

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Do Unto Others: A Behavioral Approach to Reciprocity in International Relations

Permalink

<https://escholarship.org/uc/item/2vb82636>

ISBN

9798189227233

Author

Roskos-Ewoldsen, Andrew

Publication Date

2026-06-25

Peer reviewed|Thesis/dissertation

Do Unto Others:

A Behavioral Approach to Reciprocity
in International Relations

By
ANDREW ROSKOS-EWOLDSEN
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in
Political Science
in the
OFFICE OF GRADUATE STUDIES
of the
UNIVERSITY OF CALIFORNIA
DAVIS

Approved:
Brandon Kinne, Chair
Lauren Young
Juan Tellez
Brian Rathbun
Committee in Charge

2026

Copyright © 2026 by Andrew Roskos-Ewoldsen

Acknowledgements

I could not have written this dissertation without the wisdom, guidance, and support of many great people around me.

I would first like to thank my advisor, Brandon Kinne, for the tireless support he provided me throughout my graduate career at Davis. Beyond his tremendous breadth of knowledge in international relations, his efforts to make himself available for his students is unparalleled, and were an immeasurably important part of how I moved forward with this dissertation. Additionally, being able to watch first-hand his dedication and skill as a teacher, researcher, writer, and mentor was an invaluable experience, and will always serve as an example by which I can be guided by in my own career. I feel extremely fortunate and deeply indebted for having learned so much from him. I am also very grateful to Lauren Young for her constant reassurance and encouragements to challenge myself as a researcher. My admiration for her work was one of the reasons I came to UC Davis 6 years ago, and our many discussions have been an enormous influence on how I find and approach the research puzzles I am most interested in, including this project. Juan Tellez has also been extremely gracious and helpful to me since moving to UC Davis, and apart from his instrumental contributions to the observational portions of my dissertation, I've benefited greatly from his warm and instructive mentoring style. I cannot overstate how important Brian Rathbun has been to the development of this project, and my maturation as a researcher. His research has been foundational for me dating back to my undergraduate years almost 15 years ago, and continues to shape how I see the world. I feel deeply honored to have been able to learn from him personally, and grateful for his willingness to fill the gaps of my knowledge and experience with his own.

I want to also thank my colleagues and friends who I've met at Davis. In particular, I feel lucky to have taken every step of my graduate school journey with my cohort members Anya, Alexa, Lily, Zach, and Jon, without whom I would have certainly felt lost. Thank you to Alex, Anissa, Rana, Mieke, Denon, Cam, and Kodi, who have also been unforgettable parts of my experience here.

I am also extremely grateful for the generous financial and academic support that I received that was integral to the completion of this dissertation, including the Russell J. and Dorothy S. Bilinski Educational Foundation, the Harry Frank Guggenheim Foundation, the American Political Science Association, as well as the political science department here at the University of California, Davis.

Finally, I would like to acknowledge those that have unfailingly provided me support every day that I've been in this program. My partner Linh remains inextricably rooted to any strength, motivation, or resilience that I may possess, and without her warmth, her humor, her intelligence, and her care, this entire ordeal would not have been possible. She has been an ever present lighthouse for me in both turbulent waters and calm. I would also like to thank my family for doing what they do best: providing unconditional love and support. Elliott, Nancy, Connor, Deana, Chandler, Virginia, Jason, Elizabeth, Eric, Hannah, and George – I am eternally grateful to have you in my life. I would lastly like to dedicate this dissertation to my parents, David and Beverly, and my brother, Jonathan. There is too much to thank you for, so I will just say I am grateful to be your son and your brother.

Abstract

This dissertation examines how reciprocity shapes individual behavior in international relations, with implications for cooperation and conflict at the state level. Reciprocity, or the returning of a behavior, has long been studied as both a theoretical strategy and an observed phenomenon in international relations research. Prior work has largely focused on the state level, examining how reciprocity drives or constrains security cooperation, conflict escalation, and alliance management. At the same time, a rich body of research in psychology and sociology demonstrates that reciprocity operates as a powerful behavioral heuristic at the individual level, where humans have developed psychological shortcuts that encourage reciprocal behavior as a means for survival and overcoming barriers to cooperation.

This dissertation bridges these two lines of research. Drawing on the theory of Bounded Generalized Reciprocity (BGR) from social and evolutionary psychology, I argue that individuals use reciprocity as a heuristic guide to decision-making in international contexts: they return favors and slights in a strategic but not strictly rational fashion, and this has important implications for how we understand cooperation and conflict in world politics. Rather than treating reciprocity solely as a state-level strategy, I examine it as a micro-level behavioral process that may aggregate into the cooperative and conflictual patterns observed among states.

I focus on three distinct but related dimensions of reciprocity that are important yet underexplored in international relations research: indirect reciprocity, group-based reciprocity, and costly reciprocity. Each constitutes a standalone research article, and together they constitute a broader empirical investigation of how and why individuals use reciprocity in international contexts. My

research employs a combination of methods, including survey experiments administered to nationally representative samples of the US public and observational analysis of country-level event data from the Integrated Crisis Early Warning System (ICEWS).

The first article examines indirect or “third-party” reciprocity, asking whether people withdraw support for cooperation with countries that fail to return favors to *other* countries, or not their own. I find that individuals penalize reciprocity violations even when they are uninvolved, and that this effect is driven by moral, reputational, and emotional mechanisms consistent with the BGR framework. Complementary observational analysis of directed country-pair data suggests that state-level behavior follows a similar logic.

The second article investigates whether ingroup or outgroup membership moderates the reciprocity effect, asking whether people are more forgiving of friends or harsher in punishing them for cooperative failures. Contrary to theoretical expectations, group identity does not consistently moderate the reciprocity effect. However, reciprocity itself, and group identity independently, are powerful predictors of cooperative support and trust, with outgroups that use negative reciprocity penalized most severely.

The third article explores costly reciprocity, examining whether individuals support the use of reciprocity in international security contexts even when doing so is expensive or lacks an international audience. Results indicate that the public is sensitive to costs in ways not predicted by the BGR heuristic framework, and that reputational, moral, and fairness motivations all partially drive the use of reciprocity.

Taken together, these studies offer new insights into a foundational concept in international relations and suggest that behavioral and psychological approaches to reciprocity can enrich our understanding of when and why countries cooperate or conflict.

Contents

Acknowledgements	iii
Abstract	v
1 Introduction	1
1.1 Overview and Motivation	1
1.2 The Empirical Puzzle	2
1.3 Theoretical Framework: Bounded Generalized Reciprocity	3
1.4 Research Design and Three Studies	5
1.5 Contributions and Dissertation Roadmap	7
2 Following the Golden Rule: The Effect of Indirect Reciprocity on Public Support for International Cooperation	18
3 Who Needs Friends Like These? Group Membership, Reciprocity, and Public Support for International Cooperation	66
4 The Price of Retaliation: Understanding Costly Reciprocity in International	

Relations	135
5 Reciprocity in International Relations: A Synthesis	180
5.1 Overview	180
5.2 Summary of Findings	181
5.3 Theoretical Implications	183
5.4 Implications for International Relations Theory and Practice	184
5.5 Limitations and Future Directions	185
5.6 Conclusion	186

List of Figures

2.1	Direct and Indirect Reciprocity	27
2.2	Support for Cooperation by Treatment Group + 95% Confidence Intervals	38
2.3	Response Distributions by Treatment Group	39
2.4	Indirect Reciprocity Effect on Support for Coop	40
2.5	Indirect Reciprocity Effect on Support for Cooperation with Vexa	41
2.6	Effect of Indirect Reciprocity on Cooperative Behavior	46
2.7	Effect of Indirect Reciprocity on Conflictual Behavior	56
3.1	Policy Support and Trust by Reciprocity Treatment Group	90
3.2	Policy Support and Trust by Group Treatment	91
3.3	The Effects of Reciprocity and Groups on Cooperation	92
3.4	Policy Support and Trust by All Treatment Groups	93
3.5	Interaction Coefficients	94
3.6	Mediators of the Violation Reciprocity Effect on Policy Support	123

4.1	Support for Reciprocity by Cost Treatment Group + 95% Confidence Intervals . . .	154
4.2	Support for Reciprocity by Group Treatment Group + 95% Confidence Intervals . .	155
4.3	Cost Effect on Support for Reciprocity	156
4.4	Dovishness Moderates the Visibility Effect on Public Reciprocity	160
4.5	Hawkishness Moderates the Visibility Effect on Private Reciprocity	160
4.6	Isolationism Moderates the Visibility Effect on Private Reciprocity	161
4.7	Hawkishness Moderates the Visibility Effect on General Reciprocity	161
4.8	Hawkishness Moderates the Costliness Effect on Reciprocity	163
4.9	Isolationism Moderates the Costliness Effect on Private Reciprocity	164
4.10	Costliness Effect on Public Reciprocity Towards Country A	168

List of Tables

2.1	OLS Multiple Regression Models	53
2.2	Reputation Mechanism	53
2.3	Bad Behavior Mechanism	53
2.4	Moral/Fairness Mechanism	54
2.5	Emotions Mechanism	54
3.1	Summary of Hypothesis Tests	96
3.2	Main Models: Effect of Treatment on Policy Support and Trust	118
3.3	Interaction Models: Policy Support and Trust	119
3.4	Effects of Treatment on Threat Perception (Supplementary)	120
3.5	Reputation	124
3.6	Morality	124
3.7	Fairness	124
3.8	Affect	125

3.9	Long Run Gain	125
3.10	Future Behavior	125
4.1	FPI as Moderators of Visibility Effect	165
4.2	FPI as Moderators of Cost Effect	166

Chapter 1

Introduction

1.1 Overview and Motivation

Reciprocity, or the returning of a behavior, is one of the most fundamental and widely observed dynamics in human social life, and international relations is no exception. From the tit-for-tat logic of deterrence to the cooperative norms embedded in international institutions, reciprocity shapes how states interact, how cooperation emerges, and how conflicts escalate. And yet, despite its centrality to both scholarly accounts and everyday descriptions of world politics, we still lack a comprehensive understanding of why and how *individuals* use reciprocity when making decisions in international contexts.

This is the central question this dissertation takes on. Reciprocity has long been studied in international relations as a state-level phenomenon, with foundational works examining how and when states return favors, match cooperation, or retaliate in kind (Axelrod 1980; Crescenzi, Best and Kwon 2010; Goldstein and Freeman 1991; Goldstein and Pevehouse 1997; Keohane 1986). These studies have produced important insights, but they tend to treat reciprocity as an instrument of statecraft rather than as a behavioral tendency rooted in human psychology. Separately, a large body of research in psychology and sociology has demonstrated that reciprocity operates as a powerful heuristic at the individual level, guiding behavior across a vast range of social and economic

contexts (Boyd and Richerson 1988; Fehr and Fischbacher 2004; Fehr, Fischbacher and Gächter 2002; Yamagishi, Jin and Kiyonari 1999). This individual-level research has not, however, been systematically brought to bear on international relations.

My dissertation bridges these two traditions. Drawing on the theory of Bounded Generalized Reciprocity (BGR) from evolutionary and social psychology (Yamagishi, Jin and Kiyonari 1999), I argue that individuals use reciprocity as a mental shortcut when navigating the complex strategic landscape of international relations, just like they do in everyday social life. Rather than exhaustively calculating costs and benefits, people rely on the simple rule of reciprocity: return what you receive, and punish or avoid those who do not. This rule is deeply embedded in human psychology because it is effective; over the long run, following reciprocity norms facilitates cooperation and punishes defection, leaving individuals better off. But because it operates as a heuristic rather than a fully rational calculation, it can produce behaviors that diverge from what we would expect under standard rational-actor models.

The scope of this research extends beyond security alone. While earlier formulations of this project focused primarily on security cooperation and conflict, this dissertation takes a broader view, examining reciprocity as a driver of international relations outcomes more generally; I examine cooperation across both security and non-security domains, trust in foreign counterparts, and responses to conflict-related behaviors. This approach is in keeping with the theory of BGR, where if reciprocity is a general behavioral heuristic, its effects should be observable across a range of international outcomes.

1.2 The Empirical Puzzle

Why do states and individuals seem to go out of their way to return favors or slights, even when doing so is costly or involves countries that were not the original source of the behavior? Standard state-level, rational accounts of international relations struggle with this question. And yet anecdotes from world politics tell us otherwise.

During the COVID-19 pandemic, China's failure to return medical aid to Italy and Spain, coun-

tries that had donated supplies when China was struggling, generated widespread international criticism and a measurable cooling in relations. The United States’ withdrawal from the Trans-Pacific Partnership under President Trump was interpreted by many partners not merely as a strategic recalibration but as a violation of cooperative reciprocity norms, leading to downstream consequences for American credibility. Iran’s provision of military drones to Russia during the Ukraine conflict can be understood as the reciprocation of Russian intervention in Syria on Iran’s behalf years earlier; the returning of a favor whose timing and costs make the act difficult to explain.

At the individual level, these patterns are even harder to explain without reference to the psychological mechanisms that drive reciprocal behavior. Public opinion data consistently shows that citizens’ willingness to cooperate with, support, or sanction foreign countries is shaped by perceptions of whether those countries have followed the rules of reciprocity. People want their countries to “get even,” to “pay back” cooperation, and to punish those who do not return favors. This can happen even when their country cannot stand to benefit immediately, even when the original exchange did not involve them, and even when the counterpart is nominally a friend.

Understanding why individuals use reciprocity in these ways is not merely an academic exercise. The public plays an increasingly documented role in shaping state foreign policy, both through electoral mechanisms and through the constraints that public opinion places on political elites (Aldrich, Gelpi, Feaver, Reifler and Sharp 2006; Chiba, Johnson and Leeds 2015; Tomz, Weeks and Yarhi-Milo 2020; Yarhi-Milo, Kertzer and Renshon 2018). If the public systematically uses reciprocity as a heuristic in international contexts, this has implications for when we can expect cooperative initiatives to succeed, when retaliation will be demanded, and how leaders can manage the reciprocity expectations of their domestic audiences.

1.3 Theoretical Framework: Bounded Generalized Reciprocity

The theoretical foundation of this dissertation is the theory of Bounded Generalized Reciprocity (BGR), developed by Yamagishi, Jin and Kiyonari (1999) in social psychology. BGR posits that humans have evolved to use reciprocity as a default behavioral heuristic in strategic interactions,

because in the long run following reciprocity norms produces self-interest benefits from cooperation. The theory is grounded in evolutionary psychology, which argues that humans have developed psychological mechanisms that encourage reciprocal behavior—mechanisms like feelings of moral obligation, emotional responses to violations, and sensitivity to reputation—because these mechanisms helped our ancestors survive and cooperate in complex social environments (Boyd and Richerson 1988, 1989; Fehr, Fischbacher and Gächter 2002).

Several features of BGR distinguish it from rational-actor accounts of reciprocity and make it well-suited to understanding behavior in international contexts.

First, BGR treats reciprocity as a heuristic rather than a fully calculated strategy. This means individuals use reciprocity as a rule of thumb or a cognitive shortcut that guides behavior without requiring exhaustive cost-benefit analysis. The implications are significant: individuals may follow reciprocity norms even when doing so is costly, suboptimal, or based on incomplete information. They may punish violations of reciprocity even when they are not directly affected. They may extend reciprocity to groups or collectives (like countries) that they have never interacted with directly.

Second, BGR identifies specific psychological mechanisms that drive reciprocity behavior. These center mostly on reputational concerns (the desire to be seen as a reliable cooperation partner), but I synthesize other research on reciprocity and introduce moral obligation (the sense that returning favors or punishing violations is the “right” or “fair” thing to do), concern for correcting bad behavior (the belief that punishing defection will discourage future violations), and emotional responses (anger at violations, satisfaction at reciprocation). Each of these mechanisms generates testable predictions about when and why individuals will use reciprocity.

Third, BGR makes predictions about how group membership moderates reciprocity. Because interacting with ingroup members offers greater potential cooperative benefits, and greater potential losses from defection, BGR predicts heightened reciprocity sensitivity toward ingroup members. People are expected to cooperate more generously with friends and to punish their violations more harshly.

Fourth, BGR is agnostic about the specific domain of interaction. While much of the empirical research using BGR has focused on economic games or everyday social interactions, the theory’s logic is general enough to apply to international contexts. If reciprocity is a fundamental human heuristic, there is no reason to expect it to be “switched off” when individuals think about interactions between states.

1.4 Research Design and Three Studies

I test the predictions of a BGR-based approach to reciprocity in international relations across three empirical studies, each constituting an independent research article. The three studies examine distinct but related dimensions of reciprocity: indirect reciprocity, group-based reciprocity, and costly reciprocity. Together, they cover a range of international contexts—including security and non-security cooperation, positive and negative reciprocity, and interactions with both familiar and unfamiliar foreign actors.

The primary empirical strategy across all three studies is the survey experiment, administered to nationally representative samples of US adults. Survey experiments offer important methodological advantages for studying reciprocity in international contexts: they allow causal identification of reciprocity effects through random assignment, they provide control over the specific features of reciprocity situations that BGR identifies as theoretically important (like cost, visibility, and group membership), and they use real-world vignettes grounded in plausible international scenarios. Where appropriate, I supplement experimental findings with observational analysis of country-level event data, providing a bridge between the individual-level experimental results and state-level behavioral patterns.

Chapter 2: Indirect Reciprocity

The first research article examines indirect or “third-party” reciprocity, or the tendency for individuals to condition their cooperative behavior toward a country based on how that country has treated *other* countries, not their own. I argue that people, acting as third-party observers, pay

attention to whether countries follow the “golden rule” of reciprocity, and punish those who fail to return favors even when the violation did not involve them directly.

The study employs a novel survey experiment on a nationally representative sample of 1,000 US adults, in which participants learn that a fictitious country either did or did not return a cooperative gesture to an unrelated third country, and then report their support for cooperation with that country. I also complement the experimental findings with an observational analysis of directed country-pair event data from ICEWS.

Results strongly support the indirect reciprocity hypothesis: public support for cooperation decreases significantly when participants learn that the potential partner country has violated reciprocity with others. Mediation analysis confirms that this effect operates through mechanisms associated with the BGR framework, including moral obligation, reputational concerns, and emotional responses. Observational evidence is partially consistent with the experimental results, with countries that reciprocate material cooperative or conflictual acts receiving more cooperative behavior from uninvolved third parties.

