soviet brigade crisis among elite decision makers in the summer and fall of 1979 has received less attention in international relations, here, too, the theory can account for variation that would appear puzzling from the perspective of existing theories of belief updating and signaling.

Based on these findings, the domain-specific theory of attribution I advance clearly merits further investigation. Specifically, the case studies in this paper examined decision makers whose conceptions of the Soviet Union mapped onto three of the four types of implicit theory outlined in the domain-specific theory of attribution. Future research could conduct additional case studies to test the theory's predictions for the fourth type of implicit theory (i.e. low entitativity and low mutability). Moreover, although the case studies in this paper focus on costly signals of resolve, the theory I propose is general and should apply to signals regardless of their valence (i.e. whether they are competitive or cooperative). Future research could test this empirically by examining the inferences decision makers draw from costly signals of reassurance, like unilateral decisions to disarm.

The theory and findings of this study bear on several areas of research in international relations. First, the findings offer novel empirical insights into the role of attribution in signaling between states. In doing so, they advance longstanding debates within the discipline about the role of misperception in conflict escalation, suggesting that the risk of misperception due to attributional biases is not universal; it is actor- and context-specific. In particular, the findings suggest that the risk of misperception is greatest when actors who assume agents in international politics to be immutable receive signals that were

sent in a high-constraints environment. Awareness of these scope conditions is especially valuable in crisis situations when the risk of action-reaction chains and escalatory dynamics is high.

Second, the theory and findings on the role of implicit theories in the attribution of states' behavior shed light on the role of meaning in signaling. A growing body of empirical work on the microfoundations of signaling has emphasized the role of beliefs in information processing, arguing that beliefs have a self-reinforcing and, thus, polarizing effect on assessments of motives.[216] While important in its own right, this paper suggests that the focus on beliefs can obscure the deeper role meaning plays as a pre-requisite for belief formation, shaping, not only the outcome, but the very process by which recipients update their beliefs about a state's motives.

[216] For example, Kertzer, B. C. Rathbun, and N. S. Rathbun 2020; Yarhi-Milo 2014.

References

"Address by President Carter on the State of the Union Before a Joint Session of Congress, January 23, 1980" (2014). In: *Foreign Relations of the United States, 1977-1980, Volume I, Foundations of Foreign Policy*. Ed. by Kristin L. Ahlberg. United States Government Printing Office, pp. 691–700.

Aronoff, Yael (2014). *The Political Psychology of Israeli Prime Ministers: When Hard-Liners Opt for Peace*. Cambridge University Press.

Aronoff, Yael S. (2006). "In Like a Lamb, Out Like a Lion: The Political Conversion of Jimmy Carter." In: *Political Science Quarterly* 121 (3), pp. 425–449.

"Background Briefing, September 14, 1979" (1979). In: *Box 1, Folder: Cuba, Ray Jenkins' Subject File, White House Press Office, Carter Presidential Library, Atlanta*.

Bennett, Andrew and Colin Elman (2008). "Case Study Methods." In: *The Oxford Handbook of International Relations*. Ed. by Snidal Duncan and Christian Reus-Smit. Oxford University Press, pp. 499–517.

Boulding, Kenneth E. (1959). "National Images and International Systems." In: *Journal of Conflict Resolution* 3 (2), pp. 120–131.

Brewer, Marilynn B., Ying-yi Hong, and Qiong Li (2004). "Dynamic Entitativity: Perceiving Groups as Actors." In: *The Psychology of Group Perception*. Ed. by Vincent Yzerbyt, Charles M. Judd, and Olivier Corneille. Psychology Press, pp. 19–29.

Bronfenbrenner, Urie (1961). "The Mirror Image in Soviet-American Relations: A Social Psychologist's Report." In: *Journal of Social Issues* 17 (3), pp. 45–56.

Brutger, Ryan and Joshua D. Kertzer (2018). "A Dispositional Theory of Reputation Costs." In: *International Organization* 72 (3), pp. 693–724.

Brzezinski, Zbigniew (1970). *Between Two Ages: America's Role in the Technetronic Era*. Viking Press.

Brzezinski, Zbigniew (1983). *Power and Principle: Memoirs of the National Security Adviser, 1977-1981*. Weidenfeld & Nicolson.

Byrne, Malcolm and Svetlana Savranskaya (2012). *Soviet Invasion of Afghanistan Was Not a Grand Design But a Grand Entanglement Resulting from Faulty Intelligence, Excessive Secrecy, and a Paralyzed Leadership, According to Conference of Former Decision-Makers*. URL: https://nsarchive2.gwu.edu/NSAEBB/NSAEBB409/.

Cacioppo, John T. and Richard E. Petty (1982). "The Need for Cognition." In: *Journal of Personality and Social Psychology* 42 (1), pp. 116–131.

Caldwell, Dan (1991). *The Dynamics of Domestic Politics and Arms Control: The SALT II Treaty Ratification Debate*. University of South Carolina Press.

Campbell, Donald T. (1958). "Common Fate, Similarity, and Other Indices of the Status of Aggregates of Persons as Social Entities." In: *Behavioral Science* 3 (1), pp. 14–25.

Campbell, Steven Jay (2003). "Brzezinski's Image of the USSR: Inferring Foreign Policy Beliefs from Multiple Sources over Time." University of South Carolina.

Carter, Jimmy (1982). *Keeping Faith: Memoirs of a President*. Bantam Books.

Carter, Jimmy (2010). *White House Diary*. 1st ed. Farrar, Straus and Giroux.

Chittick, William O., Keith R. Billingsley, and Rick Travis (1995). "A Three-Dimensional Model of American Foreign Policy Beliefs." In: *International Studies Quarterly* 39 (3), pp. 313–331.

Chiu, Chi-yue, Ying-yi Hong, and Carol S. Dweck (1997). "Lay Dispositionism and Implicit Theories of Personality." In: *Journal of Personality and Social Psychology* 73 (1), pp. 19–30.

Choi, Incheol and Richard E. Nisbett (1998). "Situational Salience and Cultural Differences in the Correspondence Bias and Actor-Observer Bias." In: *Personality & Social Psychology Bulletin* 24 (9), pp. 949–960.

Dickson, Eric S. (2009). "Do Participants and Observers Assess Intentions Differently During Bargaining and Conflict?" In: *American Journal of Political Science* 53 (4), pp. 910–930.

Dweck, Carol S., Ying-yi Hong, and Chi-yue Chiu (1993). "Implicit Theories: Individual Differences in the Likelihood and Meaning of Dispositional Inference." In: *Personality & Social Psychology Bulletin* 19 (5), pp. 644–656.

Edward, Jones E. and Richard E. Nisbett (1971). *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior*. General Learning Press.

Fearon, James D. (1995). "Rationalist Explanations of War." In: *International Organization* 49 (3), pp. 379–414.

Fearon, James D. (1997). "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." In: *Journal of Conflict Resolution* 41 (1), pp. 68–90.

Festinger, Leon (1957). *A Theory of Cognitive Dissonance*. Row, Peterson.

Gartzke, Erik (1999). "War Is in the Error Term." In: *International Organization* 53 (3), pp. 567–587.

George, Alexander L. (1969). "The 'Operational Code': A Neglected Approach to the Study of Political Leaders and Decision-Making." In: *International Studies Quarterly* 13 (2), pp. 190–222.

George, Alexander L. and Andrew Bennett (2005). *Case Studies and Theory Development in the Social Sciences*. MIT Press.

Gilbert, Daniel T. and Patrick S. Malone (1995). "The Correspondence Bias." In: *Psychological Bulletin* 117 (1), pp. 21–38.