Chapter 3: Group Reciprocity

The second research article examines whether ingroup or outgroup status moderates the effect of reciprocity on public support for international cooperation. BGR predicts that people hold higher cooperative expectations for ingroup members and punish their violations more severely. I test this prediction using a novel survey experiment in which participants are randomly assigned to one of three groupness conditions (control, ingroup, outgroup) crossed with one of three reciprocity conditions (positive reciprocity, violation of positive reciprocity, negative reciprocity).

Contrary to the theoretical expectation, I find that group identity does not consistently moderate the reciprocity effect: the penalty for violating reciprocity is not significantly larger for ingroup members than for control or outgroup countries. However, the results reveal an interesting amplification effect: outgroup countries that use negative reciprocity are penalized the most steeply, a finding that has implications for how group identity and reciprocity interact in international

contexts. I also find that both reciprocity and group identity independently and substantially shape public support for cooperation and trust in foreign counterparts, with reciprocity exhibiting a particularly large effect.

Chapter 4: Costly Reciprocity

The third research article examines whether individuals support the use of reciprocity in international security contexts even when doing so is costly or lacks an international audience. I employ a 2x2 factorial survey experiment in which the cost of the reciprocal act and the visibility of the original offense are independently manipulated. Results reveal that the public is sensitive to costs in ways that diverge from the BGR heuristic framework: people are more opposed to negative reciprocity when costs are higher. I also find that visibility conditionally affects the type of negative reciprocity that people prefer, and that the effects of costliness and visibility are moderated in interesting ways by foreign policy orientations. Finally, I show that reputational, moral, and fairness motivations all partially mediate the reciprocity effect.

1.5 Contributions and Dissertation Roadmap

This dissertation makes several contributions to the study of reciprocity and international relations.

First, it develops and tests a psychological, individual-level account of reciprocity in international contexts. While prior work has established that states use reciprocity, this dissertation asks *why* individuals use reciprocity and identifies the specific mechanisms through which this occurs, offering a micro-level foundation for macro-level patterns.

Second, it broadens the scope of reciprocity research to include indirect, group-based, and costly forms of reciprocity; three dimensions that are theoretically important but underexplored in international relations. Each study reveals both support for and important limits of the BGR framework, pointing towards a more nuanced understanding of how reciprocity works at the individual level.

Third, it extends the study of reciprocity beyond traditional security to international relations

outcomes more broadly, including modern security scenarios, trade-related cooperation, trust in foreign counterparts, and public opinion dynamics. This broader scope reflects the generality of the reciprocity heuristic and its relevance to international relations across multiple domains.

The remainder of this dissertation proceeds as follows. Chapter 2 presents the indirect reciprocity study. Chapter 3 presents the group reciprocity study. Chapter 4 presents the costly reciprocity study. Each chapter is a self-contained research article with its own introduction, theory, research design, results, and reference list. Chapter 5 integrates the findings across the three studies, discusses their collective implications for international relations theory, and identifies directions for future research.

Bibliography

- Aldrich, John H, Christopher Gelpi, Peter Feaver, Jason Reifler and Kristin Thompson Sharp. 2006. “Foreign policy and the electoral connection.” *Annu. Rev. Polit. Sci.* 9:477–502.
- Alexander, Richard. 2017. *The biology of moral systems*. Routledge.
- Aron, Arthur, Elaine N Aron and Danny Smollan. 1992. “Inclusion of other in the self scale and the structure of interpersonal closeness.” *Journal of personality and social psychology* 63(4):596.
- Axelrod, Robert. 1980. “Effective choice in the prisoner’s dilemma.” *Journal of conflict resolution* 24(1):3–25.
- Axelrod, Robert. 2006. *The evolution of cooperation*. Revised ed. New York: Basic Books.
- Balliet, Daniel, Junhui Wu and Carsten KW De Dreu. 2014. “Ingroup favoritism in cooperation: a meta-analysis.” *Psychological bulletin* 140(6):1556.
- Baum, Matthew A and Philip BK Potter. 2008. “The relationships between mass media, public opinion, and foreign policy: Toward a theoretical synthesis.” *Annu. Rev. Polit. Sci.* 11:39–65.
- Berg, Joyce, John Dickhaut and Kevin McCabe. 1995. “Trust, reciprocity, and social history.” *Games and economic behavior* 10(1):122–142.
- Bergé, Laurent R., Kyle Butts and Grant McDermott. 2026. “Fast and user-friendly econometrics estimations: The R package fixest.” *arXiv preprint arXiv:2601.21749*.
- Boschee, Elizabeth, Jennifer Lautenschlager, Sean O’Brien, Steve Shellman, James Starz and

- Michael Ward. 2015. "ICEWS Coded Event Data." .
URL: <https://doi.org/10.7910/DVN/28075>
- Boyd, Robert and Peter J Richerson. 1988. "The evolution of reciprocity in sizable groups." *Journal of theoretical Biology* 132(3):337–356.
- Boyd, Robert and Peter J Richerson. 1989. "The evolution of indirect reciprocity." *Social networks* 11(3):213–236.
- Brandt, Hannelore, Hisashi Ohtsuki, Yoh Iwasa and Karl Sigmund. 2007. A survey of indirect reciprocity. In *Mathematics for ecology and environmental sciences*. Springer pp. 21–49.
- Brewer, Marilynn B and Linnda R Caporael. 2013. An evolutionary perspective on social identity: Revisiting groups. In *Evolution and social psychology*. Psychology Press pp. 143–161.
- Cameron, James E. 2004. "A three-factor model of social identity." *Self and identity* 3(3):239–262.
- Caporael, Linnda R. 1997. "The evolution of truly social cognition: The core configurations model." *Personality and Social Psychology Review* 1(4):276–298.
- Carlin, Ryan E and Gregory J Love. 2013. "The politics of interpersonal trust and reciprocity: An experimental approach." *Political behavior* 35(1):43–63.
- Carson, Austin. 2016. "Facing off and saving face: covert intervention and escalation management in the Korean War." *International Organization* 70(1):103–131.
- Carson, Austin and Keren Yarhi-Milo. 2017. "Covert communication: The intelligibility and credibility of signaling in secret." *Security Studies* 26(1):124–156.
- Chiba, Daina, Jesse C Johnson and Brett Ashley Leeds. 2015. "Careful commitments: Democratic states and alliance design." *The Journal of Politics* 77(4):968–982.
- Crescenzi, Mark JC, Rebecca H Best and Bo Ram Kwon. 2010. Reciprocity in international studies. In *Oxford Research Encyclopedia of International Studies*.
- Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes." *American political science review* 88(3):577–592.

- Fearon, James D. 1997a. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1):68–90.
- Fearon, James D. 1997b. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1):68–90.
- Fearon, James D. 1998. "Bargaining, enforcement, and international cooperation." *International organization* 52(2):269–305.
- Fehr, Ernst and Simon Gächter. 2002. "Altruistic punishment in humans." *Nature* 415(6868):137–140.
- Fehr, Ernst and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and human behavior* 25(2):63–87.
- Fehr, Ernst, Urs Fischbacher and Simon Gächter. 2002. "Strong reciprocity, human cooperation, and the enforcement of social norms." *Human nature* 13:1–25.
- Frank, Robert H. 2011. "The strategic role of the emotions." *Emotion Review* 3(3):252–254.
- Gartzke, Erik. 1998. "Kant we all just get along? Opportunity, willingness, and the origins of the democratic peace." *American Journal of Political Science* pp. 1–27.
- Gartzke, Erik and Jon R Lindsay. 2015. "Weaving tangled webs: offense, defense, and deception in cyberspace." *Security Studies* 24(2):316–348.
- Gartzke, Erik and Kristian Skrede Gleditsch. 2004. "Why democracies may actually be less reliable allies." *American Journal of Political Science* 48(4):775–795.
- Gartzke, Erik and Kristian Skrede Gleditsch. 2006. "Identity and conflict: Ties that bind and differences that divide." *European Journal of International Relations* 12(1):53–87.
- Gibbs, David N. 1995. "Secrecy and international relations." *Journal of Peace Research* 32(2):213–228.
- Goldstein, Joshua S. 1991. "Reciprocity in superpower relations: An empirical analysis." *International Studies Quarterly* 35(2):195–209.

- Goldstein, Joshua S and John R Freeman. 1990. *Three-way street: Strategic reciprocity in world politics*. Chicago: University of Chicago Press.
- Goldstein, Joshua S and John R Freeman. 1991. "US-Soviet-Chinese relations: Routine, reciprocity, or rational expectations?" *American Political Science Review* 85(1):17–35.
- Goldstein, Joshua S and Jon C Pevehouse. 1997. "Reciprocity, bullying, and international cooperation: Time-series analysis of the Bosnia conflict." *American Political Science Review* 91(3):515–529.
- Gray, Kurt, Adrian F Ward and Michael I Norton. 2014. "Paying it forward: generalized reciprocity and the limits of generosity." *Journal of experimental psychology: General* 143(1):247.
- Grieco, Joseph M. 1988. "Anarchy and the limits of cooperation: a realist critique of the newest liberal institutionalism." *International organization* 42(3):485–507.
- Gummerum, Michaela, Masanori Takezawa and Monika Keller. 2009. "The influence of social category and reciprocity on adults' and children's altruistic behavior." *Evolutionary Psychology* 7(2):147470490900700212.
- Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath et al. 2001. "Cooperation, reciprocity and punishment in fifteen small-scale societies." *American Economic Review* 91(2):73–78.
- Holsti, Ole R. 2004. *Public opinion and American foreign policy*. University of Michigan Press.
- Huth, Paul K. 1988. "Extended deterrence and the outbreak of war." *American Political Science Review* 82(2):423–443.
- Huth, Paul K. 1999. "Deterrence and international conflict: Empirical findings and theoretical debates." *Annual Review of Political Science* 2(1):25–48.
- Hyde, Susan D. 2015. "Experiments in international relations: Lab, survey, and field." *Annual Review of Political Science* 18:403–424.
- Imai, Kosuke and In Song Kim. 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis* 29(3):405–415.

- Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1):51 – 71.
URL: <https://doi.org/10.1214/10-STS321>
- Jervis, Robert. 1976. *Perception and misperception in international politics: New edition*. Princeton University Press.
- Jervis, Robert, Keren Yarhi-Milo and Don Casler. 2021. "Redefining the debate over reputation and credibility in international security: Promises and Limits of New Scholarship." *World Politics* 73(1):167–203.
- Keohane, Robert O. 1986. "Reciprocity in international relations." *International organization* 40(1):1–27.
- Kertzer, Joshua D. 2022. "Re-assessing elite-public gaps in political behavior." *American Journal of Political Science* 66(3):539–553.
- Keysar, Boaz, Benjamin A Converse, Jiunwen Wang and Nicholas Epley. 2008. "Reciprocity is not give and take: Asymmetric reciprocity to positive and negative acts." *Psychological Science* 19(12):1280–1286.
- Kydd, Andrew. 1997. "Game theory and the spiral model." *World Politics* 49(3):371–400.
- Kydd, Andrew. 2000. "Trust, reassurance, and cooperation." *International Organization* 54(2):325–357.
- Kydd, Andrew H. 2007. *Trust and mistrust in international relations*. Princeton, NJ: Princeton University Press.
- Larson, Deborah Welch. 1988. "The psychology of reciprocity in international relations." *Negot. J.* 4:281.
- Leach, Colin Wayne, Martijn Van Zomeren, Sven Zebel, Michael LW Vliek, Sjoerd F Pennekamp, Bertjan Doosje, Jaap W Ouwerkerk and Russell Spears. 2008. "Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification." *Journal of personality and social psychology* 95(1):144.

- Leng, Russell J. 1993a. *Interstate Crisis Behavior, 1816-1980*. Cambridge University Press.
- Leng, Russell J. 1993b. "Reciprocating influence strategies in interstate crisis bargaining." *Journal of Conflict Resolution* 37(1):3-41.
- Lindskold, Svenn. 1978. "Trust development, the GRIT proposal, and the effects of conciliatory acts on conflict and cooperation." *Psychological Bulletin* 85(4):772.
- Marcus, George E, Michael MacKuen, Jennifer Wolak and Luke Keele. 2006. The measure and mismeasure of emotion. In *Feeling politics: Emotion in political information processing*. Springer pp. 31-45.
- McDermott, Rose, Anthony C Lopez and Peter K Hatemi. 2017. "'Blunt Not the Heart, Enrage It': The Psychology of Revenge and Deterrence (November 2017)." *Texas National Security Review*.
- Mellers, Barbara, Alan Schwartz and Ilana Ritov. 1999. "Emotion-based choice." *Journal of experimental psychology: General* 128(3):332.
- Miller, Steven V. 2022. "peacesciencer: An R Package for Quantitative Peace Science Research." *Conflict Management and Peace Science* 39(6):755-779.
- Nowak, Martin A and Karl Sigmund. 2005. "Evolution of indirect reciprocity." *Nature* 437(7063):1291-1298.
- Nowak, Martin A and Sébastien Roch. 2007. "Upstream reciprocity and the evolution of gratitude." *Proceedings of the royal society B: Biological Sciences* 274(1610):605-610.
- Oneal, John R. and Bruce Russett. 2001. "Clear and Clean: The Fixed Effects of the Liberal Peace." *International Organization* 55(2):469-485.
- Osgood, Charles E. 1962. "An alternative to war or surrender."
- Oskamp, Stuart. 1971. "Effects of programmed strategies on cooperation in the Prisoner's Dilemma and other mixed-motive games." *Journal of Conflict Resolution* 15(2):225-259.
- Perugini, Marco, Marcello Gallucci, Fabio Presaghi and Anna Paola Ercolani. 2003. "The personal norm of reciprocity." *European Journal of Personality* 17(4):251-283.

- Poznansky, Michael and Evan Perkoski. 2018. "Rethinking secrecy in cyberspace: the politics of voluntary attribution." *Journal of Global Security Studies* 3(4):402–416.
- Rabin, Matthew et al. 1992. *Incorporating fairness into game theory and economics*. University of California at Berkeley, Department of Economics.
- Rajmaira, Sheen and Michael D Ward. 1990. "Evolving foreign policy norms: Reciprocity in the superpower triad." *International Studies Quarterly* 34(4):457–475.
- Rathbun, Brian C. 2011. *Trust in international cooperation: International security institutions, domestic politics and American multilateralism*. Vol. 121 Cambridge University Press.
- Rathbun, Brian C, Christopher Sebastian Parker and Caleb Pomeroy. 2025. "Separate but unequal: Ethnocentrism and racialization explain the "democratic" peace in public opinion." *American Political Science Review* 119(2):621–636.
- Rathbun, Brian C, Joshua D Kertzer, Jason Reifler, Paul Goren and Thomas J Scotto. 2016. "Taking foreign policy personally: Personal values and foreign policy attitudes." *International Studies Quarterly* 60(1):124–137.
- Reiter, Dan and Allan C Stam. 2002. *Democracies at war*. Princeton University Press.
- Renshon, Jonathan and Ryan Powers. 2024. "Identity and the Social Construction of Reputation in World Politics."
- Rhodes, Carolyn. 1989. "Reciprocity in trade: The utility of a bargaining strategy." *International Organization* 43(2):273–299.
- Romano, Angelo, Daniel Balliet and Junhui Wu. 2017. "Unbounded indirect reciprocity: Is reputation-based cooperation bounded by group membership?" *Journal of Experimental Social Psychology* 71:59–67.
- Roskos-Ewoldsen, Andrew, Morgan Ellithorpe and Brandon J Kinne. 2026. "The Effect of Reciprocity on Trust: International Cooperation and COVID Aid." *Journal of Conflict Resolution* 70(1):115–139.

- Russett, Bruce M and John R Oneal. 2001. "Triangulating peace: Democracy, interdependence, and international organizations." (*No Title*) .
- Sanger, David E. 2012. *Confront and conceal: Obama's secret wars and surprising use of American power*. Crown.
- Schelling, Thomas C. 2008. Arms and influence. In *Strategic Studies*. Routledge pp. 96–114.
- Schroeder, David A, Julie E Steel, Andria J Woodell and Alicia F Bembenek. 2003. "Justice within social dilemmas." *Personality and Social Psychology Review* 7(4):374–387.
- Shinada, Mizuho, Toshio Yamagishi and Yu Ohmura. 2004. "False friends are worse than bitter enemies: "Altruistic" punishment of in-group members." *Evolution and Human Behavior* 25(6):379–393.
- Spence, Michael. 1978. Job market signaling. In *Uncertainty in economics*. Elsevier pp. 281–306.
- Tajfel, Henri. 1970. "Experiments in intergroup discrimination." *Scientific american* 223(5):96–103.
- Tajfel, Henri. 1974. "Social identity and intergroup behaviour." *Social science information* 13(2):65–93.
- Tomz, Michael and Jessica LP Weeks. 2021. "Military alliances and public support for war." *International Studies Quarterly* 65(3):811–824.
- Tomz, Michael, Jessica LP Weeks and Keren Yarhi-Milo. 2020. "Public opinion and decisions about military force in democracies." *International Organization* 74(1):119–143.
- Tomz, Michael R and Jessica LP Weeks. 2013. "Public opinion and the democratic peace." *American political science review* 107(4):849–865.
- Uslaner, Eric M. 2002. *The Moral Foundations of Trust*. Cambridge University Press.
- Ward, Michael D and Sheen Rajmaira. 1992. "Reciprocity and norms in US-Soviet foreign policy." *Journal of Conflict Resolution* 36(2):342–368.
- Ward, Michael Don. 1982. "Cooperation and conflict in foreign policy behavior: Reaction and memory." *International Studies Quarterly* 26(1):87–126.

- Weisiger, Alex and Keren Yarhi-Milo. 2015. "Revisiting reputation: How past actions matter in international politics." *International Organization* 69(2):473–495.
- Wilson, Warner. 1971. "Reciprocation and other techniques for inducing cooperation in the Prisoner's Dilemma game." *Journal of Conflict Resolution* 15(2):167–195.
- Wu, Junhui, Daniel Balliet and Paul AM Van Lange. 2016. "Reputation, gossip, and human cooperation." *Social and Personality Psychology Compass* 10(6):350–364.
- Yamagishi, Toshio, Nobuhito Jin and Toko Kiyonari. 1999. "Bounded generalized reciprocity: ingroup boasting and ingroup favoritism. Adv." *Group Process* 16:161–197.
- Yamagishi, Toshio and Toko Kiyonari. 2000. "The group as the container of generalized reciprocity." *Social psychology quarterly* pp. 116–132.
- Yarhi-Milo, Keren, Joshua D Kertzer and Jonathan Renshon. 2018. "Tying hands, sinking costs, and leader attributes." *Journal of Conflict Resolution* 62(10):2150–2179.
- Zheng, Wenqian, Shengli Cao, Yangyang Wang, Ke Yang, Yunyun Chen and Guoping Song. 2021. "The impact of social value orientation, game context and trust on cooperative behavior after cooperative video game play." *Psychological Reports* 124(3):1353–1369.

Chapter 2

Following the Golden Rule: The Effect of Indirect Reciprocity on Public Support for International Cooperation

Following the Golden Rule: The Effect of Indirect Reciprocity on Public Support for International Cooperation

Andrew Roskos-Ewoldsen

Department of Political Science, University of California, Davis

Abstract

States are constantly presented with opportunities to cooperate, or not, with other countries, so how do they decide whether to engage with or pass on these chances? This study focuses on an understudied but important factor in these decisions: indirect reciprocity. I argue that a country's people, acting as a third-party, pay attention to whether other countries follow "the golden rule", or reciprocate positive acts to others, and support or oppose cooperation with countries based on this past reciprocity towards others. Utilizing a novel survey experiment on a representative sample of US citizens, I find supporting evidence for the effect of indirect reciprocity on public support for cooperation: the public more strongly oppose cooperation with countries that "violate" reciprocity, or do not return favors that were given to them. This is true even if the cooperation issue is security or trade related. Further, I find evidence that the individual motivations to use indirect reciprocity are both strategic and normative; we think we can gain something from following the rules of reciprocity, but we also think punishing those who do not is the "right" or "fair" thing to do. I also complement this survey experiment with an observational analysis of directed country-pair event data from 1995–2013, and find state-level evidence consistent with the experimental results: countries that engage in material reciprocity with third parties subsequently receive more cooperative behavior from uninvolved observers. These findings suggest that individual-level indirect reciprocity may shape state-level policy, and imply that countries and leaders have to think carefully about breaking the rules of reciprocity or not returning favors, even if they feel they have a good reason, because their actions may have downstream consequences.

Introduction

States are constantly presented with opportunities to cooperate with other states, whether on security, trade or other foreign policy issues. How do states determine when to do so, and just as importantly, when to withhold their cooperation? I argue in this paper that an under-explored factor in these decisions is individual-level “indirect reciprocity”, or “third party” reciprocity, where cooperative choices are made in line with the “golden rule” logic of reciprocity: do unto others what is done to you. Or, in these third party cases, doing to countries what they did to other countries.

Reciprocity is already a known, established influence on cooperation between countries, but it has almost exclusively been studied at the “direct” level, or the interactions between two parties/countries (Goldstein 1991; Goldstein and Freeman 1991; Keohane 1986; Ward and Rajmaira 1992; Ward 1982). Countries expect their counterpart to return favors with favors, or insults with insults, and this can dictate the success of cooperation: returning a favor can get the ball rolling on cooperation, while failing to return a favor leads to cooperative break downs. There is considerably less work on whether *other* countries pay attention to whether you follow reciprocity or not, and return favors when they are due, even if these reciprocity interactions did not involve them directly (Goldstein and Freeman 1990; Rajmaira and Ward 1990; Rhodes 1989). Going even further, despite the documented use of reciprocity at the country level and the strategic logic of “tit for tat” reciprocity in abstract scenarios like the prisoner’s dilemma (Axelrod 1980), we lack sufficient explanations for why *individuals* might use reciprocity in international contexts (Larson 1988). In other words, we have yet to explore how and why real people use reciprocity when facing real world cooperative dilemmas.

Events in world politics tell us that countries do seem to pay attention to the reciprocity behavior of other countries, even if these behaviors did not originally involve them. For instance, during the COVID-19 pandemic, Italy and Spain were some of the earliest and most generous countries in helping China try to manage the virus, donating medical supplies during some of China’s roughest periods. China; however, did not return the favor. When Italy and Spain were next hit hard by the virus, China forced Italy to buy back the medical supplies Italy had originally donated,

and sent Spain nearly 60,000 faulty testing kits that had to be returned.¹ Not only did these instances strain relations between these countries directly, it brought significant criticism from other countries around the world, with one official from US President Donald Trump’s administration calling China’s attempts to frame their actions towards Italy as helpful “so disingenuous.”²

The publics of countries seem to pay attention to these instances too. Public warmth and trust towards China across 14 different countries like the US, South Korea, Australia, and Sweden plummeted in the year of the pandemic, with a majority holding negative views about whether China’s leaders would do the right thing in international affairs.³ Of course, other considerations, like the attribution of blame to China for causing the pandemic, in this example make the role of indirect reciprocity less clear. Identifying clean real-world cases of indirect reciprocity at the state level is difficult, not necessarily because the reciprocity dynamic is absent, but because states rarely justify cooperative decisions on the basis of a single factor, and reciprocity considerations are typically entangled with strategic interests, domestic politics, and prior bilateral relationships. This is precisely why indirect reciprocity is a ripe topic for study: any knowledge gained is important for unpacking a notoriously complex cooperative process.

My theoretical expectations are that people, such as members of the public, acting as a “third party”, will express less support for cooperation with another country when that country has violated reciprocity, or the returning of a positive behavior, even when this violation did not directly involve their own country. I offer a psychology-based explanation for the effect of indirect reciprocity on support for international security cooperation. Based on the theory of Bounded Generalized Reciprocity (BGR), I posit that reciprocity works as a heuristic tool that guides our behavior with others, particularly in complex and uncertain strategic interactions where calculating the best choice can be difficult (Yamagishi, Jin and Kiyonari 1999). In the long run, reciprocity is an efficient strategy for survival and accruing cooperative benefits, so human beings have developed psychological mechanisms that encourage us to use and enforce the rules of reciprocity early and

¹ “China forces Italy to buy same coronavirus supplies it had donated to Beijing a few weeks ago,” *Fox News*, 6 April 2020.

“Coronavirus: Countries reject Chinese-made equipment,” *BBC*, 30 March 2020.

² “Italy gave China PPE to help with coronavirus — then China made them buy it back,” *The Spectator*, 4 April 2020.

³ “Unfavorable Views of China Reach Historic Highs in Many Countries,” *Pew Research Center*, 6 October 2020.

often. This is why we sometimes act on the shorthand logic of “what goes around comes around”, and that returning what is owed will pay off for us eventually, even if we are not quite sure how in the moment. This is also why people consider their reputations with others when returning favors, or withhold judgment on those who reciprocate negative acts to “get even”; people think this is the “right” or “fair” thing to do.

To test whether indirect reciprocity affects public support for international cooperation, I field a survey experiment on a nationally representative sample of 1,000 US adults. Participants read a vignette describing a hypothetical cooperation request to the United States from a fictitious country, set in either a security or trade context and based on real-world events. They are then randomly assigned to one of two conditions: a reciprocity violation condition, in which they learned that the requesting country had previously failed to return a cooperative gesture from an unrelated third country, or a control condition, in which no such information was provided. I measure individual support for cooperation as the primary outcome, both because it is a standard measure of public sentiment in international relations research and because prior work suggests public opinion can shape state-level foreign policy decisions (Aldrich et al. 2006; Tomz, Weeks and Yarhi-Milo 2020).

The results strongly support the hypothesis that indirect reciprocity affects public support for cooperation with other countries; public support significantly decreases when participants are exposed to a “reciprocity violation” by another country, even when this violation does not involve their own country. Finally, using mediation analysis, I find empirical support for a psychological approach to indirect reciprocity; mechanisms associated with the use of reciprocity as a heuristic guide to behavior like reputational concerns, long run gains, fairness, and emotion drive the indirect reciprocity effect on public support. I take these findings as initial support for applying this particular theoretical approach to reciprocity in international relations, and the argument that people use the everyday concept of reciprocity - the “golden rule” - in international contexts too. At the individual-level, reciprocity acts as a mental guideline that helps us make difficult strategic choices, and can thus explain why reciprocity behaviors like “third party” reciprocity or tit-for-tat are apparent in world politics today.

To assess whether the individual-level indirect reciprocity effect identified in the experiment has

observable parallels at the state level, I complement the survey with an observational analysis of directed country-pair interactions using event data from the Integrated Crisis Early Warning System (ICEWS). While the experiment establishes a causal effect of indirect reciprocity on public attitudes under controlled conditions, the observational study asks a different but related question: do states that engage in reciprocal behavior toward third parties subsequently receive more cooperative behavior from uninvolved observers? This study provides an initial proof of concept test of whether the same logic scales up to actual state behavior across a large panel of country pairs over time.

I find partial evidence that countries do indeed condition their cooperative behavior towards another country based on whether that country reciprocates to others. Reciprocating material cooperative acts and material conflictual acts is associated with an increase in the likelihood of receiving a cooperative act from a third party. The convergence across both designs, which differ in their level of analysis, data sources, and methodological approach, provide compelling evidence for the broader relevance of indirect reciprocity in international cooperation.

In the following sections of the paper, I first review the literature on reciprocity and cooperation. I follow this by briefly defining and discussing indirect reciprocity for the purposes of this study. I then lay out my hypotheses and theoretical expectations for the effect of indirect reciprocity on public support for security cooperation. I then explain my research strategy and the design of my experimental survey. In the results section, I describe the main findings of my survey experiment. I then discuss and interpret these results, and provide a brief supplementary analysis on potential mechanisms. I then proceed to describe the design, modeling strategy, and results of the observational study, and provide brief takeaways from these findings. Finally, I conclude with a discussion of limitations and thoughts on the implications and future directions of my research.

Literature Review

What factors make the difference between successful cooperation and failure, between getting involved in the affairs of other states or not? While there are many known influences on cooperation, reciprocity is a key feature of international relations research on both cooperation and

conflict (Goldstein and Pevehouse 1997; Keohane 1986). Prior work most commonly focuses on reciprocity as an instrumental strategy for countries to gain the benefits of cooperation; tit-for-tat, or players returning cooperation with cooperation and defection with defection (Axelrod 1980). In both formal and experimental approaches, reciprocity successfully induces cooperation from players in prisoner's-dilemma-style settings (Axelrod 2006; Oskamp 1971; Wilson 1971). Goldstein and Pevehouse (1997) also highlights how reciprocity can be used as a strategy in instances of conflict, where the threat of conflict is tied to the prospects of cooperation with other actors.

Still, most prior research on reciprocity focuses solely on "direct" reciprocity, or the behavior of two actors towards one another. "Indirect" reciprocity, or using reciprocity as a third party, has been studied far less in international relations, and if done so has mostly been restricted to state-level interactions between superpowers or within confined regions (Goldstein and Freeman 1990; Goldstein and Pevehouse 1997). Throughout human society; however, people constantly demonstrate their willingness to "pay it forward" as third party observers of reciprocity (Fehr and Fischbacher 2004). Using examples from everyday life, when people see someone do something kind to another person, they often go out of their way to reward them for their kindness voluntarily. The same goes for violations of reciprocity; people punish those who did not return favors after receiving one. People even reward or punish others despite suffering costs to do so. These types of behaviors have been repeatedly demonstrated in lab experiments and in other social settings (Fehr, Fischbacher and Gächter 2002; Shinada, Yamagishi and Ohmura 2004) It is thus worth exploring whether this behavior also applies to international relations.

Most formal or laboratory studies also tend to focus on generic or highly abstract outcomes, like "cooperate" versus "defect", that make it difficult to determine when and why reciprocity promotes cooperation in the real-world. Because they use such highly stylized or abstract examples of reciprocity, they often suffer from a lack of external validity. There is thus room for experimental and observational approaches that study reciprocity in international settings, with real-world behaviors like agreeing to a joint military operation or investing in the development of another country's infrastructure or economic industries.

Prior reciprocity research also tends to focus on state-level analysis, and ends up ignoring the un-

derlying motivations for reciprocity at the individual level. While state-level behavior is obviously important, it is worth understanding how real human beings use fundamental concepts like reciprocity by testing micro-level explanations. Representative studies on reciprocity often focus on the interactions between major powers, or at most the “triangular” interactions of three actors (Goldstein 1991; Goldstein and Freeman 1990, 1991; Rajmaira and Ward 1990; Rhodes 1989; Ward and Rajmaira 1992; Ward 1982). These studies find that almost all actors use reciprocity at rates higher than expected, but they tell us very little about why states, or people, actually reciprocate. Rational approaches to reciprocity also do not adequately account for seemingly irrational reciprocity-related behaviors, such as indirect “upstream” reciprocity. This also known as “paying it forward” to someone else without expecting anything immediate in return, as opposed to behaviors towards a direct interaction partner in direct reciprocity (Gray, Ward and Norton 2014; Nowak and Sigmund 2005; Nowak and Roch 2007).

Research in international relations derived from social psychology draws attention to some of these issues and highlights the opportunity for more work that explores reciprocity at an individual level. Larson (1988), for instance, explicitly calls for more experimental work focusing on individual perceptions of reciprocity in international relations. Citing empirical evidence from psychology, Larson argues that individuals use reciprocity to discern the motives and intentions of other actors, which can lead to outcomes that are not easily explained by game theoretic approaches where the payoffs of a given scenario are assumed to make every actor’s intentions obvious. Since humans use subjective criteria such as first impressions or heuristics to infer intentions, Larson argues this can lead to varying interpretations of behavior from international actors, where real world acts are not always easy to discern as “cooperative” or “defective.”

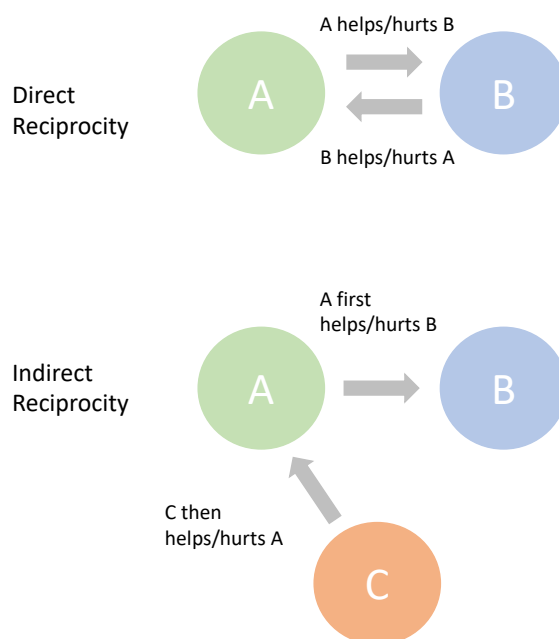
Focusing on the individual-level also offers the benefit of building on previous research on potential micro-level channels for how reciprocity affects state security decisions. First, public opinion has been shown under many conditions to be an important influence on elite and state foreign policy decision-making. Democratic leaders are sensitive to public sentiment when making decisions about international cooperation, and expect to pay domestic political costs for pursuing policies the public opposes (Aldrich et al. 2006; Reiter and Stam 2002; Tomz, Weeks and Yarhi-Milo 2020).

These incentives might especially constrain behavior when elite and public preferences diverge (Holsti 2004). They also matter even when they align; public opinion can reinforce leaders' existing incentives to act, or reflect a selection process in which leaders with preferences similar to the public's are more likely to attain office in the first place (Tomz and Weeks 2021). Second, recent research shows that while elites and the public often have very different attitudes and traits, the decision-making processes between elites and the public are oftentimes strikingly similar (Kertzer 2022; Yarhi-Milo, Kertzer and Renshon 2018). Focusing on individual support for cooperation is thus not just a convenient way to test causal claims; it is a theoretically motivated choice that connects micro-level reciprocity dynamics to the broader question of when and how public attitudes constrain state-level foreign policy behavior. While I am agnostic about how and when individual decision-making ultimately influences state-level decisions, further research beyond the scope of this study might investigate how these reciprocity-specific processes occur.

What is Indirect Reciprocity?

Reciprocity, or the returning of a behavior, has been studied for decades in international relations, and the logic is simple and intuitive. Countries do to other countries what other countries do to them, or what they expect to be done to them. However; as previously discussed, reciprocity is most often studied in its direct, dyadic form, where two countries directly return the acts of the other. Another common form in international relations, but one that is far less studied, is indirect reciprocity. Countries often find themselves playing the role of a third party, observing two other actors and deciding how to behave towards them both. This more accurately depicts “downstream” indirect reciprocity, or when a third party decides to reward or punish an actor based on how it behaved towards a second actor. Figure 2.1 visually lays out the logic of direct reciprocity, where two actors reciprocate a helpful or hurtful act, and also “downstream” indirect reciprocity, where a third party (C) returns a helpful or hurtful act to party A based on whether A helped or hurt party B. In this study, I focus on indirect reciprocity when a first or second party has violated reciprocity, and a third party has an opportunity to cooperate with this party or not. As noted in the introduction, countries often find themselves in the position to play this third party role.

Figure 2.1: Direct and Indirect Reciprocity



A Psychological Approach to Indirect Reciprocity in IR

My theoretical explanation of reciprocity in international relations is derived from the theory of Bounded Generalized Reciprocity (BGR) in psychology (Yamagishi, Jin and Kiyonari 1999), which maps out the conditions under which people use reciprocity. This approach builds on prior research positing that humans have evolved to use reciprocity behaviors as a basis for making cooperative decisions, and we do so because reciprocity is one of the most efficient strategies for overcoming barriers to cooperation and ultimately human survival (Boyd and Richerson 1988, 1989; Fehr and Fischbacher 2004; Fehr, Fischbacher and Gächter 2002). Essentially, reciprocity functions as a sort of rule of thumb that guides our behavior. We also have developed psychological mechanisms that encourage the use of this rule of thumb. The logic goes that in the long run following the rules of reciprocity brings self-interest benefits from cooperation, so people have evolved in various ways that encourage this behavior. Using a simple example of direct reciprocity between two actors, the logic of BGR is simple and intuitive: returning a kind gesture with an equally kind gesture or

returning an offense with an equal offense generally leaves us better off. We are mentally wired to follow the unwritten logic that “what goes around, comes around”, and following the rules of reciprocity will be a net benefit over the course of our interactions with others.