Glad, Betty (2009). *An Outsider in the White House: Jimmy Carter, His Advisors, and the Making of American Foreign Policy*. Cornell University Press.

Glaser, Charles L. (1995). "Realists as Optimists: Cooperation as Self-help." In: *International Security* 19 (3), pp. 50–90.

Glaser, Charles L. (2010). *Rational Theory of International Politics: The Logic of Competition and Cooperation*. Princeton University Press.

Gries, Peter Hays (2014). *The Politics of American Foreign Policy: How Ideology Divides Liberals and Conservatives Over Foreign Affairs*. Stanford University Press.

Gwertzman, Bernard (1979). "Brzezinski Cautions Soviet on Cuba Unit." In: *New York Times*.

Hall, Todd and Keren Yarhi-Milo (2012). "The Personal Touch: Leaders' Impressions, Costly Signaling, and Assessments of Sincerity in International Affairs." In: *International Studies Quarterly* 56 (3), pp. 560–573.

Hamilton, David L. and Steven J. Sherman (1996). "Perceiving Persons and Groups." In: *Psychological Review* 103 (2), pp. 336–355.

Hamilton, David L., Steven J. Sherman, and Luigi Castelli (2002). "A Group By Any Other Name–The Role of Entitativity in Group Perception." In: *European Review of Social Psychology* 12 (1), pp. 139–166.

Hamilton, David L., Steven J. Sherman, and Julie S. Rodgers (2004). "Perceiving the Groupness of Groups: Entitativity, Homogeneity, Essentialism, and Stereotypes." In: *The Psychology of Group Perception*. Ed. by Vincent Yzerbyt, Charles M. Judd, and Olivier Corneille. Psychology Press, pp. 30–46.

Heider, Fritz (1958). *The Psychology of Interpersonal Relations*. Wiley.

Herrmann, Richard K. et al. (1997). "Images in International Relations: An Experimental Test of Cognitive Schemata." In: *International Studies Quarterly* 41 (3), pp. 403–433.

Hewstone, Miles (1989). *Causal Attribution : From Cognitive Processes to Collective Beliefs*. Basil Blackwell.

Hewstone, Miles (1990). "The 'Ultimate Attribution Error'? A Review of the Literature on Intergroup Causal Attribution." In: *European Journal of Social Psychology* 20 (4), pp. 311–335.

Holsti, Ole (1970). "The "Operational Code" Approach to the Study of Political Leaders: John Foster Dulles' Philosophical and Instrumental Beliefs." In: *Canadian Journal of Political Science* 3 (1), pp. 123–157.

Holsti, Ole R. (1962). "The Belief System and National Images: A Case Study." In: *Journal of Conflict Resolution* 6 (3), pp. 244–252.

Holsti, Ole R. and James N. Rosenau (1986). "Consensus Lost. Consensus Regained?: Foreign Policy Beliefs of American Leaders, 1976-1980." In: *International Studies Quarterly* 30 (4), pp. 375–409.

Holsti, Ole R. and James N. Rosenau (1990). "The Structure of Foreign Policy Attitudes among American Leaders." In: *Journal of Politics* 52 (1), pp. 94–125.

Hurwitz, Jon and Mark Peffley (1990). "Public Images of the Soviet Union: The Impact on Foreign Policy Attitudes." In: *Journal of Politics* 52 (1), pp. 3–28.

Jervis, Robert (1976). *Perception and Misperception in International Politics*. Princeton University Press.

Johnson, Dominic D. P. (2020). *Strategic Instincts: The Adaptive Advantages of Cognitive Biases in International Politics*. Princeton University Press.

Johnson, Dominic D. P. and Dominic Tierney (2011). "The Rubicon Theory of War: How the Path to Conflict Reaches the Point of No Return." In: *International Security* 36 (1), pp. 7–40.

Jones, Edward E. and Victor A. Harris (1967). "The Attribution of Attitudes." In: *Journal of Experimental Social Psychology* 3 (1), pp. 1–24.

Kai, He and Feng Huiyun (2013). "Xi Jinping's Operational Code Beliefs and China's Foreign Policy." In: *Chinese Journal of International Politics* 6 (3), pp. 209–231.

Kaufmann, Chaim D. (1994). "Out of the Lab and into the Archives: A Method for Testing Psychological Explanations of Political Decision Making." In: *International Studies Quarterly* 38 (4), pp. 557–586.

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

Kertzer, Joshua D. (2016). *Resolve in International Politics*. Princeton University Press.

Kertzer, Joshua D., Brian C. Rathbun, and Nina Srinivasan Rathbun (2020). "The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations." In: *International Organization* 74 (1), pp. 95–118.

Kruglanski, Arie W. and Donna M. Webster (1996). "Motivated Closing of the Mind: "Seizing" and "Freezing"." In: *Psychological Review* 103 (2), pp. 263–283.

Kuehnen, Ulrich et al. (2013). "How Self-construal Affects Dispositionalism in Attributions." In: *Social Cognition* 31 (2), pp. 237–259.

Kydd, Andrew H. (2005). *Trust and Mistrust in International Relations*. Princeton University Press.

Larson, Deborah Welch (1994). "The Role of Belief Systems and Schemas in Foreign Policy Decision-Making." In: *Political Psychology* 15 (1), pp. 17–33.

Larson, Deborah Welch (1997a). *Anatomy of Mistrust : U.S.-Soviet Relations During the Cold War*. Cornell University Press.

Larson, Deborah Welch (1997b). "Trust and Missed Opportunities in International Relations." In: *Political Psychology* 18 (3), pp. 701–734.

Laucella, Melchiore J. (2004). "A Cognitive-Psychodynamic Perspective to Understanding Secretary of State Cyrus Vance's Worldview." In: *Presidential Studies Quarterly* 34 (2), pp. 227–271.

Lebow, Richard Ned and Janice Gross Stein (2019). "Afghanistan, Carter, and Foreign Policy Change: The Limits of Cognitive Models." In: *Diplomacy, Force, and Leadership: Essays in Honor of Alexander L. George*. Ed. by Dan Caldwell and Timothy J. McKeown. Westview.

Leites, Nathan (1951). *The Operational Code of the Politburo*. McGraw-Hill Book Company.

Levy, Sheri R., Chi-yue Chiu, and Ying-yi Hong (2006). "Lay Theories and Intergroup Relations." In: *Group Processes & Intergroup Relations* 9 (1), pp. 5–24.

Lickel, Brian, David L. Hamilton, and Steven J. Sherman (2001). "Elements of a Lay Theory of Groups: Types of Groups, Relational Styles, and the Perception of Group Entitativity." In: *Personality and Social Psychology Review* 5 (2), pp. 129–140.

Lickel, Brian, David L. Hamilton, A. Neville Uhles, et al. (2000). "Varieties of Groups and the Perception of Group Entitativity." In: *Journal of Personality and Social Psychology* 78 (2), pp. 223–246.

Markus, Hazel Rose and Shinobu Kitayama (1991). "Culture and the Self: Implications for Cognition, Emotion, and Motivation." In: *Psychological Review* 98 (2), p. 224.

Masuda, Takahiko and Shinobu Kitayama (2004). "Perceiver-Induced Constraint and Attitude Attribution in Japan and the US: A Case for the Cultural Dependence of the Correspondence Bias." In: *Journal of Experimental Social Psychology* 40 (3), pp. 409–416.

McConnell, Allen R., Steven J. Sherman, and David L. Hamilton (1997). "Target Entitativity: Implications for Information Processing about Individual and Group Targets." In: *Journal of Personality and Social Psychology* 72 (4), pp. 750–762.

Mearsheimer, John J. (2014). *The Tragedy of Great Power Politics*. Updated edition. W. W. Norton & Company.

Mercer, Jonathan (1996). *Reputation and International Politics*. Cornell University Press.

Miller, Joan G (1984). "Culture and the Development of Everyday Social Explanation." In: *Journal of Personality and Social Psychology* 46 (5), pp. 961–978.

Miyamoto, Yuri and Shinobu Kitayama (2002). "Cultural Variation in Correspondence Bias: The Critical Role of Attitude Diagnosticity of Socially Constrained Behavior." In: *Journal of Personality and Social Psychology* 83 (5), pp. 1239–1248.

Molden, Daniel C. and Carol S. Dweck (2006). "Finding "Meaning" in Psychology: A Lay Theories Approach to Self-Regulation, Social Perception, and Social Development." In: *American Psychologist* 61 (3), pp. 192–203.

Morris, Michael W., Richard E. Nisbett, and Kaiping Peng (1996). "Causal Attribution across Domains and Cultures." In: *Causal Cognition: A Multidisciplinary Debate*. Ed. by Dan Sperber, David Premack, and Ann James Premack. Oxford University Press.

Morris, Michael W. and Kaiping Peng (1994). "Culture and Cause: American and Chinese Attributions for Social and Physical Events." In: *Journal of Personality and Social Psychology* 67 (6), pp. 949–971.

Morrow, James D. (1999). "The Strategic Setting of Choices: Signaling, Commitment, and Negotiation in International Politics." In: *Strategic Choice and International Relations*. Ed. by David A. Lake and Robert Powell. Princeton University Press, pp. 77–114.

Nisbett, Richard E. (2003). *The Geography of Thought: How Asians and Westerners Think Differently - And Why*. Nicholas Brealey.

Norenzayan, Ara, Incheol Choi, and Richard E. Nisbett (2002). "Cultural Similarities and Differences in Social Inference: Evidence from Behavioral Predictions and Lay Theories of Behavior." In: *Personality & Social Psychology Bulletin* 28 (1), pp. 109–120.

Patman, Robert G. (2009). *The Soviet Union in the Horn of Africa: The Diplomacy of Intervention and Disengagement*. Cambridge University Press.

Pettigrew, Thomas F. (1979). "The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice." In: *Personality & Social Psychology Bulletin* 5 (4), pp. 461–476.

Plaks, Jason E., Sheri R. Levy, and Carol S. Dweck (2009). "Lay Theories of Personality: Cornerstones of Meaning in Social Cognition." In: *Social and Personality Psychology Compass* 3 (6), pp. 1069–1081.

Plaks, Jason E., Sheri R. Levy, Carol S. Dweck, and Steven J. Stroessner (2004). "In the Eye of the Beholder: Lay Theories and the Perception of Group Entitativity, Variability, and Essence." In: *The Psychology of Group Perception*. Ed. by Vincent Yzerbyt, Charles M. Judd, and Olivier Corneille. Psychology Press, pp. 96–109.

Quek, Kai (2021). "Four Costly Signaling Mechanisms." In: *American Political Science Review* 115 (2), pp. 537–549.

Rathbun, Brian C. (2019). *Reasoning of State: Realists, Romantics and Rationality in International Relations*. Cambridge University Press.

Rathbun, Brian C., Joshua D. Kertzer, and Mark Paradis (2017). "Homo Diplomaticus: Mixed-Method Evidence of Variation in Strategic Rationality." In: *International Organization* 71 (S1), S33–S60.

Renshon, Jonathan (2008). "Stability and Change in Belief Systems: The Operational Code of George W. Bush." In: *Journal of Conflict Resolution* 52 (6), pp. 820–849.

Rokeach, Milton (1960). *The Open and Closed Mind: Investigations into the Nature of Belief Systems and Personality Systems*. Basic Books.

Rosati, Jerel A. (1988). "Continuity and Change in the Foreign Policy Beliefs of Political Leaders: Addressing the Controversy over the Carter Administration." In: *Political Psychology* 9 (3), pp. 471–505.

Rosato, Sebastian (2015). "The Inscrutable Intentions of Great Powers." In: *International Security* 39 (3), pp. 48–88.

Rosato, Sebastian (2021). *Intentions in Great Power Politics: Uncertainty and the Roots of Conflict*. Yale University Press.

Rosenberg, Shawn W. and Gary Wolfsfeld (1977). "International Conflict and the Problem of Attribution." In: *Journal of Conflict Resolution* 21 (1), pp. 75–103.

Ross, Lee (1977). "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process." In: *Advances in Experimental Social Psychology* 10, p. 174.

Rydell, Robert J. et al. (2007). "Implicit Theories About Groups and Stereotyping." In: *Personality & Social Psychology Bulletin* 33 (4), pp. 549–558.

Sabini, John, Michael Siepmann, and Julia Stein (2001). "The Really Fundamental Attribution Error in Social Psychological Research." In: *Psychological Inquiry* 12 (1), pp. 1–15.

Schelling, Thomas C. (1960). *The Strategy of Conflict*. Harvard University Press.

Shimko, Keith L. (1991). *Images and Arms Control: Perceptions of the Soviet Union in the Reagan Administration*. University of Michigan Press.

Silverstein, Brett and Catherine Flamenbaum (1989). "Biases in the Perception and Cognition of the Actions of Enemies." In: *Journal of Social Issues* 45 (2), pp. 51–72.

Simon, Herbert A. (2017). "Rationality, Bounded." In: *The New Palgrave Dictionary of Economics*. Palgrave Macmillan UK, pp. 1–4. URL: https://doi.org/10.1057/978-1-349-95121-5_472-2.

Slantchev, Branislav L. (2005). "Military Coercion in Interstate Crises." In: *American Political Science Review* 99 (4), pp. 533–547.

Snidal, Duncan (1991). "Relative Gains and the Pattern of International Cooperation." In: *American Political Science Review* 85 (3), pp. 701–726.

Strong, Robert A. (2000). *Working in the World: Jimmy Carter and the Making of American Foreign Policy*. Louisiana State University Press.

Susskind, Joshua et al. (1999). "Perceiving Individuals and Groups: Expectancies, Dispositional Inferences, and Causal Attributions." In: *Journal of Personality and Social Psychology* 76 (2), pp. 181–191.

Taber, Charles S. and Milton Lodge (2006). "Motivated Skepticism in the Evaluation of Political Beliefs." In: *American Journal of Political Science* 50 (3), pp. 755–769.

Taylor, Donald M. and Vaishna Jaggi (1974). "Ethnocentrism and Causal Attribution in a South Indian Context." In: *Journal of Cross-Cultural Psychology* 5 (2), pp. 162–171.

"Summary of Conclusions of a Mini-Special Coordination Committee Meeting, August 29, 1979" (2013). In: *Foreign Relations of the United States, 1977-1980, Volume VI, Soviet Union*. Ed. by Jane Melissa Taylor. United States Government Printing Office, pp. 639–640.

"Letter From Senator Richard Stone to President Carter, July 24, 1979" (2013a). In: *Foreign Relations of the United States, 1977-1980, Volume VI, Soviet Union*. Ed. by Melissa Jane Taylor. United States Government Printing Office, pp. 630–631.

"Telegram From the Department of State to the Embassy in the Soviet Union, August 29, 1979" (2013b). In: *Foreign Relations of the United States, 1977-1980, Volume VI, Soviet Union*. Ed. by Melissa Jane Taylor. United States Government Printing Office, p. 641.

Tetlock, Philip (1999). "Theory-Driven Reasoning about Possible Pasts and Probable Futures in World Politics: Are We Prisoners of our Preconceptions?" In: *American Journal of Political Science* 43 (2), p. 335.