According to this theory, the use of reciprocity is based on heuristics or automatic psychological processes as opposed to strictly rational evaluations of payoffs, and thus at the individual-level studies consistently find that people use reciprocity in ways and situations that would not be expected or explained sufficiently by formal or rational actor models. Research using BGR has demonstrated, for instance, that people are more likely to behave cooperatively (i.e., be kind, allocate more rewards) towards ingroup members than outgroup members or strangers, even when it is not welfare-improving to do so (Balliet, Wu and De Dreu 2014; Tajfel 1974; Yamagishi and Kiyonari 2000). It is worth noting again that this approach does not advocate for viewing irrational behaviors as non-strategic or idiosyncratic; reciprocity is a strategy for accruing cooperative benefits, and BGR makes the distinction that the micro-level processes behind reciprocity at the individual-level are not always fully calculated, but also potentially automatic or semi-deliberative.

But why do we use “downstream” reciprocity, or return favors or slights on behalf of others, even if we are not directly involved? According to this theoretical approach, we need it for survival. As humans, mutual cooperation in social and economic exchanges that involve large groups is essential, but difficult because of the ubiquity of material incentives to cheat our exchange partners (Shinada, Yamagishi and Ohmura 2004). Scholars argue that indirect reciprocity, or third parties rewarding (or punishing) good or (bad behavior), leads to almost universal cooperation in situations where selfish, egoist behavior would cause a cooperation breakdown and where direct reciprocity alone is insufficient to maintain cooperation (Balliet, Wu and De Dreu 2014; Boyd and Richerson 1988; Fehr and Fischbacher 2004; Romano, Balliet and Wu 2017). Consistent with the explanation that heuristical or other psychological processes drive reciprocity behavior, this research shows that people engage in third-party punishment or rewarding even when it is costly to themselves and there is no welfare-improving reward for doing so, with the point of this strategy being that eventually they will benefit. These studies find that there is an indirect reciprocity effect on behavior for both positive and negative actions, and across a wide variety of issue domains, but as of yet no research

investigates whether this approach also applies to international contexts.

Even if reciprocity is a strong strategy for cooperation and survival, people need mechanisms that encourage its use. Based on this theoretical approach, I describe four theoretical mechanisms for how indirect reciprocity might facilitate the effect of indirect reciprocity on individual support for cooperation with another country: reputational concerns, desires to correct bad behavior, fairness/morality concerns, and emotion.

The first mechanism is based on the idea that reputations drive reciprocity behavior in human beings (Romano, Balliet and Wu 2017; Yamagishi, Jin and Kiyonari 1999). This is the central mechanism explicitly laid out in BGR. Whether in direct reciprocity, by returning favors or slights, or indirect reciprocity, by punishing or rewarding others as a third party, people signal to other potential cooperative partners that they are trustworthy and reliable. We generally follow the rules of reciprocity, even when it does not directly involve us, and thus others can be sure that they can rely on us to return favors or cooperate without cheating or exploitation. Punishing or refusing to cooperate with others who have violated reciprocity shows commitment to cooperation.

In experimental and formal games, merely making individuals aware that their behavior is being monitored by others is enough to trigger reputational concerns and higher use of third party, indirect reciprocity (Brandt, Ohtsuki, Iwasa and Sigmund 2007; Romano, Balliet and Wu 2017; Wu, Balliet and Van Lange 2016). People are even willing to punish as third parties under these conditions when there is no prospect of future interactions with their observers, and are willing to incur costs to themselves in order to punish or reward others' behavior (Henrich, Boyd, Bowles, Camerer, Fehr, Gintis, McElreath et al. 2001). Put another way, when faced with cooperative dilemmas as a third party, people have a tendency to simplify the interaction by using reciprocity as a heuristic about reputation: rewarding those who practice reciprocity with cooperation improves our reputation, and so does punishing those who violate reciprocity.

A second mechanism taken from other psychological work on reciprocity is the desire to correct bad behavior from potential future partners. As a third party, we may not be dealing with the parties we are observing immediately or in the near future, but refusing to cooperate with or sanctioning

someone who violates “good” reciprocity might incentivize these parties and other observers to avoid violations in the future, and thus we are more likely to benefit from cooperation down the road. Romano, Balliet and Wu (2017) demonstrate this empirically in experimental economic games, showing that people punish as a third party when there is an expectancy of future interaction with the two other parties. Again, indirect reciprocity may function as a strategic heuristic, telling us that our cooperative behavior can help us down the line, albeit without knowing exactly when or how.

Concerns about morality and fairness might also induce us to use reciprocity; we often feel morally compelled to give an “eye for an eye” or “follow the golden rule”, thinking that it is the “right” or “fair” thing to do. Fehr, Fischbacher and Gächter (2002) find in their study on reciprocity that people cooperate or not based on how “fairly” they are being treated, and are willing to punish as a third party on behalf of others who are treated unfairly. While not explicitly writing within the framework of BGR, this mechanism may apply to reciprocity processes generally. This fairness motive is strong enough that people even show a willingness to sacrifice resources to do this; they are willing to pay costs to punish unfairness or accept that doing so will bring no material rewards. Prior work has also demonstrated how fairness and morality concerns can be rooted in a self-serving purpose consistent with the human goal of accruing cooperation benefits: doing the “right” thing or adhering to moral rules may lead others to give more than they otherwise would, for instance, or eliminate cheaters of reciprocity by outing them as immoral (Alexander 2017). People thus have a biological incentive to integrate considerations of fairness and morality into their use of reciprocity, and there are formal models that demonstrate how actors can incorporate fairness into their behavior at equilibrium when interacting with others (Rabin et al. 1992).

Finally, we may have automatic, emotional reactions that may trigger us to use reciprocity without full, conscious deliberation. We have all likely experienced an angry need to “get even” to someone who has harmed us first, or felt the satisfaction and joy that comes from “paying it forward” to someone else. Emotions have the ability to induce certain behaviors. Fehr and Fischbacher (2004) find in their study on reciprocity that in one-shot prisoner’s dilemma and dictator games, third party observers whose economic payoffs were unaffected by the games consistently punished those

who violated reciprocity, even if it was costly to themselves. They attribute these results to self-reported emotions and personal values, arguing that the results are consistent with an emotional motive and moral motive for the use of reciprocity. This conclusion stems from evolutionary game theory and psychology, where emotions like anger, hate, or joy have been demonstrated to benefit human survival and the accrual of self-interest benefits by inducing people to behave in ways that a conscious, selfish, rational actor would not (Frank 2011). Behavior based on emotion can theoretically lend credibility to commitments or threats that would otherwise not appear legitimate, like revenge in international relations (McDermott, Lopez and Hatemi 2017). In the case of indirect reciprocity, this might make a commitment to cooperation and returning favors credible and increase the chances of cooperation with others, even when it might not appear rational to do so. It also might make threats to punish as third parties more legitimate, and discourage cheaters from using exploitive behavior.

These are familiar concepts in international relations, where states and leaders commonly use the language of these mechanisms in justifying various foreign policy decisions. Countries are constantly concerned about their reputation with other countries, for instance, and often cite reputation as a motive for their actions (Jervis, Yarhi-Milo and Casler 2021; Weisiger and Yarhi-Milo 2015). Former Prime Minister of Great Britain Winston Churchill once famously said: “We do not covet anything from any nation, except their respect.”⁴

Countries also commonly abandon opportunities for cooperation in the name of correcting bad behavior of other states, like in 2010 when Canada broke off all diplomatic relations with North Korea following the sinking of a South Korean naval vessel, with a Canadian official stating: “The adoption of a controlled engagement policy and the imposition of special economic measures send a clear message to the North Korean government that its aggressive actions will not be tolerated.”⁵

Morality and fairness justifications for state actions are frequent, where countries often claim that doing the “right” thing guides foreign policy as much as interests do. Fittingly, former US president Barack Obama once declared during a speech at the United Nations: “I believe that America

⁴ “Vive la France! SLEEP TO GATHER STRENGTH FOR THE MORNING.” 19, October, 1940, Retrieved from *International Churchill Society*.

⁵ See: “Canada cuts ties with North Korea” *ABC News*, 28 October 2010.

is exceptional – in part because we have shown a willingness, through the sacrifice of blood and treasure, to stand up not only for our own narrow self-interest but for the interests of all.”⁶ Similarly, emotion is also easily observable in world affairs, like in instances of negative reciprocity, or taking an “eye for an eye”. In the aftermath of a terrorist attack on India in 2019, for instance, Indian Prime Minister Narendra Modi expressed a deep-seated emotional response: “The amount of anger that is there in the country, people’s blood is boiling. I can very well understand this. I want to tell the terror outfits and their patrons that they have committed a big mistake. They have to pay a heavy price for this.”⁷ Research on public opinion has also produced evidence that some of these mechanisms affect public support for foreign policies, like going to war. Tomz, Weeks and Yarhi-Milo (2020) find that reputational concerns and personal commitments to fairness mediated public support for going to war on behalf of military allies.

Based on this theoretical approach, I argue that indirect reciprocity may play an influential role on individuals whose countries take on the role of a “third party”, particularly on their attitudes towards security cooperation with other countries; people are supportive of cooperation with states that practice reciprocity and oppose cooperation with those that violate reciprocity. People do this because our brains have developed mechanisms, out of a necessity for human survival, that encourage us to use and enforce reciprocity whenever possible.⁸ Thus, my main hypothesis is as follows:

Hypothesis 1: *Respondents exposed to information about a reciprocity violation by a country towards another country will report lower support compared to those not exposed to a reciprocity violation.*

⁶ See: “Obama tells the world: America is exceptional” *USA Today*, 24 September 2013.

⁷ See: “India’s blood is boiling, says Modi” *Telegraph India*, 15 February 2019.

⁸ It is also worth noting that since this is to my knowledge one of the first studies to apply this approach to reciprocity to international relations, it is not theoretically clear which, or what combination of, these mechanisms is most likely to drive an indirect reciprocity effect in international cooperation contexts. While I do not hold any a priori expectations about these mechanisms, there is a clear opportunity for future research to both improve the specificity of this theory and explore the empirical question of what mechanisms drive the use of reciprocity in international relations contexts at the micro-level.

Research Design

To test my hypothesis, I employ a novel survey experiment measuring individual-level support for cooperation with another country in a hypothetical scenario. Experiments based on the public have several limitations, such as differences between ordinary citizens and leaders/state decision-makers or incoherent attitudes on foreign policy (Baum and Potter 2008; Hyde 2015). However, as an initial test of this micro-level approach to reciprocity, experiments offer comparative advantages of other methods for exploring the relationship between the main variables of interest: indirect reciprocity and public support for cooperation. By manipulating precisely individual exposure to reciprocity, and measuring individual support for cooperation, I can easily observe a causal effect from one onto the other and thus cleanly test whether reciprocity works as theorized.

Experiment and Treatments

The survey was conducted in the summer of 2024 on a nationally representative sample of 1000 US adults, recruited through the survey company Prolific.⁹ Before treatment, I collected participant information on several demographic and attitudinal variables (gender, age, education, party identification, political orientation, foreign policy orientation,) that might interact with indirect reciprocity or directly affect support for cooperation.¹⁰ Participants were then randomly assigned to one of four conditions in a 2×2 between-subjects design. In the first treatment dimension, participants were either assigned to a cooperative scenario on security or trade. In the second treatment dimension, participants were either assigned to a reciprocity violation condition or a control condition.

Participants were then instructed to read about a situation the United States could face in the future, and given a vignette describing a cooperation scenario between the US and another fictional country, called “Vexa”.¹¹ Stimuli for the treatments was delivered through these vignettes.

⁹ An a priori power analysis, based on results from a pilot study of 415 respondents, determined that a sample size of at least 320 would be sufficient to detect an effect at the standard power benchmark of .80.

¹⁰ Foreign policy orientation questions were taken from (Rathbun, Kertzer, Reifler, Goren and Scotto 2016).

¹¹ This name was generated using ChatGPT when prompted to create a fictional name for a country that would not

Participants then read the vignettes about the cooperative scenario: The country Vexa asks the United States for a cooperation deal to help deal with problems they are facing, which do not involve the US. All vignettes contained this basic premise, and were designed to be as identical as possible barring details about the specific cooperative issue the participants were assigned to. In the security condition group, participants read about Vexa asking the United States for a deal to share security records after suffering terrorist attacks in the security condition, and in the trade scenario, Vexa asks the United States for a deal to increase imports from Vexa in the US market after suffering economic problems. These vignettes were based on real examples of “third party” international cooperation opportunities, as a way to increase mundane realism and present cooperative decisions in a similar way to how people would encounter the concept in real life.¹² The vignettes were carefully workshopped with the goal of maximizing external validity and real scenarios that respondents can find plausible, while attempting to minimize the presence of internal confounders. “Vexa”, the country asking for cooperation with the United States in these scenarios, for instance, was given this generic name to avoid confounding by prior attitudes on real countries. Another obvious motive that might confound attitudes towards cooperation would be the benefits that the country or its people could get from it, or the material costs that go with doing another country a favor, so in the vignettes participants are told that the United States are not expected to gain or lose anything from the deal. For simplicity, the following examples use the security scenario, but respondents were randomly assigned to read either the trade or the security vignettes. Stimuli are provided in the appendix and on Open Science Framework (OSF), where the study was preregistered. The security scenario given to participants appears as follows:

Vexa is a country far from the United States, and does not interact with the US very often. Vexa has faced several costly terrorist attacks lately. These attacks do not affect the US or involve the US directly. The government of Vexa is taking steps to make its

introduce priors about real countries, and was chosen after reviewing potential candidate names with colleagues.

¹² For the security scenario, see: “Iran condemns ‘terrorist’ attack on gas pipelines.” *Al Jazeera*, 14 February, 2024. “USA seeks bilateral deals for access to European “criminal, terrorist, and identity records”.” *Statewatch*, 14 November, 2023.

For the trade scenarios, see: “U.S.-Cambodian Textile Agreement Links Increasing Trade with Improving Workers’ Rights” *Office of the United States Trade Representative*, 1 January, 2002.

“The political economy of tourism in Cambodia”.” *Asia Pacific Journal of Tourism Research*, 2008.

borders safer. Here are some important points about the situation:

Terrorist groups from nearby areas managed to get into Vexa and carry out attacks on important places like power stations and gas pipelines. Even though Vexa’s government caught the terrorists responsible, they’re still working to keep the borders safe from more attacks.

Vexa has asked the United States for a security agreement. Vexa wants the US government to share records on the identities of suspected terrorists. If the US agrees to share these records, security officials in Vexa will be able to check the fingerprints of people coming into Vexa against records from the US. If security officials find a match, they could use this information to identify suspected terrorists trying to enter the country. From the US perspective, sharing this information with Vexa is not expected to help or hurt the United States.

The vignette then either ends with no further information (*control*), or provides additional information about Vexa violating reciprocity with a separate country generically named “Skylis” on an issue that does not directly involve or affect the United States (*reciprocity violation*). The reciprocity violation treatment is shown below:

Last year, Vexa made a similar request to another country, Skylis, and received a database of terrorist identity records in response. However, when Skylis requested that Vexa reciprocate by sharing its own terrorist identity records with Skylis, Vexa denied this request and refused to share its records.

The experimental conditions allow for an assessment of the direct effect of indirect reciprocity on public support for cooperation proposed by Hypothesis 1, as well as exploring whether this effect is conditional on type of international cooperation, respectively. The design of the vignettes, and the precise measurement of public support for cooperation, are intended to address some of the common limitations of prior research on reciprocity in IR, such as highly simplified game-theoretic scenarios with “cooperate/defect” outcomes, or connections to real-world, substantive foreign policy issues.

Measures, Outcomes, and Empirical Strategy

After reading the vignette, participants responded to the focal outcome measure of support for cooperation with Vexa, as well as supplementary and manipulation check questions to assess whether respondents recognized and understood the content of our treatments. Participants were then debriefed at the end of the study about the purpose of the study and the nature of the manipulations.

The main outcome of interest, public support for cooperation, was operationalized as a participant's individual support for cooperation with Vexa, and was measured using an 11-point Likert scale where participants rated how much they support or oppose the proposed agreement with Vexa. The scale ranged from 0 (strongly oppose) to 10 (strongly support), and was based on previous research on public support (Tomz, Weeks and Yarhi-Milo 2020). Because participants are asked about support for cooperation with Vexa on one of two different issues, security or trade, I also create a combined scale measuring support for cooperation with Vexa generally.

For supplementary analysis purposes, I also asked respondents questions that I could use to test whether the theoretical mechanisms specified by the BGR approach mediate the effect of indirect reciprocity on public support for cooperation.¹³ For reputation, I ask respondents to what extent they agree that if the U.S. does not agree to the deal with Vexa, other countries will doubt America's willingness to cooperate on issues in the future. For correcting bad behavior, participants rate how much they agree with the statement that "The U.S. should use its cooperation to discourage other countries from breaking promises in the future." Moral and fairness concerns were measured by asking respondents to rate how much they thought the US had a moral obligation to act regarding Vexa or whether Vexa "played fair" with other countries. Emotions are notoriously difficult to measure in a survey setting, but I include feeling thermometers (0-100 point scale, with 0 indicating "coldest" feelings and 100 being "warmest" towards Vexa) as a basic measure, since they are a common method of approximating positive and negative emotion in political science and other fields like psychology (Marcus, MacKuen, Wolak and Keele 2006).

I use multiple linear regression to estimate the effect of the reciprocity violation treatment on

¹³ Exact measurement procedures and survey instrumentation can be found in the appendix.

support for cooperation, and also account for any variable imbalances in the condition groups. Covariates were normalized to range from 0 to 1 for ease of interpretation and comparison.