Tingley, Dustin H. and Barbara F. Walter (2011). "Can Cheap Talk Deter? An Experimental Analysis." In: *Journal of Conflict Resolution* 55 (6), pp. 996–1020.

Trachtenberg, Marc (2012). "Audience Costs: An Historical Analysis." In: *Security Studies* 21 (1), pp. 3–42.

Trager, Robert F. (2017). *Diplomacy: Communication and the Origins of International Order*. Cambridge University Press.

Turner, Stansfield (1986). "The Stupidity of 'Intelligence': The Real Story Behind That Soviet Combat Brigade in Cuba." In: *Washington Monthly*, p. 29.

"Updated Report on Soviet Ground Forces Brigade in Cuba: Interagency Intelligence Memorandum, September 18, 1979" (1979). In: *Container 34a, Cuba-USSR Brigade, Hamilton Jordan's Confidential Files, Office of the Chief of Staff Files, Carter Presidential Library, Atlanta*.

Vaisse, Justin (2018). *Zbigniew Brzezinski: America's Grand Strategist*. Harvard University Press.

Vance, Cyrus R. (1983). *Hard Choices: Critical Years in America's Foreign Policy*. Simon and Schuster.

Vertzberger, Yaacov (1990). *The World in Their Minds: Information Processing, Cognition and Perception in Foreign Policy Decisionmaking*. Stanford University Press.

Walker, Stephen G. (1977). "The Interface between Beliefs and Behavior: Henry Kissinger's Operational Code and the Vietnam War." In: *Journal of Conflict Resolution* 21 (1), pp. 129–168.

Walker, Stephen G. and Mark Schafer (2000). "The Political Universe of Lyndon B. Johnson and His Advisors: Diagnostic and Strategic Propensities in Their Operational Codes." In: *Political Psychology* 21 (3), pp. 529–543.

Waltz, Kenneth N. (1979). *Theory of International Politics*. McGraw-Hill.

"War and Peace in the Nuclear Age; Missile Experimental; Interview with Jimmy Carter, 1987" (1987). In: *GBH Archives*. WGBH Educational Foundation. URL: http://openvault.wgbh.org/catalog/V_332175FFE7084DEDA75E815B914A6192.

Weiner, Bernard (1985). "An Attributional Theory of Achievement Motivation and Emotion." In: *Psychological Review* 92 (4), pp. 548–573.

Weiner, Bernard (2008). "Reflections on the History of Attribution Theory and Research: People, Personalities, Publications, Problems." In: *Social Psychology* 39 (3), pp. 151–156.

Welbourne, Jennifer L. (1999). "The Impact of Perceived Entitivity on Inconsistency Resolution for Groups and Individuals." In: *Journal of Experimental Social Psychology* 35 (5), pp. 481–508.

"Letter from Marshall Shulman to Cyrus Vance on US-Soviet relations after Afghanistan, February 15, 1980" (1997a). In: *The Fall of Detente: Soviet-American Relations during the Carter Years*. Ed. by Odd Arne Westad. Scandinavian University Press, pp. 351–357.

"Special Coordination Committee Meeting on the Horn of Africa, March 2, 1978" (1997b). In: *The Fall of Detente: Soviet-American Relations during the Carter Years*. Ed. by Odd Arne Westad. Scandinavian University Press, pp. 260–272.

Westad, Odd Arne (1997c). "The Fall of Detente and the Turning Tides of History." In: *The Fall of Detente: Soviet-American Relations During the Carter Years*. Ed. by Odd Arne Westad. Scandinavian University Press, pp. 351–357.

Westad, Odd Arne (2007). *The Global Cold War: Third World Interventions and the Making of Our Times*. Cambridge University Press.

White, R. K. (1965). "Images in the Context of International Conflict: Soviet Perceptions of the U.S. and the U.S.S.R." In: *International Behavior: A Social Psychological Analysis*. Ed. by H. C. Kelman. Holt, Rinehart & Winston, pp. 238–276.

Wittkopf, Eugene R. (1990). *Faces of Internationalism: Public Opinion and American Foreign Policy*. Duke University Press.

Woodroofe, Louise (2013). *"Buried in the sands of the Ogaden": The United States, the Horn of Africa, and the Demise of Detente*. Kent State University Press.

Yarhi-Milo, Keren (2013). "In the Eye of the Beholder: How Leaders and Intelligence Communities Assess the Intentions of Adversaries." In: *International Security* 38 (1), pp. 7–51.

Yarhi-Milo, Keren (2014). *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations*. Princeton University Press.

Yarhi-Milo, Keren (2018). *Who Fights for Reputation: The Psychology of Leaders in International Conflict*. Princeton University Press.

Yarhi-Milo, Keren, Joshua D. Kertzer, and Jonathan Renshon (2018). "Tying Hands, Sinking Costs, and Leader Attributes." In: *Journal of Conflict Resolution* 62 (10), pp. 2150–2179.

Yoder, Brandon K. and Kyle Haynes (2021). "Signaling under the Security Dilemma: An Experimental Analysis." In: *Journal of Conflict Resolution* 65 (4), pp. 672–700.

Yzerbyt, Vincent Y., Anouk Rogier, and Susan T. Fiske (1998). "Group Entitativity and Social Attribution: On Translating Situational Constraints into Stereotypes." In: *Personality & Social Psychology Bulletin* 24 (10), pp. 1089–1103.

"Memorandum From the President's Assistant for National Security Affairs (Brzezinski) to President Carter, December 26, 1979" (2018). In: *Foreign Relations of the United States, 1977-1980, Volume XII, Afghanistan*. Ed. by David Zierler. United States Government Printing Office.

6 Conclusion

6.1 Summary of Arguments and Findings

States' ability to convey their motives credibly, and infer other states' motives accurately, can make the difference between bargaining and war, policies of mutual reassurance and arms races, a state of perpetual conflict and the pursuit of a shared vision for international society. Accordingly, motives, and the question of how actors assess them, are central to international relations theory. However, despite their centrality, empirical research on the communication of motives in international politics has long lagged behind theoretical scholarship, and fundamental questions about the microfoundations of interstate communication remain. In particular, a recent wave of experimental and psychological research on signaling and belief updating in international relations finds significant heterogeneity in how individuals respond to costly signals, challenging a core tenet of mainstream, rationalist theories of interstate communication.

In this thesis, I have sought to explain this heterogeneity and understand how individuals infer a state's motives from its behavior by investigating the attributional processes that underpin signaling between states. Integrating rationalist insights into the strategic nature of interstate interactions with attribution research in social and cultural psychology, I have argued that costly signaling depends on particular assumptions about individuals' attributional processes and have offered a novel experimental test of whether these assumptions are met for sunk-cost signals of resolve. The experimental findings provide partial support for the costly signaling hypothesis in international relations: based on responses from more than 800 adults in the U.S. and UK, respectively, they show that a state's security environment affects the inferences individuals draw from that state's actions. Specifically, participants attributed a state's sunk-cost signal of resolve less to expansionist motives, and more to defensive motives, when the state's environment was insecure than when it was secure. Importantly, this finding holds for third-party observers as well as individuals whose own state poses the strategic threat in question. These results are consistent with costly signaling and show that canonical findings from attribution research in psychology do not necessarily extend to international relations, increasing the risk of arms races as some scholars have suggested. Thereby, they underscore the need for

a domain-specific theory of attribution that provides scholars with a theoretical framework for understanding how individuals attribute the behavior of states, as distinct from the behavior of individuals.

However, the findings also challenge the costly signaling hypothesis. First, although a state's strategic environment moderates the effect that the cost of its actions has on participants' inferences about motives in the U.S. sample, the cost of a state's actions has no effect on participants' inferences in the UK sample. These findings mirror the results from previous studies, which provide mixed support for the effect of a signal's cost on perceptions of resolve and the credibility of threats.[1] Second, and more importantly for this thesis, the effect that a state's strategic environment has on participants' inferences about motives systematically varies depending on those participants' implicit theories of individuals and states, as well as their self-construal. In other words, the study finds systematic variation in attributional patterns, depending on individuals' assumptions about the constitution of social agents in international politics and the self in relation to its community.

A central argument I have made in this thesis is that psychological research in international relations faces severe limitations in providing an account of this heterogeneity because it tacitly adopts a conception of the mind as computer, relegating meaning to the role of individual preferences and beliefs that figure, alongside information, as a form of input into uniform mental processes. While this approach has yielded important insights into the way individual preferences and beliefs shape strategic interactions in international politics, the emphasis that its advocates place on providing a data-driven, causal account of human perception and decision making has obscured the way in which this account is shaped, and limited, by its underlying conception of the mind. In particular, I discuss how the conception of the mind as computer sidelines the deeper role meaning plays in accounts of cognition: although the computational processes that constitute cognition are thought to operate on mental states in virtue of their syntactic properties, in philosophy and psychology today, mental states are widely assumed to have representational content, i.e. to refer to objects or states of affairs in the world. Representational content in turn forms the basis for the propositional content of preferences and beliefs and, thus,

[1] Quek 2016; Quek 2021.

is essential to processes of belief formation and updating. However, psychology, and the computational conception of the mind, presumes representational content; it does not provide an account of it. Consequently, international relations scholars must integrate psychological research on information processing with constructivist scholarship on social facts and socially-constructed meaning in order to develop a microfoundational account of belief formation and updating in international politics.

While the experimental findings suggest the need for a domain-specific theory of attribution, this theoretical argument about the limits of the conception of the mind as a computer gives us a starting point for building such a theory: individuals' socially-constructed representations of the target of attribution in international politics. I draw on the notion of 'implicit theories' in psychology to conceptualize these representations and theorize their impact on belief updating about states' motives. Building on attribution research, as well as research on implicit theories and group perception in psychology, I propose a typology of 'implicit theories' and offer predictions for how each implicit theory will shape an individual's attributional processes in international interactions. Crucially, I argue that individuals' attributions, and their belief updating about a state's motives, will depend on the interaction between an individual's implicit theory of a state and the informational structure of the state's signal, notably the salience of situational factors in the state's actions.

As an initial test of the theory, I examined three U.S. elite decision makers' attributions of Soviet behavior during the Carter Administration. I focused on three sunk-cost signals of resolve that differed in the extent to which situational factors were salient: the Soviet intervention in the Ogaden War in 1977, the (supposed) stationing of a Soviet brigade in Cuba in the late 1970s, and the Soviet invasion of Afghanistan in December 1979. The findings support the theory: implicit theories and differences in attributional processes can explain why the three decision makers' varied in the extent to which they attributed a signal to Soviet dispositions, the extent to which they made dispositional attributions about individual state leaders, and their timing of belief change about Soviet motives. More broadly, the findings demonstrate the utility of a theory of attribution for explaining why individuals' beliefs about a state's motives dynamically vary, sometimes diverging and other times converging in response to costly signals.

In developing a domain-specific theory of attribution, this thesis builds on insights from existing theories of signaling and perception in international relations. Specifically, it builds on the rationalist insight that uncertainty about motives is essential to understanding when states succeed or fail at finding mutually acceptable agreements, and that costly signals provide states with an important mechanism to reduce this uncertainty. It also builds on psychological insights into the role of perception in producing international outcomes and, particularly, the role of attributional processes in hampering, or facilitating, communication about motives. Finally, it builds on the constructivist insight that socially-constructed meaning is a pre-requisite for understanding how actors process cues in social interactions.

At the same time, the theory makes several contributions to international relations theory. For rationalist scholars, the arguments and findings presented in this thesis offer novel insights into the microfoundations of costly signaling and identify two new sources of recipient-driven variation in individuals' responses to costly signals – implicit theories and self-construal. For scholars working on psychological approaches, this thesis systematically integrates attribution research in psychology with costly signaling arguments, and develops a domain-specific theory of attribution that places the target of attribution in international politics – states and individual state actors – front and center. In doing so, it addresses fundamental gaps in psychological theories of signaling and, in particular, theories that import findings from attribution research in psychology into international relations. For constructivist scholars, the arguments and findings of this thesis elaborate the role of socially-constructed meaning in information processing, bringing rationalist and psychological work on belief updating into dialogue with interpretivist efforts to understand the emergence of social agents and complex systems of meaning.

Beyond these contributions to substantive and theoretical debates within the discipline, the arguments I advance also bear on the methods international relations scholars use to study information processing and the microfoundations of signaling. While a recent wave of psychological research within the discipline has mainstreamed the use of experiments to study signaling and it has become commonplace to combine experimental and qualitative methods, the need to incorporate representational content into the study of belief updating puts a new twist on this disciplinary trend: qualitative analyses not only

facilitate the study of information processing among elite decision makers in real-world settings, allowing researchers to gauge the importance of psychological mechanisms in producing foreign policy outcomes; they are also necessary in order to develop a descriptive basis for theorizing variation in representational content and, thus, the information processing mechanisms that experiments investigate.

6.2 Avenues for Future Research

There are several ways in which future research could build on, and elaborate, the domain-specific theory of attribution I have developed and, thereby, advance our understanding of the mechanisms that underlie conflict and cooperation between states. Broadly, future research could investigate the role of implicit theories in individuals' attributional processes further, extend the theoretical framework of this thesis to types of signals not examined here, and leverage the theory of attribution to explore when individuals ascribe behavior at one level of analysis in international relations to motives at another.

6.2.1 The Causal Role of Implicit Theories

A central claim of the domain-specific theory of attribution developed here is that implicit theories explain heterogeneity in individuals' attributional patterns, providing scholars with an account of why individuals' beliefs about a state's motives dynamically vary, sometimes diverging and other times converging in response to that state's signals. In support of the theory, I have presented findings from a survey experiment and a qualitative analysis of three elite decision makers' evolving perceptions of Soviet motives. However, the survey experiment observes rather than randomizes individuals' implicit theories. Consequently, its findings on the link between implicit theories and attribution are correlational, not causal. Building on the qualitative analysis presented here, future research could develop implicit theory primes and employ a treatment-by-treatment design to test the causal role of implicit theories in the attribution of states' behavior experimentally, thereby supplementing the qualitative findings of this thesis.

Moreover, the qualitative analysis presented in this thesis included case studies of three out of four types of implicit theory outlined in the domain-specific theory of attribution. Future research could extend this analysis by identifying, and analyzing the attributions

of, decision makers whose assumptions about actors in international politics map onto the fourth type (i.e. low entitativity and low mutability), thereby offering a more complete test of the theory I advance in this thesis.

6.2.2 Other Types of Signals

This thesis has focused on sunk-cost signals of resolve. However, other types of signals are equally important in developing a domain-specific theory of attribution in international politics, including tying-hands signals, costless signals, and signals of reassurance.[2] Unlike sunk-cost signals, tying-hands signals do not incur a cost to the sender upfront; they incur a cost conditional on the recipient's subsequent behavior – if the recipient subsequently challenges the sender, they increase the cost of backing down relative to defending. In other words, these signals are credible because they narrow the sender's strategic options, figuratively tying the sender's hands.[3] The classic example are verbal threats made in public which incur a reputational cost to the sender if the sender is challenged but fails to follow through on the threat.[4] Recent research finds that individuals respond differentially to sunk-cost signals and tying-hands signals, depending on whether they are foreign policy 'hawks' or 'doves'.[5] Since the cost, and thus credibility, of tying-hands signals does not reside in the sender's actions *per se*, but in changes to the sender's strategic calculus in light of its environment, the differences in individuals' attributional patterns examined in this thesis may provide an alternative explanation of why individuals respond differentially to sunk-cost and tying-hands signals.