For supplementary analysis about what mechanisms may drive this indirect reciprocity effect, I use mediation analysis (Imai, Keele and Yamamoto 2010) to estimate the Average Causal Mediation Effects (ACME) of each potential mechanism on the main outcome variable.¹⁴

Empirical Results

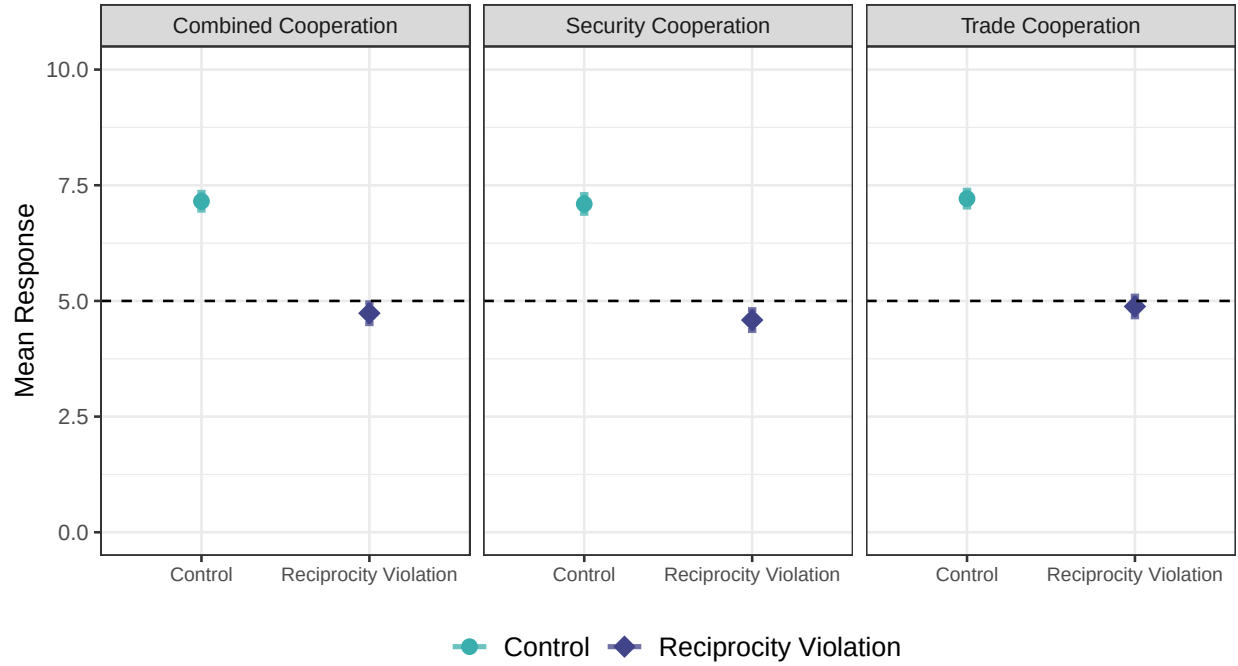
A simple visual representation of my findings by treatment group are presented below in Figures 4.1 and 2.3, and offer a preliminary picture of what the effect of indirect reciprocity might look like. On average, on the 0-10 point scale of support or opposition to cooperation with Vexa, exposure to a violation of reciprocity by Vexa dropped support from 7 to 5, or substantively the change of support for cooperation to neither support or opposition to cooperation. Looking at the distribution of responses for combined support further illustrates the indirect reciprocity effect; the number of strong supporters of cooperation drop significantly in the reciprocity violation condition group, while the number of “strongly oppose” responses goes up.

This is true across all issue conditions, and results from a t-test indicate that exposure to a violation of reciprocity works similarly across issue types like security or trade. Support for cooperation with Vexa was not significantly different in the security condition ($M=5.8$) compared to the trade condition ($M=6.05$), $t(994.23)=1.22$, $p\text{-value} = 0.22$.¹⁵ I interpret this as support for the idea that the treatment affects individual support for international cooperation in general, and thus in the following discussion of the results from regression analysis, the combined measure of public support for cooperation is used in the main model.

¹⁴ I use the Imai et. al. procedure for mediation, where the effect of some X variable on Y is mediated by variable M through three steps. First, the effect of X on Y is estimated using a regression model. Second, the effect of X on M is estimated using another regression model. Finally, the outcome Y is regressed on both X and M. To test for significance, the data is resampled many times to create a multitude of simulations, otherwise known as bootstrapping. If there is an observed effect across these simulations, we can be confident that the effect is not by chance. I follow this process for each survey item related to these potential mechanisms.

¹⁵ Results from additional manipulation checks indicate that respondents sufficiently understood the content of the scenarios and the treatments

Figure 2.2: Support for Cooperation by Treatment Group + 95% Confidence Intervals



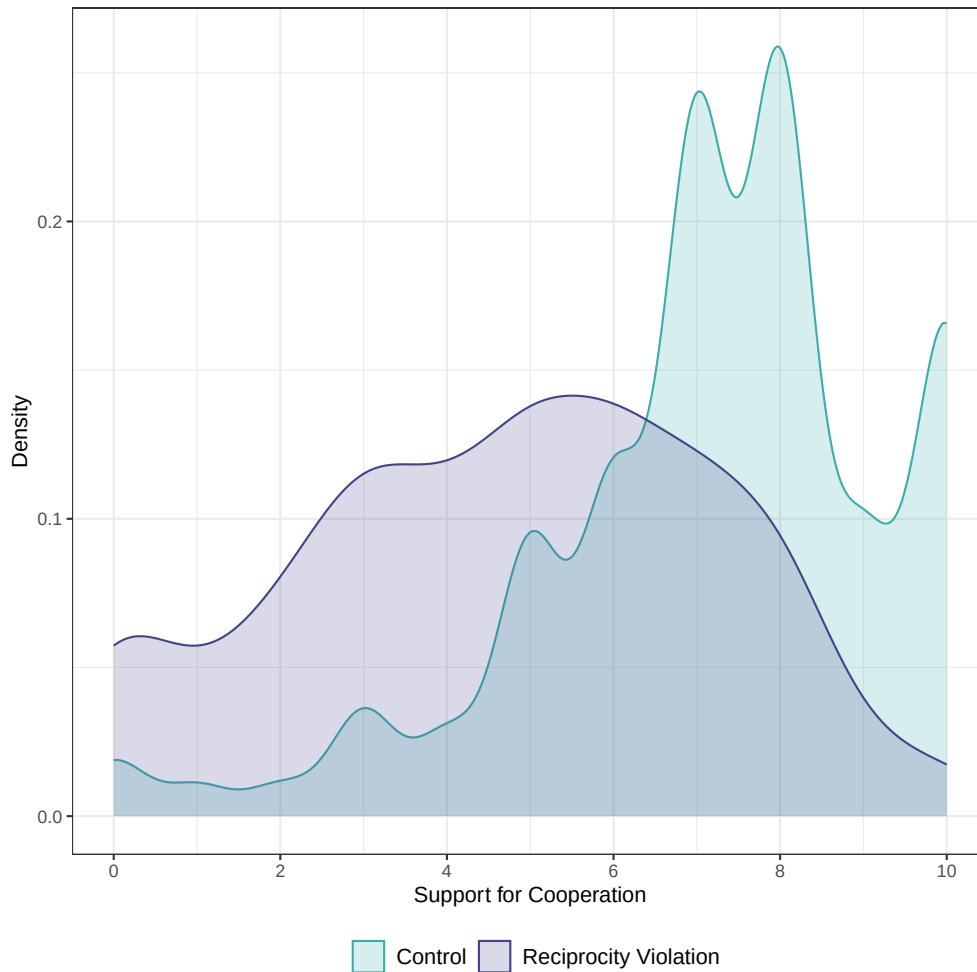
Note: The dashed line indicates the scale midpoint.

The Effect of Indirect Reciprocity on Public Support for Cooperation

Hypothesis 1 predicts that individuals will be less supportive of cooperation with a country when that country violates reciprocity, even if the individual's country was not directly involved with that interaction. In the survey experiment, this is simulated by the treatments: Vexa either violates reciprocity with Skylis by not returning the security or trade favor Skylis gave (*reciprocity violation*) or no further information is given (*control*). If “third party”, indirect reciprocity exerts an effect on public support for cooperation as predicted, participants in the reciprocity violation condition will be less supportive of cooperation with Vexa than those in the control group.

The findings strongly support Hypothesis 1. Respondents in the reciprocity violation condition were significantly less likely to support cooperation with Vexa than those in the control group. Figure 2.4 shows the estimated treatment effect on the three different measures of support for cooperation with Vexa, as well as estimates for the covariates used in these regression models. According to the primary model using the combined support outcome, on average, respondents

Figure 2.3: Response Distributions by Treatment Group



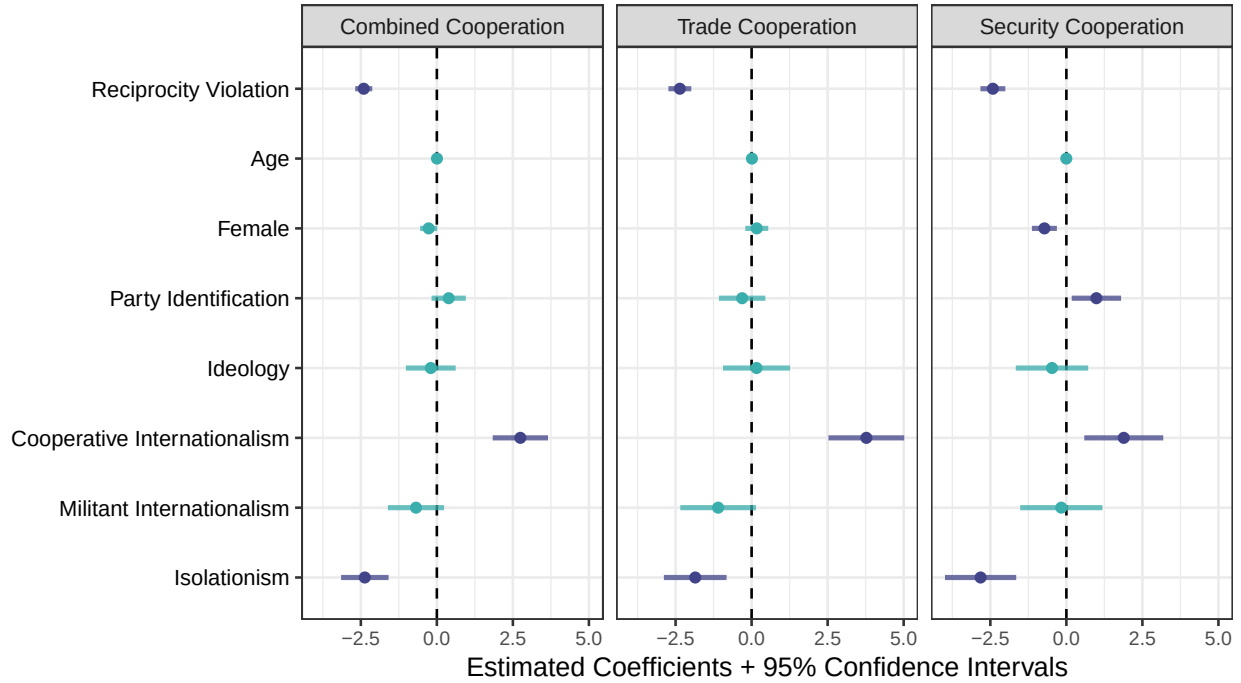
exposed to a reciprocity violation are estimated to be 2.4 points less supportive of cooperation with Vexa on the 11-point scale than those in the control condition. I interpret this as a substantial effect, and substantively is enough to move, on average, a respondent in the control condition group from supporting cooperation with Vexa (7.1) to a “neither supportive or opposed” or even slightly opposed position (4.7). This is a 34% decrease, or a drop of a third, in support for cooperation.¹⁶

A Supplementary Analysis of Mechanisms

While the results offer strong evidence that individuals use indirect reciprocity when making cooperative decisions, they do not tell us if people are doing so in the ways that BGR expects. Were

¹⁶ The effect of reciprocity remained significant and sizable in all model specifications.

Figure 2.4: Indirect Reciprocity Effect on Support for Coop



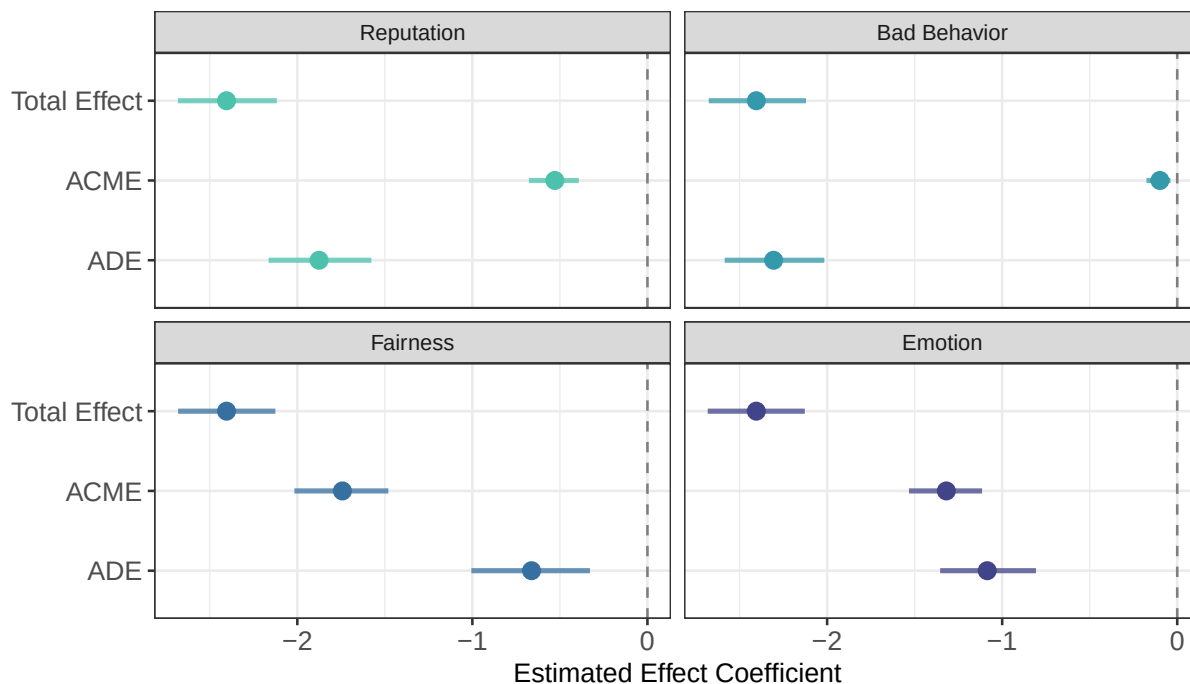
the participants less willing to cooperate with Vexa after a reciprocity violation because they used reciprocity as heuristic or mental shortcut to decision-making? In other words, the question of how reciprocity works and what drives it at the micro-level is still open-ended. This study's primary goal was to first establish that individuals use indirect reciprocity when thinking about international cooperation as a third party, and thus cannot fully unpack this question, but mediation analysis can at least offer compelling insights into how reciprocity works in international contexts.

This is an important question to explore for two reasons: first, understanding how exactly reciprocity works at the individual-level deepens our general understanding of one of the most important concepts in the study of international relations; and second, more specificity in the theorizing of important processes in international relations can lead to improved predictions and explanations. For example, if people use reciprocity out of an automatic emotional response as opposed to an exhaustive calculation, we might expect people to support the enforcement of reciprocity in some circumstances, even when it is strictly irrational to do so.

So why do people care about whether other countries violate reciprocity or not? As discussed in the theory section of this paper, there are four mechanisms that could potentially drive the effect

of indirect reciprocity: reputational concerns, correcting bad behavior, fairness/morality concerns, and emotions. Leveraging the survey questions related to these mechanisms, I statistically test whether any of these mechanisms significantly mediate the effect of indirect reciprocity on support for cooperation with Vexa. The results of the mediation analysis are presented below in figure 4.10.¹⁷ For each mechanism's corresponding panel in the figure, the Total Effect of indirect reciprocity on the main outcome (combined support for cooperation with Vexa) is shown. It also shows the Averaged Causal Mediated Effect (ACME), or how of this Total Effect was mediated by that specific mechanism. Finally, the Average Direct Effect (ADE) shows how much of the remaining Total Effect was not mediated by that mechanism. In terms of interpretation, the ACME and ADE coefficients are proportions of the Total Effect. In other words, as the ACME coefficient for a mediator gets closer to the Total Effect, the more that specific mediator is estimated to be responsible for that total effect. The farther away the ACME coefficient is from the Total Effect, the less that mechanism is responsible for the Total Effect of indirect reciprocity on support for cooperation with Vexa. The inverse is true of ADE.

Figure 2.5: Indirect Reciprocity Effect on Support for Cooperation with Vexa



¹⁷ Main models can be found in the Appendix, and all results were robust to specifications that included different covariates.

Two points are worth noting. First, all four mechanisms suggested by different psychological approaches to reciprocity partially and significantly mediate the indirect reciprocity effect. Second, the mechanisms linked to more automatic mental processes like fairness and emotion were estimated to mediate large portions of the total indirect reciprocity effect (0.73 and 0.55 respectively), while more explicitly strategic considerations like reputation and correcting bad behavior were responsible for much smaller portions of the effect (0.22 and 0.04 respectively).

Conclusions based on these results should be taken with some caution. There are obvious issues in trying to determine the exact mental process behind reciprocity without a research design that controls and manipulates more cognitive dimensions.

Nonetheless, these are fascinating findings, and lend support to the idea that individuals do likely use indirect reciprocity through these mechanisms as theorized, even in the realm of international affairs. People often act strategically, emotionally, and righteously in a myriad of social settings, and this may be true in international relations as well. These findings also likely mean that the mental processes that facilitate the use of reciprocity are very complicated, with multiple factors influencing this behavior simultaneously, and it is unclear to what extent these mechanisms interact with one another or complement/combat the others. Through the prism of every day life, it is easy to imagine how these mechanisms can operate at the same time when exposed to reciprocity; we might be thinking about our reputation with others when we punish someone who broke the rules, but we may also simultaneously feel moral outrage over someone not doing the “right thing”.

This study is not without limitations. First, as there are with all hypothetical experiments in international relations, there is the question of whether the research design accurately captures the most crucial aspects of the public opinion process. If people fundamentally think differently when using hypothetical scenarios with generic country names than when they are faced with these issues in the real world, the lessons from these types of studies are potentially flawed. While I tried to maximize the mundane realism of the experiment as much as possible, this is still a threat to the generalizability of the findings. Another limitation is the sparsity of the design. Due to resource constraints, many potential factors that might interact with an indirect reciprocity effect or that are relevant to public opinion about security decisions were not manipulated, like the power status

of the counterpart nation or the costs and benefits of the potential cooperation. Future research should explore these possibilities, and more rigorously test the boundary conditions of the theory in IR contexts.

Observational Evidence: Indirect Reciprocity and State-Level Cooperation

The survey experiment establishes a clear individual-level effect of indirect reciprocity on support for international cooperation. To assess whether this micro-level mechanism has observable correlates at the state level, I complement the experimental evidence with an observational analysis of directed country-pair interactions over time. I treat this analysis as a proof-of-concept test: while the experiment identifies a causal effect of indirect reciprocity on public attitudes, the observational study asks whether patterns consistent with indirect reciprocity are visible in actual state behavior. The two studies together span different levels of analysis, data sources, and methodological approaches, making convergent findings especially informative.

Data and Research Design

The observational study is based on a directed dyad-year panel dataset constructed from the Integrated Crisis Early Warning System (ICEWS) event data, which uses automated machine-coding to catalog interactions between countries over time (Boschee, Lautenschlager, O'Brien, Shellman, Starz and Ward 2015). Each observation represents the directed behavior of Country A toward Country B in a given year. The unit of analysis is the directed dyad-year for all country pairs from 1995–2013, yielding an analytic dataset of approximately 429,778 dyad-years due to missing values in control variables.¹⁸

¹⁸ The analytic sample covers 1995–2013 because some of the dyadic control variables, particularly bilateral trade and shared IGO memberships, do not cover beyond 2013. Dyad-years with no ICEWS events recorded were coded as zero on all dependent variables, representing true zeros (no recorded interactions) rather than missing data. As a robustness check, I created an indicator variable for observations that were previously recorded as NA's due to a lack of interactions, and use this indicator to verify that the main results are not driven by the zero imputation.

Dependent Variables

Using ICEWS quad codes, I construct four behavioral categories for events directed from Country A to Country B: Verbal Cooperation (Quad 1), Material Cooperation (Quad 2), Verbal Conflict (Quad 3), and Material Conflict (Quad 4). I also examine two aggregate outcomes: Total Cooperation (Quads 1+2) and Total Conflict (Quads 3+4). Each dependent variable is the count of events from Country A to Country B in a given year.

Key Independent Variables

The main explanatory variables are state-year level measures of Country B's lagged indirect reciprocity behavior toward all other countries excluding Country A. These are dichotomous indicators reflecting whether Country B reciprocated a cooperative or conflictual behavior with another country in the prior year, creating a one year lag. Separate indirect reciprocity measures are constructed for each quad code (q1–q4). Critically, Country A's interactions with Country B are excluded from these measures, and thus strong indicators of third-party indirect reciprocity behavior.

The one-year lag reflects a deliberate modeling choice, albeit a limiting one. It may take time for indirect reciprocity to be observed and affect outcomes, and I think it is a reasonable assumption to expect that a third-party observer's cooperative behavior in year t is shaped by what it observed about its partner's reciprocity behavior in year $t - 1$. Granted, there is no fixed optimal lag length, and the choice of one year is a limitation that future research should investigate more systematically using alternative lag structures.

Controls

All models include monadic controls for both the sender (A) and receiver (B): GDP, CINC (military capabilities), and democracy, along with their pairwise interactions to capture dyadic versions of these influences. I also include four time-varying dyadic controls drawn from the `peacesciencer` package (Miller 2022): lagged bilateral trade (log-transformed), a defense alliance dummy, UN

voting similarity (Tau-b), and shared IGO memberships, all lagged one year.

Modeling Strategy

Following a common empirical strategy for analyzing panel data (Imai and Kim 2021; Oneal and Russett 2001), I estimate four specifications for each dependent variable using the `fixest` package in R (Bergé, Butts and McDermott 2026): pooled OLS with heteroskedasticity-robust standard errors, dyad fixed effects, year fixed effects, and two-way fixed effects (TWFE) with standard errors clustered by both dyad and year. The two-way fixed effects models, which should theoretically absorb all time-invariant dyadic characteristics and time-specific global shocks while accounting for serial correlation within dyads, are treated as the primary specification. Full regression tables for all outcomes are presented in the Appendix.

Results

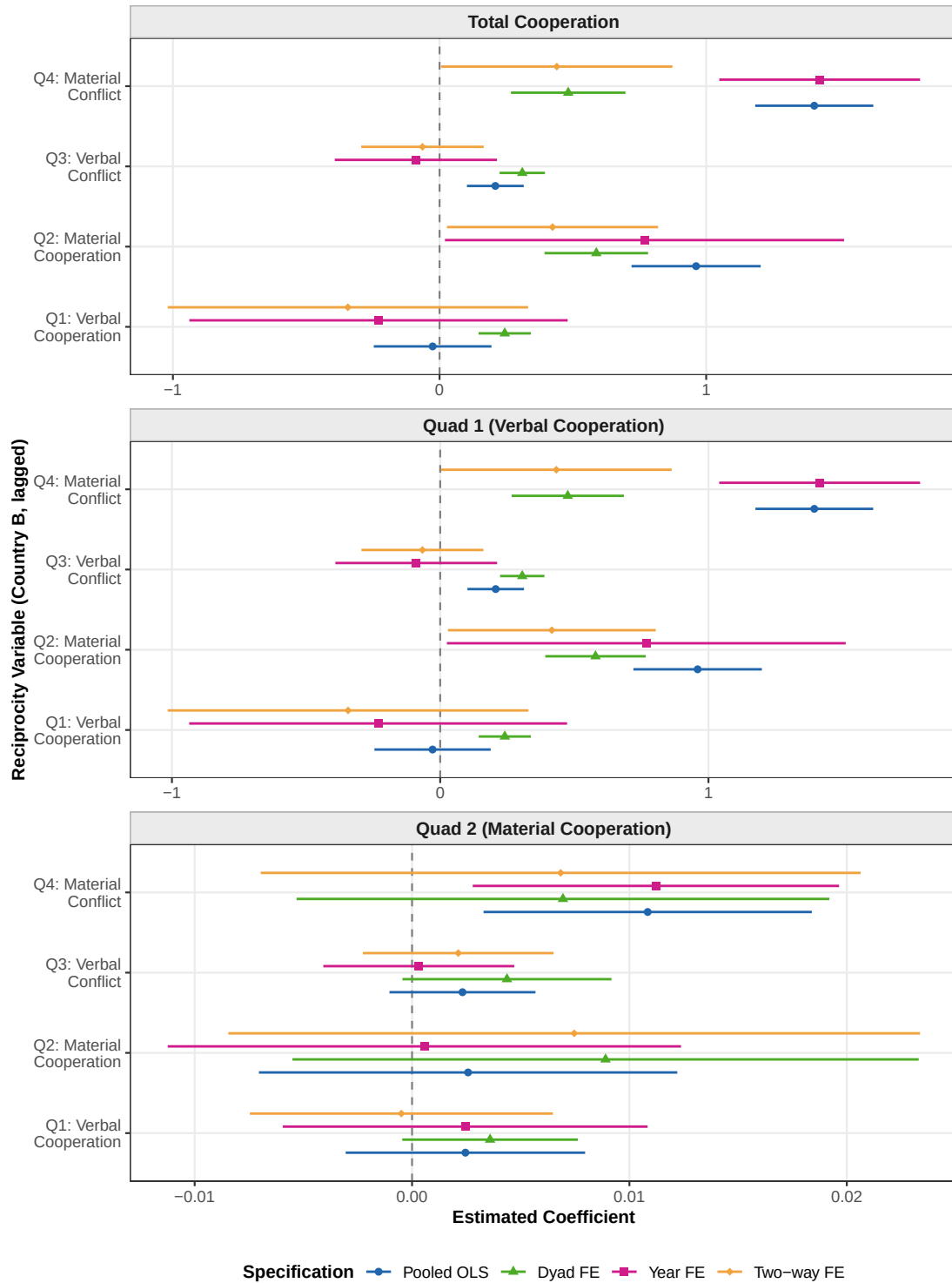
Figure 2.6 presents coefficient plots for the indirect reciprocity variables across cooperation-related outcomes, showing estimates from all four model specifications. The two-way fixed effects results, which provide the most conservative test, are summarized here.

Across cooperation outcomes, there is mixed but substantively interesting evidence for an indirect reciprocity effect at the state level. The clearest finding is for material indirect reciprocity: Country B's lagged material cooperative reciprocity (Quad 2) is positively and significantly associated with subsequent cooperative behavior from Country A in two-way fixed effects models, for both total cooperation ($\hat{\beta} = 0.42$, $p = 0.037$) and verbal cooperation ($\hat{\beta} = 0.41$, $p = 0.034$). Country B's material conflict reciprocity (Quad 4) is similarly positive and significant ($\hat{\beta} = 0.44$, $p = 0.047$). This finding is consistent with the interpretation that countries which reciprocate behaviors, regardless of whether cooperative or conflictual, signal reliability and thus attract more cooperative engagement from third parties. Verbal cooperative reciprocity (Quad 1) and verbal conflict reciprocity (Quad 3) are not significantly associated with subsequent cooperation in the TWFE model.

Figure 2.6: Effect of Indirect Reciprocity on Cooperative Behavior

Effect of Indirect Reciprocity on Cooperative Behavior

Point estimates with 95% confidence intervals



Two-way FE models include dyad and year fixed effects with two-way clustered SEs.
 All models include monadic controls (GDP, CINC, Democracy for A and B) and dyadic controls (bilateral trade, defense alliance, UNGA ideal point similarity, shared IGO memberships), all lagged one year.

For conflict outcomes, there is little evidence of a significant indirect reciprocity effect in the TWFE models, suggesting that different considerations and processes are at play when deciding when to behave adversarially towards others. These conflict results are also presented in the Appendix.

These results offer a state-level picture of a similar pattern to the one I explore in the survey experiment: third parties are attentive to whether other countries return favors or not. While the pathways for how individual-level reciprocity may have affected these decisions goes beyond the scope of this paper, this study indicates that as an initial test there is room for further research on why indirect reciprocity behavior by states follows similar patterns to reciprocity behavior by people.

Also notable is the finding that material behaviors (military cooperation, economic exchange) are costlier and therefore more credible signals of a country’s willingness to reciprocate is also noteworthy; it does not appear that “cheap talk” in this context independently influences cooperative behavior (Fearon 1997a).

These findings should be interpreted with appropriate caution. The observational design cannot establish causation; the association between Country B’s past reciprocity and Country A’s subsequent cooperative behavior could reflect unmeasured confounders, even with the inclusion of dyad and year fixed effects. The analytic window (1995-2013) is also constrained by data availability for the dyadic controls, and the results may not generalize to earlier or later periods. Furthermore, the measurement of indirect reciprocity and the outcomes linked to it could be made more precise in the future. Nevertheless, the direction and pattern of the observational findings are consistent with the experimental evidence, and together they support the plausibility of indirect reciprocity as a meaningful factor in international cooperation.

Conclusion

There are several substantive takeaways from the results of these studies. The first and most important is that both people and states *do* pay attention to and care about reciprocity, even when it does not *directly* involve them. This is a substantial finding because indirect, or third-

party, reciprocity to this point has gone largely unstudied despite having a potentially considerable impact on important cooperative decisions. One assumption when focusing on direct reciprocity, particularly when using traditional tit-for-tat and formal approaches, for instance, is that the parties by necessity must have “stakes in the game” or they must potentially benefit in some tangible way if they are to use reciprocity, since the goal is welfare improvement. The BGR approach alternatively explains why people may want their countries to get involved despite the lack of a foreseeable payoff, or to refuse seemingly beneficial cooperation: “paying it forward” or refusing to help a reciprocity violator may have long term, or “downstream” benefits.

Second, evidence from this study indicates that the BGR explanation of reciprocity is a potentially good fit for explaining micro-level behaviors in international relations. Reciprocity is a fundamental behavior in human societies, and is found in virtually every culture around the globe. It is considered a building block for cooperation, and theoretically should be adept at explaining reciprocity behaviors across a host of issue domains. Indirect reciprocity exerting an effect as expected across both security and trade issues, for instance, bolsters the notion that this account of reciprocity is well-suited for explaining behavior in complex strategic environments like in foreign affairs.

Relatedly, the findings from the survey experiment are consistent with the theoretical expectation that people use reciprocity in a way that they are familiar with in their every day lives. Reciprocity simplifies complex situations with a simple behavioral guideline: follow the “golden rule”, and punish or avoid those who do not. The mediation analysis, while not a comprehensive exploration of the mental processes that drive reciprocity, provides some confirmatory evidence that reciprocity serves as a short-hand heuristic. Mechanisms associated with automatic or semi-deliberative mental processes, like emotions or concerns about fairness, substantially mediate the indirect reciprocity effect. People care about doing the “right thing” by returning favors, and have powerful emotional responses when other countries do not. These feelings and intuitions do not make us any less strategic; on the contrary, the evolutionary foundations of BGR point to the idea that using reciprocity as heuristic tool is one of our strongest cooperative strategies. Thus, the theory offers a strong explanation for why respondents exposed to the reciprocity violation still supported cooperation less than those in the control condition, despite there explicitly being no costs or benefits for the

United States.

This approach to reciprocity in international relations is intuitive and easy to understand, and can potentially bridge an important gap between the analytical utility of rational approaches to international relations that focus on strategic gain and the plethora of evidence and perspectives showing that actors often act irrationally or in ways that do not strictly improve their welfare. Reciprocity is strategic; our goal is to accrue cooperation benefits and improve our chances of survival, but people have developed psychological mechanisms beyond exhaustive calculation that facilitate this end. Using reciprocity as a heuristic might be more efficient in some instances, and under some conditions lead to unexpected behavior, but other times may lead us to the same behaviors as predicted by formal or rational models.

Finally, the results from both the survey and the state-level analysis imply that failing to return favors is more costly than some countries might expect. Violating reciprocity may hurt your chances of cooperation with other, unrelated parties too, and finding forgiveness from either a country's leaders or the public for these transgressions might be difficult. If third party punishment or refusal to cooperate with reciprocity violators at the individual-level is based on fairness or emotion or reputations, for instance, countries may still not cooperate with violators even if the violators believe they had strong justifications or extenuating circumstances that would excuse their behavior.

This ties into the research on how the public is important to a country's foreign policy choices. Leaders may have incentives to either do what the public wants or avoid what they do not want even if their own preferences differ from the public's, or they may even think and behave like the public themselves (Aldrich et al. 2006; Chiba, Johnson and Leeds 2015; Gartzke and Gleditsch 2004; Kertzer 2022; Tomz, Weeks and Yarhi-Milo 2020). Thus, knowing that citizens care about whether other countries follow the "golden rule" may help scholars predict when and how cooperation occurs at the state-level too.

Appendix

Instrumentation

Vignettes

Participants received the following vignettes based on their assignment to either the trade or security scenarios.

Trade Scenario

Vexa is a country far from the United States, and does not interact with the US very often. Vexa is facing economic problems because of disruptions to regional supply chains. These problems do not affect the US or involve the US directly. The government of Vexa is trying to fix its economic problems. Here are some important points about the situation:

Recently, a regional shortage of key parts has stopped factories in Vexa from making small-scale machinery, causing big financial losses. Because industries in Vexa have not been able to find new suppliers, the government of Vexa is trying to help the country's economy in other ways.

Vexa has asked the United States for a trade agreement. Vexa wants the US to allow more rubber to be imported into the US from Vexa. Rubber is one of Vexa's major exports. If the United States agrees, Vexa hopes exporting more rubber will help get their economy back in shape. From the US perspective, the trade agreement with Vexa is not expected to help or hurt the United States.

Security Scenario

Vexa is a country far from the United States, and does not interact with the US very often. Vexa has faced several costly terrorist attacks lately. These attacks do not affect the US or involve the US directly. The government of Vexa is taking steps to make its borders safer. Here are some important points about the situation:

Terrorist groups from nearby areas managed to get into Vexa and carry out attacks on important places like power stations and gas pipelines. Even though Vexa's government caught the terrorists responsible, they're still working to keep the borders safe from more attacks.

Vexa has asked the United States for a security agreement. Vexa wants the US government to share records on the identities of suspected terrorists. If the US agrees to share these records, security officials in Vexa will be able to check the fingerprints of people coming into Vexa against records from the US. If security officials find a match, they could use this information to identify suspected terrorists trying to enter the country. From the US perspective, sharing this information with Vexa is not expected to help or hurt the United States.

Reciprocity Violation Treatments

Respondents were either given no information, or the following added paragraphs if they were assigned to the reciprocity violation treatment group.

Reciprocity Violation x Trade Scenario

Last year, Vexa made the same request to another country, Skylis. In that case, Skylis agreed to Vexa's request and allowed more products from Vexa to be sold within their country. However, when Skylis requested that Vexa reciprocate by allowing more goods from Skylis to be sold in Vexa, Vexa denied this request and refused to allow more Skylis goods to be sold in their country.

Reciprocity Violation x Security Scenario

Last year, Vexa made a similar request to another country, Skylis, and received a database of terrorist identity records in response. However, when Skylis requested that Vexa reciprocate by sharing its own terrorist identity records with Skylis, Vexa denied this request and refused to share its records.

Main Models

Mediation Models

The following are results from the mediation models described in the main paper.