Future research could also extend the theory of attribution to costly signals of reassurance, like unilateral decisions to disarm. Signals of reassurance are essential to understanding how states reduce uncertainty about motives, avoid or overcome security dilemmas, and form trust.[6] However, a number of studies in psychology and international relations suggest that attributional processes work against costly signals of reassurance: they find that individuals respond differentially to positive and negative behavior by outgroup members, attributing negative behavior more to underlying motives than positive behavior.[7]

[2] For more types of signals, see Quek 2021.
[3] Fearon 1997; Quek 2021.
[4] Fearon 1997.
[5] Yarhi-Milo, Kertzer, and Renshon 2018.
[6] Glaser 2010; Kydd 2005.
[7] For example Hewstone 1990; Mercer 1996; Rosenberg and Wolfsfeld 1977; Silverstein and Flamenbaum

In contrast, the theoretical framework developed in this thesis should apply to signals regardless of their valence, cross-cutting these predictions. By investigating empirically whether the theoretical framework does, in fact, extend to costly signals of reassurance, future research could shed light on how these distinct mechanisms intersect and when attributional processes facilitate, rather than hinder, efforts to transform ingrained enemy images and increase cooperation between states.

6.2.3 Attribution and the Levels of Analysis Problem

Though scholarship on signaling has long considered costly signals to be credible indicators of states' unobservable motives, recent theoretical and empirical scholarship on signaling shows that costless signals can be credible, too – particularly in interpersonal interactions.[8] This thesis has focused on how individuals infer states' motives from state behavior. However, the theoretical framework it develops also lends itself to investigating how recipients infer a state's motives from the behavior of individual state actors, like leaders or diplomats, and to explaining when individuals are more or less likely to attribute the behavior of individual state actors to the state as a whole. In doing so, it could help address important gaps in the burgeoning literatures on trust-building and reputation formation in international relations, notably the question of when recipients form country-specific reputations based on leader-specific actions,[9] and explain how mechanisms at one level of analysis in international relations shape outcomes at another.

⋆

At a time when international events have brought the possibility of great power war into focus again, the stakes of inferring states' motives accurately could not be higher. However, empirical scholarship on signaling has found that actors vary in the inferences they draw from states' behavior, with significant implications for the risk of costly military conflicts. This thesis has sought to understand this variation by developing a theory of attribution that recognizes a key insight from psychology: the tacit assumptions individuals hold about the targets of attribution fundamentally shape the way they infer motives

1989.

8 Hall and Yarhi-Milo 2012; Kurizaki 2007; Tingley and Walter 2011; Trager 2017; Wheeler 2018; Yarhi-Milo 2014.

9 Solimena 2022.

from behavior. Individuals' notions of individual and group agency in international politics thus emerge as a potent area of research that bears as much on foundational questions of cognition and the mind as on foundational questions of international relations. This thesis is but the first step toward addressing these questions.

Bibliography

Aronoff, Yael (2014). *The Political Psychology of Israeli Prime Ministers: When Hard-Liners Opt for Peace*. Cambridge University Press.

Bruni, Luigino and Robert Sugden (2007). "The Road Not Taken: How Psychology Was Removed From Economics, and How It Might Be Brought Back." In: *The Economic Journal* 117 (516), pp. 146–173.

Brutger, Ryan et al. (2022). "Abstraction and Detail in Experimental Design." In: *American Journal of Political Science*, pp. 1–16.

Bullock, John G. (2009). "Partisan Bias and the Bayesian Ideal in the Study of Public Opinion." In: *Journal of Politics* 71 (3), pp. 1109–1124.

Campbell, Donald T. (1958). "Common Fate, Similarity, and Other Indices of the Status of Aggregates of Persons as Social Entities." In: *Behavioral Science* 3 (1), pp. 14–25.

Choi, Incheol and Richard E. Nisbett (1998). "Situational Salience and Cultural Differences in the Correspondence Bias and Actor-Observer Bias." In: *Personality & Social Psychology Bulletin* 24 (9), pp. 949–960.

Crawford, Neta C. (2000). "The Passion of World Politics: Propositions on Emotion and Emotional Relationships." In: *International Security* 24 (4), pp. 116–156.

Dickson, Eric S. (2009). "Do Participants and Observers Assess Intentions Differently During Bargaining and Conflict?" In: *American Journal of Political Science* 53 (4), pp. 910–930.

Dixit, Avinash K and Jörgen W Weibull (2007). "Political Polarization." In: *Proceedings of the National Academy of Sciences (PNAS)* 104 (18), pp. 7351–7356.

Dweck, Carol S., Ying-yi Hong, and Chi-yue Chiu (1993). "Implicit Theories: Individual Differences in the Likelihood and Meaning of Dispositional Inference." In: *Personality & Social Psychology Bulletin* 19 (5), pp. 644–656.

Edelstein, David M. (2002). "Managing Uncertainty: Beliefs about Intentions and the Rise of Great Powers." In: *Security Studies* 12 (1), pp. 1–40.

Edward, Jones E. and Richard E. Nisbett (1971). *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior*. General Learning Press.

Fearon, James D. (1995). "Rationalist Explanations of War." In: *International Organization* 49 (3), pp. 379–414.

Fearon, James D. (1997). "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." In: *Journal of Conflict Resolution* 41 (1), pp. 68–90.

Festinger, Leon (1957). *A Theory of Cognitive Dissonance*. Row, Peterson.

Fey, Mark and Kristopher W. Ramsay (2006). "The Common Priors Assumption: A Comment on "Bargaining and the Nature of War"." In: *Journal of Conflict Resolution* 50 (4), pp. 607–613.

Frieden, Jeffry A. (1999). "Actors and Preferences in International Relations." In: *Strategic Choice and International Relations*. Ed. by David A. Lake and Robert Powell. Princeton University Press, pp. 39–76.

Fuhrmann, Matthew and Todd S. Sechser (2014). "Signaling Alliance Commitments: Hand-Tying and Sunk Costs in Extended Nuclear Deterrence." In: *American Journal of Political Science* 58 (4), pp. 919–935.

Gerber, Alan S. and Donald P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. W.W. Norton & Company.

Gilbert, Daniel T. and Patrick S. Malone (1995). "The Correspondence Bias." In: *Psychological Bulletin* 117 (1), pp. 21–38.

Glaser, Charles L. (2010). *Rational Theory of International Politics: The Logic of Competition and Cooperation*. Princeton University Press.