Table 2.1: OLS Multiple Regression Models

	<i>Dependent Variable: Types of Support for Cooperation with Vexa</i>		
	Combined Support	Support for Security Deal	Support for Trade Deal
	(1)	(2)	(3)
Constant	7.09*** (0.54)	7.67*** (0.77)	6.02*** (0.72)
Reciprocity Violation	-2.40*** (0.14)	-2.42*** (0.21)	-2.36*** (0.19)
Security Condition	-0.21 (0.14)		
Age	0.004 (0.005)	-0.002 (0.01)	0.01 (0.01)
Female	-0.27 (0.14)	-0.72*** (0.21)	0.17 (0.19)
Party Identification	0.37 (0.29)	0.99* (0.41)	-0.31 (0.39)
Political Ideology	-0.19 (0.42)	-0.47 (0.61)	0.16 (0.56)
MI	-0.72 (0.47)	-0.17 (0.69)	-1.10 (0.63)
CI	2.71*** (0.46)	1.89** (0.66)	3.77*** (0.64)
Isolationism	-2.39*** (0.40)	-2.82*** (0.60)	-1.85*** (0.53)
Observations	1,000	498	502
R ²	0.28	0.29	0.32
Adjusted R ²	0.28	0.27	0.31
Residual Std. Error	2.26 (df = 990)	2.32 (df = 489)	2.14 (df = 493)
F Statistic	43.49*** (df = 9; 990)	24.47*** (df = 8; 489)	29.38*** (df = 8; 493)

Note:

*p<0.05; **p<0.01; ***p<0.001

Table 2.2: Reputation Mechanism

Effect	Estimate	CI.Lower	CI.Upper	p-value
ACME	-0.53	-0.68	-0.39	0.00
ADE	-1.88	-2.16	-1.58	0.00
Total Effect	-2.40	-2.68	-2.12	0.00
Prop. Mediated	0.22	0.16	0.29	0.00

Table 2.3: Bad Behavior Mechanism

Effect	Estimate	CI.Lower	CI.Upper	p-value
ACME	-0.10	-0.18	-0.04	0.00
ADE	-2.31	-2.59	-2.02	0.00
Total Effect	-2.40	-2.68	-2.12	0.00
Prop. Mediated	0.04	0.02	0.07	0.00

Table 2.4: Moral/Fairness Mechanism

Effect	Estimate	CI_Lower	CI_Upper	p-value
ACME	-1.74	-2.02	-1.48	0.00
ADE	-0.66	-1.01	-0.33	0.00
Total Effect	-2.40	-2.68	-2.13	0.00
Prop. Mediated	0.72	0.61	0.85	0.00

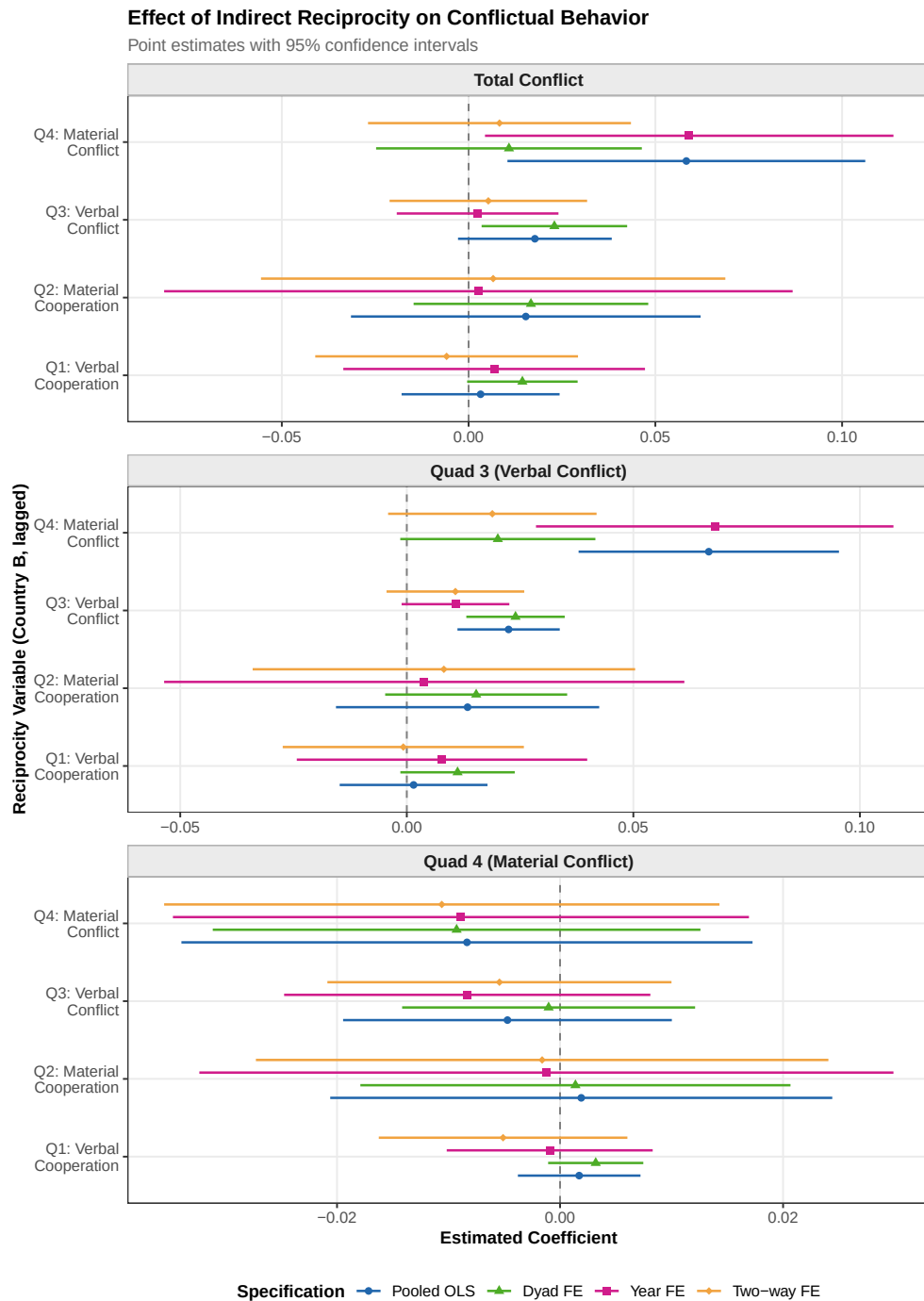
Table 2.5: Emotions Mechanism

Effect	Estimate	CI_Lower	CI_Upper	p-value
ACME	-1.32	-1.53	-1.11	0.00
ADE	-1.09	-1.35	-0.81	0.00
Total Effect	-2.40	-2.68	-2.13	0.00
Prop. Mediated	0.55	0.47	0.64	0.00

Appendix: Observational Study

Results

Figure 2.7: Effect of Indirect Reciprocity on Conflictual Behavior



Two-way FE models include dyad and year fixed effects with two-way clustered SEs.
All models include monadic controls (GDP, CINC, Democracy for A and B) and dyadic controls (bilateral trade, defense alliance, UNGA ideal point similarity, shared IGO memberships), all lagged one year.

Bibliography

- Aldrich, John H, Christopher Gelpi, Peter Feaver, Jason Reifler and Kristin Thompson Sharp. 2006. "Foreign policy and the electoral connection." *Annu. Rev. Polit. Sci.* 9:477–502.
- Alexander, Richard. 2017. *The biology of moral systems*. Routledge.
- Aron, Arthur, Elaine N Aron and Danny Smollan. 1992. "Inclusion of other in the self scale and the structure of interpersonal closeness." *Journal of personality and social psychology* 63(4):596.
- Axelrod, Robert. 1980. "Effective choice in the prisoner's dilemma." *Journal of conflict resolution* 24(1):3–25.
- Axelrod, Robert. 2006. *The evolution of cooperation*. Revised ed. New York: Basic Books.
- Balliet, Daniel, Junhui Wu and Carsten KW De Dreu. 2014. "Ingroup favoritism in cooperation: a meta-analysis." *Psychological bulletin* 140(6):1556.
- Baum, Matthew A and Philip BK Potter. 2008. "The relationships between mass media, public opinion, and foreign policy: Toward a theoretical synthesis." *Annu. Rev. Polit. Sci.* 11:39–65.
- Berg, Joyce, John Dickhaut and Kevin McCabe. 1995. "Trust, reciprocity, and social history." *Games and economic behavior* 10(1):122–142.
- Bergé, Laurent R., Kyle Butts and Grant McDermott. 2026. "Fast and user-friendly econometrics estimations: The R package fixest." *arXiv preprint arXiv:2601.21749*.
- Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and

- Michael Ward. 2015. "ICEWS Coded Event Data." .
URL: <https://doi.org/10.7910/DVN/28075>
- Boyd, Robert and Peter J Richerson. 1988. "The evolution of reciprocity in sizable groups." *Journal of theoretical Biology* 132(3):337–356.
- Boyd, Robert and Peter J Richerson. 1989. "The evolution of indirect reciprocity." *Social networks* 11(3):213–236.
- Brandt, Hannelore, Hisashi Ohtsuki, Yoh Iwasa and Karl Sigmund. 2007. A survey of indirect reciprocity. In *Mathematics for ecology and environmental sciences*. Springer pp. 21–49.
- Brewer, Marilynn B and Linnda R Caporael. 2013. An evolutionary perspective on social identity: Revisiting groups. In *Evolution and social psychology*. Psychology Press pp. 143–161.
- Cameron, James E. 2004. "A three-factor model of social identity." *Self and identity* 3(3):239–262.
- Caporael, Linnda R. 1997. "The evolution of truly social cognition: The core configurations model." *Personality and Social Psychology Review* 1(4):276–298.
- Carlin, Ryan E and Gregory J Love. 2013. "The politics of interpersonal trust and reciprocity: An experimental approach." *Political behavior* 35(1):43–63.
- Carson, Austin. 2016. "Facing off and saving face: covert intervention and escalation management in the Korean War." *International Organization* 70(1):103–131.
- Carson, Austin and Keren Yarhi-Milo. 2017. "Covert communication: The intelligibility and credibility of signaling in secret." *Security Studies* 26(1):124–156.
- Chiba, Daina, Jesse C Johnson and Brett Ashley Leeds. 2015. "Careful commitments: Democratic states and alliance design." *The Journal of Politics* 77(4):968–982.
- Crescenzi, Mark JC, Rebecca H Best and Bo Ram Kwon. 2010. Reciprocity in international studies. In *Oxford Research Encyclopedia of International Studies*.
- Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes." *American political science review* 88(3):577–592.

- Fearon, James D. 1997a. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1):68–90.
- Fearon, James D. 1997b. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1):68–90.
- Fearon, James D. 1998. "Bargaining, enforcement, and international cooperation." *International organization* 52(2):269–305.
- Fehr, Ernst and Simon Gächter. 2002. "Altruistic punishment in humans." *Nature* 415(6868):137–140.
- Fehr, Ernst and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and human behavior* 25(2):63–87.
- Fehr, Ernst, Urs Fischbacher and Simon Gächter. 2002. "Strong reciprocity, human cooperation, and the enforcement of social norms." *Human nature* 13:1–25.
- Frank, Robert H. 2011. "The strategic role of the emotions." *Emotion Review* 3(3):252–254.
- Gartzke, Erik. 1998. "Kant we all just get along? Opportunity, willingness, and the origins of the democratic peace." *American Journal of Political Science* pp. 1–27.
- Gartzke, Erik and Jon R Lindsay. 2015. "Weaving tangled webs: offense, defense, and deception in cyberspace." *Security Studies* 24(2):316–348.
- Gartzke, Erik and Kristian Skrede Gleditsch. 2004. "Why democracies may actually be less reliable allies." *American Journal of Political Science* 48(4):775–795.
- Gartzke, Erik and Kristian Skrede Gleditsch. 2006. "Identity and conflict: Ties that bind and differences that divide." *European Journal of International Relations* 12(1):53–87.
- Gibbs, David N. 1995. "Secrecy and international relations." *Journal of Peace Research* 32(2):213–228.
- Goldstein, Joshua S. 1991. "Reciprocity in superpower relations: An empirical analysis." *International Studies Quarterly* 35(2):195–209.

- Goldstein, Joshua S and John R Freeman. 1990. *Three-way street: Strategic reciprocity in world politics*. Chicago: University of Chicago Press.
- Goldstein, Joshua S and John R Freeman. 1991. "US-Soviet-Chinese relations: Routine, reciprocity, or rational expectations?" *American Political Science Review* 85(1):17–35.
- Goldstein, Joshua S and Jon C Pevehouse. 1997. "Reciprocity, bullying, and international cooperation: Time-series analysis of the Bosnia conflict." *American Political Science Review* 91(3):515–529.
- Gray, Kurt, Adrian F Ward and Michael I Norton. 2014. "Paying it forward: generalized reciprocity and the limits of generosity." *Journal of experimental psychology: General* 143(1):247.
- Grieco, Joseph M. 1988. "Anarchy and the limits of cooperation: a realist critique of the newest liberal institutionalism." *International organization* 42(3):485–507.
- Gummerum, Michaela, Masanori Takezawa and Monika Keller. 2009. "The influence of social category and reciprocity on adults' and children's altruistic behavior." *Evolutionary Psychology* 7(2):147470490900700212.
- Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath et al. 2001. "Cooperation, reciprocity and punishment in fifteen small-scale societies." *American Economic Review* 91(2):73–78.
- Holsti, Ole R. 2004. *Public opinion and American foreign policy*. University of Michigan Press.
- Huth, Paul K. 1988. "Extended deterrence and the outbreak of war." *American Political Science Review* 82(2):423–443.
- Huth, Paul K. 1999. "Deterrence and international conflict: Empirical findings and theoretical debates." *Annual Review of Political Science* 2(1):25–48.
- Hyde, Susan D. 2015. "Experiments in international relations: Lab, survey, and field." *Annual Review of Political Science* 18:403–424.
- Imai, Kosuke and In Song Kim. 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis* 29(3):405–415.

- Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1):51 – 71.
URL: <https://doi.org/10.1214/10-STS321>
- Jervis, Robert. 1976. *Perception and misperception in international politics: New edition*. Princeton University Press.
- Jervis, Robert, Keren Yarhi-Milo and Don Casler. 2021. "Redefining the debate over reputation and credibility in international security: Promises and Limits of New Scholarship." *World Politics* 73(1):167–203.
- Keohane, Robert O. 1986. "Reciprocity in international relations." *International organization* 40(1):1–27.
- Kertzer, Joshua D. 2022. "Re-assessing elite-public gaps in political behavior." *American Journal of Political Science* 66(3):539–553.
- Keysar, Boaz, Benjamin A Converse, Jiunwen Wang and Nicholas Epley. 2008. "Reciprocity is not give and take: Asymmetric reciprocity to positive and negative acts." *Psychological Science* 19(12):1280–1286.
- Kydd, Andrew. 1997. "Game theory and the spiral model." *World Politics* 49(3):371–400.
- Kydd, Andrew. 2000. "Trust, reassurance, and cooperation." *International Organization* 54(2):325–357.
- Kydd, Andrew H. 2007. *Trust and mistrust in international relations*. Princeton, NJ: Princeton University Press.
- Larson, Deborah Welch. 1988. "The psychology of reciprocity in international relations." *Negot. J.* 4:281.
- Leach, Colin Wayne, Martijn Van Zomeren, Sven Zebel, Michael LW Vliek, Sjoerd F Pennekamp, Bertjan Doosje, Jaap W Ouwerkerk and Russell Spears. 2008. "Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification." *Journal of personality and social psychology* 95(1):144.

- Leng, Russell J. 1993a. *Interstate Crisis Behavior, 1816-1980*. Cambridge University Press.
- Leng, Russell J. 1993b. "Reciprocating influence strategies in interstate crisis bargaining." *Journal of Conflict Resolution* 37(1):3-41.
- Lindskold, Svenn. 1978. "Trust development, the GRIT proposal, and the effects of conciliatory acts on conflict and cooperation." *Psychological Bulletin* 85(4):772.
- Marcus, George E, Michael MacKuen, Jennifer Wolak and Luke Keele. 2006. The measure and mismeasure of emotion. In *Feeling politics: Emotion in political information processing*. Springer pp. 31-45.
- McDermott, Rose, Anthony C Lopez and Peter K Hatemi. 2017. "'Blunt Not the Heart, Enrage It': The Psychology of Revenge and Deterrence (November 2017)." *Texas National Security Review*.
- Mellers, Barbara, Alan Schwartz and Ilana Ritov. 1999. "Emotion-based choice." *Journal of experimental psychology: General* 128(3):332.
- Miller, Steven V. 2022. "peacesciencer: An R Package for Quantitative Peace Science Research." *Conflict Management and Peace Science* 39(6):755-779.
- Nowak, Martin A and Karl Sigmund. 2005. "Evolution of indirect reciprocity." *Nature* 437(7063):1291-1298.
- Nowak, Martin A and Sébastien Roch. 2007. "Upstream reciprocity and the evolution of gratitude." *Proceedings of the royal society B: Biological Sciences* 274(1610):605-610.
- Oneal, John R. and Bruce Russett. 2001. "Clear and Clean: The Fixed Effects of the Liberal Peace." *International Organization* 55(2):469-485.
- Osgood, Charles E. 1962. "An alternative to war or surrender."
- Oskamp, Stuart. 1971. "Effects of programmed strategies on cooperation in the Prisoner's Dilemma and other mixed-motive games." *Journal of Conflict Resolution* 15(2):225-259.
- Perugini, Marco, Marcello Gallucci, Fabio Presaghi and Anna Paola Ercolani. 2003. "The personal norm of reciprocity." *European Journal of Personality* 17(4):251-283.

- Poznansky, Michael and Evan Perkoski. 2018. "Rethinking secrecy in cyberspace: the politics of voluntary attribution." *Journal of Global Security Studies* 3(4):402–416.
- Rabin, Matthew et al. 1992. *Incorporating fairness into game theory and economics*. University of California at Berkeley, Department of Economics.
- Rajmaira, Sheen and Michael D Ward. 1990. "Evolving foreign policy norms: Reciprocity in the superpower triad." *International Studies Quarterly* 34(4):457–475.
- Rathbun, Brian C. 2011. *Trust in international cooperation: International security institutions, domestic politics and American multilateralism*. Vol. 121 Cambridge University Press.
- Rathbun, Brian C, Christopher Sebastian Parker and Caleb Pomeroy. 2025. "Separate but unequal: Ethnocentrism and racialization explain the "democratic" peace in public opinion." *American Political Science Review* 119(2):621–636.
- Rathbun, Brian C, Joshua D Kertzer, Jason Reifler, Paul Goren and Thomas J Scotto. 2016. "Taking foreign policy personally: Personal values and foreign policy attitudes." *International Studies Quarterly* 60(1):124–137.
- Reiter, Dan and Allan C Stam. 2002. *Democracies at war*. Princeton University Press.
- Renshon, Jonathan and Ryan Powers. 2024. "Identity and the Social Construction of Reputation in World Politics."
- Rhodes, Carolyn. 1989. "Reciprocity in trade: The utility of a bargaining strategy." *International Organization* 43(2):273–299.
- Romano, Angelo, Daniel Balliet and Junhui Wu. 2017. "Unbounded indirect reciprocity: Is reputation-based cooperation bounded by group membership?" *Journal of Experimental Social Psychology* 71:59–67.
- Roskos-Ewoldsen, Andrew, Morgan Ellithorpe and Brandon J Kinne. 2026. "The Effect of Reciprocity on Trust: International Cooperation and COVID Aid." *Journal of Conflict Resolution* 70(1):115–139.