Glaser, Charles L. et al. (2016). "Correspondence: Can Great Powers Discern Intentions?" In: *International Security* 40 (3), pp. 197–215.
Goldgeier, J. M. and P. E. Tetlock (2001). "Psychology and International Relations Theory." In: *Annual Review of Political Science* 4 (1), pp. 67–92.
Hafner-Burton, Emilie M. et al. (2017). "The Behavioral Revolution and International Relations." In: *International Organization* 71 (S1), S1–S31.
Hall, Todd and Keren Yarhi-Milo (2012). "The Personal Touch: Leaders' Impressions, Costly Signaling, and Assessments of Sincerity in International Affairs." In: *International Studies Quarterly* 56 (3), pp. 560–573.
Hamilton, David L. and Steven J. Sherman (1996). "Perceiving Persons and Groups." In: *Psychological Review* 103 (2), pp. 336–355.
Hamilton, David L., Steven J. Sherman, and Luigi Castelli (2002). "A Group By Any Other Name–The Role of Entitativity in Group Perception." In: *European Review of Social Psychology* 12 (1), pp. 139–166.
Haslam, Nick, Louis Rothschild, and Donald Ernst (2004). "Essentialism and Entitativity: Structures of Beliefs about the Ontology of Social Categories." In: *The Psychology of Group Perception*. Ed. by Vincent Yzerbyt, Charles M. Judd, and Olivier Corneille. Psychology Press, pp. 47–59.
Heider, Fritz (1958). *The Psychology of Interpersonal Relations*. Wiley.
Hewstone, Miles (1990). "The 'Ultimate Attribution Error'? A Review of the Literature on Intergroup Causal Attribution." In: *European Journal of Social Psychology* 20 (4), pp. 311–335.
Holmes, Marcus and Nicholas J. Wheeler (2020). "Social Bonding in Diplomacy." In: *International Theory* 12 (1), pp. 133–161.
Holsti, Ole R. (1962). "The Belief System and National Images: A Case Study." In: *Journal of Conflict Resolution* 6 (3), pp. 244–252.
Hopf, Ted (2010). "The Logic of Habit in International Relations." In: *European Journal of International relations* 16 (4), pp. 539–561.
Hurd, Ian (2008). "Constructivism." In: *The Oxford Handbook of International Relations*. Ed. by Christian Reus-Smit and Duncan Snidal. Oxford University Press, pp. 298–316.
Hyde, Susan D. (2015). "Experiments in International Relations: Lab, Survey, and Field." In: *Annual Review of Political Science* 18 (1), pp. 403–424.
Jackman, Simon (2009). *Bayesian Analysis for the Social Sciences*. Wiley.
Jackson, Patrick Thaddeus and Daniel H. Nexon (1999). "Relations Before States." In: *European Journal of International Relations* 5 (3), pp. 291–332.
Jervis, Robert (1970). *The Logic of Images in International Relations*. Princeton University Press.
Jervis, Robert (1976). *Perception and Misperception in International Politics*. Princeton University Press.
Jones, Edward E. and Victor A. Harris (1967). "The Attribution of Attitudes." In: *Journal of Experimental Social Psychology* 3 (1), pp. 1–24.
Kaufmann, Chaim D. (1994). "Out of the Lab and into the Archives: A Method for Testing Psychological Explanations of Political Decision Making." In: *International Studies Quarterly* 38 (4), pp. 557–586.
Keohane, Robert O. (1984). *After Hegemony: Cooperation and Discord in the World Political Economy*. Princeton University Press.
Kertzer, Joshua D. (2016). *Resolve in International Politics*. Princeton University Press.
Kertzer, Joshua D. (2017). "Microfoundations in International Relations." In: *Conflict Management and Peace Science* 34 (1), pp. 81–97.

Kertzer, Joshua D., Marcus Holmes, et al. (2022). "Hawkish Biases and Group Decision Making." In: *International Organization* 76 (3), pp. 513–548.

Kertzer, Joshua D. and Kathleen M. McGraw (2012). "Folk Realism: Testing the Microfoundations of Realism in Ordinary Citizens." In: *International Studies Quarterly* 56 (2), pp. 245–258.

Kertzer, Joshua D., Brian C. Rathbun, and Nina Srinivasan Rathbun (2020). "The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations." In: *International Organization* 74 (1), pp. 95–118.

Kertzer, Joshua D. and Dustin H. Tingley (2018). "Political Psychology in International Relations: Beyond the Paradigms." In: *Annual review of political science* 21 (1), pp. 319–339.

Kruglanski, Arie W. and Donna M. Webster (1996). "Motivated Closing of the Mind: "Seizing" and "Freezing"." In: *Psychological Review* 103 (2), pp. 263–283.

Krull, Douglas S. et al. (1999). "The Fundamental Fundamental Attribution Error: Correspondence Bias in Individualist and Collectivist Cultures." In: *Personality & Social Psychology Bulletin* 25 (10), pp. 1208–1219.

Kuehnen, Ulrich et al. (2013). "How Self-construal Affects Dispositionalism in Attributions." In: *Social Cognition* 31 (2), pp. 237–259.

Kurizaki, Shuhei (2007). "Efficient Secrecy: Public versus Private Threats in Crisis Diplomacy." In: *American Political Science Review* 101 (3), pp. 543–558.

Kydd, Andrew H. (1997). "Game Theory and the Spiral Model." In: *World Politics* 49 (3), pp. 371–400.

Kydd, Andrew H. (2005). *Trust and Mistrust in International Relations*. Princeton University Press.

Larson, Deborah Welch (1997a). *Anatomy of Mistrust : U.S.-Soviet Relations During the Cold War*. Cornell University Press.

Larson, Deborah Welch (1997b). "Trust and Missed Opportunities in International Relations." In: *Political Psychology* 18 (3), pp. 701–734.

Lebow, Richard Ned and Janice Gross Stein (2019). "Afghanistan, Carter, and Foreign Policy Change: The Limits of Cognitive Models." In: *Diplomacy, Force, and Leadership: Essays in Honor of Alexander L. George*. Ed. by Dan Caldwell and Timothy J. McKeown. Westview.

Legro, Jeffrey W. (2007). "What China Will Want: The Future Intentions of a Rising Power." In: *Perspectives on Politics* 5 (3), pp. 515–534.

Levy, Sheri R., Chi-yue Chiu, and Ying-yi Hong (2006). "Lay Theories and Intergroup Relations." In: *Group Processes & Intergroup Relations* 9 (1), pp. 5–24.

MacDonald, Paul K. (2003). "Useful Fiction or Miracle Maker: The Competing Epistemological Foundations of Rational Choice Theory." In: *American Political Science Review* 97 (4), pp. 551–565.

Markus, Hazel Rose and Shinobu Kitayama (1991). "Culture and the Self: Implications for Cognition, Emotion, and Motivation." In: *Psychological Review* 98 (2), p. 224.

Masuda, Takahiko and Shinobu Kitayama (2004). "Perceiver-Induced Constraint and Attitude Attribution in Japan and the US: A Case for the Cultural Dependence of the Correspondence Bias." In: *Journal of Experimental Social Psychology* 40 (3), pp. 409–416.

McDermott, Rose (2002). "Experimental Methodology in Political Science." In: *Political analysis* 10 (4), pp. 325–342.

Mearsheimer, John J. (2014). *The Tragedy of Great Power Politics*. Updated edition. W. W. Norton & Company.

Mercer, Jonathan (1996). *Reputation and International Politics*. Cornell University Press.

Mercer, Jonathan (2005). "Rationality and Psychology in International Politics." In: *International Organization* 59 (1), pp. 77–106.
Miller, Joan G (1984). "Culture and the Development of Everyday Social Explanation." In: *Journal of Personality and Social Psychology* 46 (5), pp. 961–978.
Molden, Daniel C. and Carol S. Dweck (2006). "Finding "Meaning" in Psychology: A Lay Theories Approach to Self-Regulation, Social Perception, and Social Development." In: *American Psychologist* 61 (3), pp. 192–203.
Morris, Michael W., Richard E. Nisbett, and Kaiping Peng (1996). "Causal Attribution across Domains and Cultures." In: *Causal Cognition: A Multidisciplinary Debate*. Ed. by Dan Sperber, David Premack, and Ann James Premack. Oxford University Press.
Morris, Michael W. and Kaiping Peng (1994). "Culture and Cause: American and Chinese Attributions for Social and Physical Events." In: *Journal of Personality and Social Psychology* 67 (6), pp. 949–971.
Morrow, James D. (1999). "The Strategic Setting of Choices: Signaling, Commitment, and Negotiation in International Politics." In: *Strategic Choice and International Relations*. Ed. by David A. Lake and Robert Powell. Princeton University Press, pp. 77–114.
Moscovici, Serge (1984). "The Phenomenon of Social Representations." In: *Social Representations*. Ed. by R. M. Farr and Serge Moscovici. Cambridge University Press, pp. 3–69.
Moscovici, Serge (1988). "Notes Towards a Description of Social Representations." In: *European Journal of Social Psychology* 18 (3), pp. 211–250.
Nisbett, Richard E. (2003). *The Geography of Thought: How Asians and Westerners Think Differently - And Why*. Nicholas Brealey.
Nisbett, Richard E. et al. (2001). "Culture and Systems of Thought: Holistic versus Analytic Cognition." In: *Psychological Review* 108 (2), pp. 291–310.
Norenzayan, Ara, Incheol Choi, and Richard E. Nisbett (2002). "Cultural Similarities and Differences in Social Inference: Evidence from Behavioral Predictions and Lay Theories of Behavior." In: *Personality & Social Psychology Bulletin* 28 (1), pp. 109–120.
Pettigrew, Thomas F. (1979). "The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice." In: *Personality & Social Psychology Bulletin* 5 (4), pp. 461–476.
Plaks, Jason E., Sheri R. Levy, and Carol S. Dweck (2009). "Lay Theories of Personality: Cornerstones of Meaning in Social Cognition." In: *Social and Personality Psychology Compass* 3 (6), pp. 1069–1081.
Powell, Robert (2017). "Research Bets and Behavioral IR." In: *International Organization* 71 (S1), S265–S277.
Quek, Kai (2016). "Are Costly Signals More Credible? Evidence of Sender-Receiver Gaps." In: *The Journal of Politics* 78 (3), pp. 925–940.
Quek, Kai (2021). "Four Costly Signaling Mechanisms." In: *American Political Science Review* 115 (2), pp. 537–549.
Ramirez, Luisa et al. (2013). "The Study of Lay Theories: A Piece of the Puzzle for Understanding Prejudice." In: *Navigating the Social World: What Infants, Children, and Other Species Can Teach Us*. Ed. by Mahzarin R. Banaji and Susan A. Gelman. Oxford University Press.
Rathbun, Brian C. (2007). "Uncertain about Uncertainty: Understanding the Multiple Meanings of a Crucial Concept in International Relations Theory." In: *International Studies Quarterly* 51 (3), pp. 533–557.
Rathbun, Brian C. (2019). *Reasoning of State: Realists, Romantics and Rationality in International Relations*. Cambridge University Press.

Rathbun, Brian C., Joshua D. Kertzer, and Mark Paradis (2017). "Homo Diplomaticus: Mixed-Method Evidence of Variation in Strategic Rationality." In: *International Organization* 71 (S1), S33–S60.

Renshon, Jonathan (2008). "Stability and Change in Belief Systems: The Operational Code of George W. Bush." In: *Journal of Conflict Resolution* 52 (6), pp. 820–849.

Rosati, Jerel A. (1988). "Continuity and Change in the Foreign Policy Beliefs of Political Leaders: Addressing the Controversy over the Carter Administration." In: *Political Psychology* 9 (3), pp. 471–505.

Rosato, Sebastian (2015). "The Inscrutable Intentions of Great Powers." In: *International Security* 39 (3), pp. 48–88.

Rosato, Sebastian (2021). *Intentions in Great Power Politics: Uncertainty and the Roots of Conflict*. Yale University Press.

Rosenberg, Shawn W. and Gary Wolfsfeld (1977). "International Conflict and the Problem of Attribution." In: *Journal of Conflict Resolution* 21 (1), pp. 75–103.

Ross, Lee (1977). "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process." In: *Advances in Experimental Social Psychology* 10, p. 174.

Schelling, Thomas C. (1960). *The Strategy of Conflict*. Harvard University Press.

Silverstein, Brett and Catherine Flamenbaum (1989). "Biases in the Perception and Cognition of the Actions of Enemies." In: *Journal of Social Issues* 45 (2), pp. 51–72.

Solimena, Paola M. (2022). "Seen One, Seen Them All? Group Perception and Reputation Formation in International Politics."

Stein, Janice Gross (2017). "The Micro-Foundations of International Relations Theory: Psychology and Behavioral Economics." In: *International Organization* 71 (S1), S249–S263.

Tang, Shiping (2008). "Fear in International Politics: Two Positions." In: *International Studies Review* 10 (3), pp. 451–471.

Tang, Shiping (2009). "The Security Dilemma: A Conceptual Analysis." In: *Security Studies* 18 (3), pp. 587–623.

Taylor, Donald M. and Vaishna Jaggi (1974). "Ethnocentrism and Causal Attribution in a South Indian Context." In: *Journal of Cross-Cultural Psychology* 5 (2), pp. 162–171.

Tetlock, Philip (1999). "Theory-Driven Reasoning about Possible Pasts and Probable Futures in World Politics: Are We Prisoners of our Preconceptions?" In: *American Journal of Political Science* 43 (2), p. 335.

Thaler, Richard H. and Cass R. Sunstein (2009). *Nudge*. Penguin Books.

Tingley, Dustin H. and Barbara F. Walter (2011). "Can Cheap Talk Deter? An Experimental Analysis." In: *Journal of Conflict Resolution* 55 (6), pp. 996–1020.

Trager, Robert F. (2017). *Diplomacy: Communication and the Origins of International Order*. Cambridge University Press.

Varnum, Michael E. W. et al. (2010). "The Origin of Cultural Differences in Cognition: The Social Orientation Hypothesis." In: *Current Directions in Psychological Science* 19 (1), pp. 9–13.

Vertzberger, Yaacov (1990). *The World in Their Minds: Information Processing, Cognition and Perception in Foreign Policy Decisionmaking*. Stanford University Press.

Waltz, Kenneth N. (1959). *Man, the State, and War: A Theoretical Analysis*. Columbia University Press.

Waltz, Kenneth N. (1979). *Theory of International Politics*. McGraw-Hill.

Wendt, Alexander (1992). "Anarchy is What States Make of It: The Social Construction of Power Politics." In: *International Organization* 46 (2), pp. 391–425.

Wendt, Alexander (1995). "Constructing International Politics." In: *International Security* 20 (1), pp. 71–81.

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

Wheeler, Nicholas J. (2018). *Trusting Enemies: Interpersonal Relationships in International Conflict*. Oxford University Press.

Wolfers, Arnold (1962). *Discord and Collaboration : Essays on International Politics*. Johns Hopkins University Press.

Yarhi-Milo, Keren (2013). "In the Eye of the Beholder: How Leaders and Intelligence Communities Assess the Intentions of Adversaries." In: *International Security* 38 (1), pp. 7–51.

Yarhi-Milo, Keren (2014). *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations*. Princeton University Press.

Yarhi-Milo, Keren (2018). *Who Fights for Reputation: The Psychology of Leaders in International Conflict*. Princeton University Press.

Yarhi-Milo, Keren, Joshua D. Kertzer, and Jonathan Renshon (2018). "Tying Hands, Sinking Costs, and Leader Attributes." In: *Journal of Conflict Resolution* 62 (10), pp. 2150–2179.

Yoder, Brandon K. and Kyle Haynes (2021). "Signaling under the Security Dilemma: An Experimental Analysis." In: *Journal of Conflict Resolution* 65 (4), pp. 672–700.