- Russett, Bruce M and John R Oneal. 2001. "Triangulating peace: Democracy, interdependence, and international organizations." (*No Title*) .
- Sanger, David E. 2012. *Confront and conceal: Obama's secret wars and surprising use of American power*. Crown.
- Schelling, Thomas C. 2008. Arms and influence. In *Strategic Studies*. Routledge pp. 96–114.
- Schroeder, David A, Julie E Steel, Andria J Woodell and Alicia F Bembenek. 2003. "Justice within social dilemmas." *Personality and Social Psychology Review* 7(4):374–387.
- Shinada, Mizuho, Toshio Yamagishi and Yu Ohmura. 2004. "False friends are worse than bitter enemies: "Altruistic" punishment of in-group members." *Evolution and Human Behavior* 25(6):379–393.
- Spence, Michael. 1978. Job market signaling. In *Uncertainty in economics*. Elsevier pp. 281–306.
- Tajfel, Henri. 1970. "Experiments in intergroup discrimination." *Scientific american* 223(5):96–103.
- Tajfel, Henri. 1974. "Social identity and intergroup behaviour." *Social science information* 13(2):65–93.
- Tomz, Michael and Jessica LP Weeks. 2021. "Military alliances and public support for war." *International Studies Quarterly* 65(3):811–824.
- Tomz, Michael, Jessica LP Weeks and Keren Yarhi-Milo. 2020. "Public opinion and decisions about military force in democracies." *International Organization* 74(1):119–143.
- Tomz, Michael R and Jessica LP Weeks. 2013. "Public opinion and the democratic peace." *American political science review* 107(4):849–865.
- Uslaner, Eric M. 2002. *The Moral Foundations of Trust*. Cambridge University Press.
- Ward, Michael D and Sheen Rajmaira. 1992. "Reciprocity and norms in US-Soviet foreign policy." *Journal of Conflict Resolution* 36(2):342–368.
- Ward, Michael Don. 1982. "Cooperation and conflict in foreign policy behavior: Reaction and memory." *International Studies Quarterly* 26(1):87–126.

- Weisiger, Alex and Keren Yarhi-Milo. 2015. "Revisiting reputation: How past actions matter in international politics." *International Organization* 69(2):473–495.
- Wilson, Warner. 1971. "Reciprocation and other techniques for inducing cooperation in the Prisoner's Dilemma game." *Journal of Conflict Resolution* 15(2):167–195.
- Wu, Junhui, Daniel Balliet and Paul AM Van Lange. 2016. "Reputation, gossip, and human cooperation." *Social and Personality Psychology Compass* 10(6):350–364.
- Yamagishi, Toshio, Nobuhito Jin and Toko Kiyonari. 1999. "Bounded generalized reciprocity: ingroup boasting and ingroup favoritism. Adv." *Group Process* 16:161–197.
- Yamagishi, Toshio and Toko Kiyonari. 2000. "The group as the container of generalized reciprocity." *Social psychology quarterly* pp. 116–132.
- Yarhi-Milo, Keren, Joshua D Kertzer and Jonathan Renshon. 2018. "Tying hands, sinking costs, and leader attributes." *Journal of Conflict Resolution* 62(10):2150–2179.
- Zheng, Wenqian, Shengli Cao, Yangyang Wang, Ke Yang, Yunyun Chen and Guoping Song. 2021. "The impact of social value orientation, game context and trust on cooperative behavior after cooperative video game play." *Psychological Reports* 124(3):1353–1369.

Chapter 3

Who Needs Friends Like These? Group Membership, Reciprocity, and Public Support for International Cooperation

Who Needs Friends Like These? Group Membership, Reciprocity, and Public Support for International Cooperation

Andrew Roskos-Ewoldsen

Department of Political Science, University of California, Davis

Abstract

Do people react more harshly to friends than strangers for failing to return a favor? Or are they more willing to forgive them for cooperative trespasses? This paper examines how group membership moderates the effect of reciprocity on public support for international cooperation. Drawing on the theory of Bounded Generalized Reciprocity (BGR), I argue that ingroup status raises the cooperative expectations individuals hold for a counterpart country, so that violations of reciprocity, and purely negative reciprocity, are penalized more severely when the offending country is an ingroup member than when it is an outgroup member or an unknown country. Outgroup countries that act non-cooperatively do not violate expectations and thus should not be punished more harshly or deemed less trustworthy than a country in a neutral control condition. I test these predictions using a novel survey experiment on a nationally representative sample of U.S. adults. Participants are randomly assigned to one of three groupness conditions (control, ingroup, outgroup) and one of three reciprocity conditions (positive reciprocity, violation of positive reciprocity, negative reciprocity), and report their support for an international cooperation agreement with and their trust in the fictional country Lurana. Contrary to my theoretical expectations, I find that when it comes to reciprocity, people have no grace for friends nor mercy for enemies. While groupness and reciprocity independently influence policy support for cooperation and trust, groupness does not significantly moderate the effect of reciprocity save for in the case of outgroups who use negative reciprocity; outgroups who reciprocate slights suffer the harshest cooperative penalties.

Introduction

When the United States imposed sweeping tariffs on Canada in early 2025, the dominant reaction among Canadians was not anger or fear, but betrayal. As Canada's Prime Minister Justin Trudeau announced its plans for its own retaliatory tariffs, he reminded Americans that Canadian troops had fought alongside them in Afghanistan and helped with wildfires in California.¹ The implicit message was clear: we did our part, and you did not do yours. The same logic animated American frustration when NATO allies did not immediately support US operations during the war with Iran: US President Donald Trump declared that "You'll have to start learning how to fight for yourself, the USA won't be there to help you anymore, just like you weren't there for us", and the Pentagon drafted plans to punish allies on the "naughty list" who were insufficiently supportive.² US remonstrations on this situation and NATO collective defense spending being "unfair" echoed sentiments from decades earlier, when France, a country many Americans considered a close partner with deep historical ties, refused to back the Iraq War. Americans infamously coined the term "freedom fries" to both replace the term "french fries" and voice their dissatisfaction, while reminding their French counterparts that the US had bailed out France in both World Wars.³ These events share a common thread: people value the act of giving and returning favors, they hold friendly countries to a higher standard, and react more sharply when those countries fall short. But is this generally true in world politics? And if so, why?

It is well established in international relations that reciprocity, or giving and returning behaviors, is a foundational component of international cooperation (Axelrod 1980; Crescenzi, Best and Kwon 2010; Goldstein and Pevehouse 1997; Keohane 1986). We also know that countries that are perceived as part of the same group are more likely to cooperate with and return favors to one another (Gartzke 1998; Russett and Oneal 2001; Tomz, Weeks and Yarhi-Milo 2020). Yet we have surprisingly few theoretical explanations and sparse empirical evidence about how real people interpret

¹ See: "Trudeau Details Canada's Retaliation Plans in Emotional Rebuke of Trump Tariffs" *The New York Times*, 5 February 2025.

² See: "Trump lashes out at Europe as growing number of allies reject US calls for help" *CNN*, 31 March 2026. "Trump weighs consequences for NATO allies on 'naughty' list" *Politico*, 22 April 2026.

³ See: "Americans Just Say 'Non' to French Products" *Fox News*, 19 February 2003.

friendly countries failing to hold up their end of a bargain, or even acting negatively towards each other. Understanding how reciprocity works at the *individual*-level in international contexts is still largely a mystery, despite the mountain of evidence of its ubiquity at the state-level (Larson 1988). This type of puzzle is particularly important considering the emergence of work demonstrating that public opinion is a meaningful influence on elite and state foreign policy decision-making (Aldrich et al. 2006; Chiba, Johnson and Leeds 2015; Tomz, Weeks and Yarhi-Milo 2020), and the decision-making processes of ordinary citizens and political elites are often strikingly similar despite differences in knowledge and experience (Kertzer 2022; Yarhi-Milo, Kertzer and Renshon 2018).

I argue that group membership moderates how reciprocity affects people’s attitudes towards international cooperation. Specifically, I expect that countries considered a part of the same ingroup are penalized more for violating reciprocity expectations, such as failing to return a favor. Conversely, I expect that countries perceived as members of an outgroup receive greater cooperative benefits when they behave unexpectedly: returning favors should increase support for cooperation relative to other countries. This is based on the theory of Bounded Generalized Reciprocity (BGR) (Yamagishi, Jin and Kiyonari 1999), which explains cooperation, ingroup favoritism, and reciprocal behavior at the individual level through a reputational mechanism. People have a strategic incentive to garner reputations as “good” reciprocators, because returning behaviors in kind leads to cooperative benefits. People also have higher reciprocity expectations of others in their ingroup, since these are the people you are most likely to interact with and potentially receive cooperative benefits from, or in the worst case, be taken advantage of by. Thus, the motives to maintain a strong reputation within your group, and policing the reciprocity behavior of others in your group, are mechanisms by which groupness can moderate reciprocity. For outgroups, because reciprocity expectations are already low, I expect that people will not penalize exhibitions of reciprocity violations and negative reciprocity more harshly than other countries.

I test these theoretical expectations by administering a novel survey experiment to a representative sample of the American public. The experiment follows a 3×3 between-subjects factorial design, crossing three groupness conditions (control, ingroup, and outgroup) with three reciprocity conditions (positive reciprocity, violation of positive reciprocity, and negative reciprocity) for a total of

nine experimental conditions. Participants are presented with a vignette describing a hypothetical international cooperation scenario involving a fictional country, Lurana, which requests an international cooperation agreement regarding a trade or security issue. Groupness is manipulated by providing participants with information about whether Lurana shares key characteristics with the United States, such as regime type and demographic similarity, making it an ingroup member, or differs from the United States on these dimensions, making it an outgroup, or by providing no additional information about Lurana in the control condition. Reciprocity is manipulated through information about a prior interaction between the United States and Lurana on a separate but related issue. In the positive reciprocity condition, the United States previously extended assistance to Lurana, and Lurana returned the favor when asked. In the violation condition, Lurana received the same assistance but refused to reciprocate when the United States made a similar request. In the negative reciprocity condition, the United States reduces cooperation with Lurana, and Lurana responded in kind. I measure two primary outcomes: individual support for a cooperation agreement with Lurana on an 11-point scale ranging from strongly oppose to strongly support, and trust in Lurana on a similar scale. I test my hypotheses using OLS regression, with the interaction between groupness and reciprocity as the primary research interest.

The results provide mixed support for my theoretical expectations, but nonetheless present an interesting pattern in reciprocity behavior in international relations contexts. Reciprocity and groupness both independently influence policy support for cooperation and trust in expected ways: violations and negative acts of reciprocity lower support and trust, ingroup identification increases support and trust while outgroup status lowers it. However, contrary to the central prediction of BGR, I find that groupness does not significantly moderate the effect of reciprocity on either outcome. Respondents generally did not favor or punish “bad” behavior by ingroup countries anymore than control or outgroup countries. The only exception is that outgroup countries who engaged in negative reciprocity faced steeper penalties than the other group conditions; a finding that cuts in the opposite direction of my theoretical expectations.

In the following sections, I first present an overview of relevant literature on reciprocity, group identity, and international cooperation. I then explain my theoretical approach to group-based

reciprocity and derive hypotheses from BGR. I follow this with a description of my research design, survey experiment, and analytical strategy. In the results section, I showcase and detail the main findings while evaluating my hypotheses. I then provide a discussion of these findings, and conclude with a review of the limitations and implications for the future study of reciprocity and group dynamics in international relations.

Literature Review

What determines whether countries cooperate with one another? Two bodies of literature speak directly to this question: research on reciprocity in international relations, and research on group membership and intergroup bias. This paper sits at the intersection of both, and I draw on a third body of work, namely Bounded Generalized Reciprocity, to connect them.

Reciprocity

Reciprocity is a foundational concept in international relations research on both cooperation and conflict (Goldstein and Pevehouse 1997; Keohane 1986). The basic logic is intuitive: countries do to other countries what other countries do to them, and this tit-for-tat dynamic is one of the most influential and well-observed mechanisms for sustaining cooperation under anarchy (Axelrod 1980). In both formal and experimental approaches, reciprocity successfully induces cooperation in prisoner's dilemma-style settings (Axelrod 2006; Oskamp 1971; Wilson 1971), and observational studies find that countries reciprocate both cooperative and conflictual acts at rates higher than chance (Goldstein and Freeman 1991; Rajmaira and Ward 1990; Ward and Rajmaira 1992). Reciprocity also underlies much of the logic of deterrence and costly signaling in security contexts, where returning a negative act credibly communicates resolve and willingness to retaliate (Fearon 1994; Leng 1993b; Schelling 2008). Yet despite this rich literature, most prior research on reciprocity focuses on direct dyadic interactions between two states, and largely treats reciprocity as a strategic, rational calculation.

This leaves two important gaps. First, it tells us relatively little about the individual-level motiva-

tions and mental processes that drive reciprocity behavior (Larson 1988). Research in other fields, such as social psychology, have repeatedly demonstrated that reciprocity in other social contexts follows a different psychological process than a strictly rational one, which can lead to unexpected behaviors and outcomes, like individuals who engage in altruistic punishment or cooperate in one-shot prisoner's dilemma games (Fehr and Gächter 2002; Fehr, Fischbacher and Gächter 2002). It is unclear whether these processes are similar for individuals in international contexts. Second, it says little about how other factors, like group membership, might condition reciprocity. If people apply the rules of reciprocity differently depending on whether the other party is a friend, an enemy, or a stranger, this may have significant implications for our expectations in how fundamental interactions between countries are interpreted and reacted to.

Groups

Research on group membership and identity in international relations has established that perceived groups shape cooperation and conflict in important ways. The democratic peace literature, for instance, demonstrates that shared regime type significantly reduces the likelihood of conflict and increases cooperative behavior between states (Russett and Oneal 2001). Beyond regime type, countries that share similar foreign policy preferences, cultural characteristics, or ethnic and linguistic ties are also more likely to cooperate (Gartzke 1998; Gartzke and Gleditsch 2006). At the individual level, Tomz and Weeks (2021, 2013) show that public support for military cooperation and avoiding war are significantly shaped by whether a counterpart country is an ally or share a democratic government, suggesting that ordinary citizens use group membership cues when forming foreign policy attitudes. Similarly, Rathbun, Parker and Pomeroy (2025) argue that perceived group membership in the form of ethnocentrism shapes public attitudes towards going to war with other countries; people are more reluctant to fight countries that are ethno-culturally similar to themselves.

These findings are consistent with a broader pattern documented across many social contexts: people cooperate more with, extend more trust to, and return favors more readily for members of their own group than for strangers or outgroup members (Balliet, Wu and De Dreu 2014; Tajfel

1974; Yamagishi, Jin and Kiyonari 1999). Much of the work on group effects has thus concentrated on an important and related question: where does group membership come from? In psychology, research on the Minimal Group Paradigm (MGP) and Social Identity Theory (SIT) have been foundational in explaining how and why groups form, and thus how intergroup discrimination effects like ingroup bias or outgroup derogation come to pass (Tajfel 1970, 1974). The SIT line of research argues that group membership and categorization comes from the fundamental psychological need to view oneself positively; groups are a story of self-esteem. In addition to wanting to view our own personal characteristics positively, we also desire to maintain a positive *social* identity, and thus we categorize others around us into “us” and “them” so that we can compare ourselves favorably to the “them”. This is such a strong psychological motivation, people will even engage in this group categorization behavior when the groups are arbitrary or previously meaningless (e.g. fans of an abstract painter). Thus, perceived social status and self-esteem are important drivers to the perception of group membership and also its downstream effects. In international relations, this approach has been used to explain how certain group identities, like regime type, and the reputations associated with these identities are perceived by the public (Renshon and Powers 2024).

Another important explanation to the origins of group membership and group-based behaviors comes from evolutionary and social psychology (Brewer and Caporael 2013; Caporael 1997; Yamagishi and Kiyonari 2000), where in lieu of a story about self-esteem, groups are a story about survival. This line of research argues that humans are fundamentally social creatures where small groups, like dyads or work groups, are the primary context for human mental evolution. Small groups have almost always been the setting where cooperation has lead to future payoffs and better survival. Thus, humans have evolved to view social exchange by default within the context of groups, and will instinctively categorize and identify with groups when interacting with others.

Despite this evidence that groups and reciprocity affect international cooperation independently, the question of whether group and reciprocity effects interact with one another has received surprisingly little attention in international relations. In particular, while we know that friendly countries cooperate more when everything is going well, we know far less about what happens when a country considered “close” fails to return a favor, or acts against you. This is a significant gap

given how important and numerous these situations are in the real world, as the examples in the introduction illustrates.

Public Opinion, Trust, and International Cooperation

This paper examines two related but distinct outcomes: policy support for cooperation and trust. Both are theoretically important but distinct outcomes related to international cooperation at the state-level, and together offer a more complete picture of how people evaluate potential cooperation with a counterpart country.

A growing body of research demonstrates that public opinion is not merely a passive reflection of elite preferences but an active constraint on foreign policy decision-making, making individual-level policy support a theoretically meaningful outcome in its own right. Democratic leaders are sensitive to public opinion when making decisions about the use of military force, and expect to pay significant domestic political costs for pursuing policies the public opposes (Aldrich et al. 2006; Reiter and Stam 2002; Tomz, Weeks and Yarhi-Milo 2020). These incentives are particularly important when leaders' and the public's opinion about what to do are at odds (Holsti 2004). Even when elites and masses agree, public opinion may augment other incentives for leaders to act in line with public sentiment, or reflect a selection mechanism where leaders with shared preferences and attributes are chosen (Tomz and Weeks 2021). Policy support is thus the most direct way explore the conditions under which group and reciprocity effects in the public are likely to constrain cooperative foreign policy, and in the bigger picture, evaluate the implications of group-based reciprocity for sustained cooperation within and between groups of states.

Perceived trust is a well-established precondition for cooperation in international relations; countries need credible information about another country's intentions or reliability (Kydd 2000, 2007). Rathbun (2011) demonstrates that in order to build and sustain robust international institutions, states must trust in other states in a way that goes beyond "strategic" trust, or a belief in others' desire to sustain a beneficial relationship of exchange. In order to overcome the perceived risk of opportunism and free-riding in such large multilateral institutions, countries need to have a moral-

istic, “generalized” trust, or a belief in the character and attributes of an actor’s trustworthiness; they must believe that other countries “trust” others as a disposition. At the individual level, trust has been studied extensively as both a determinant and an outcome of cooperative behavior, from the investment game tradition in experimental economics (Berg, Dickhaut and McCabe 1995). It is also one of the most common outcomes measured for studying the relationship between cooperation and reciprocity across social contexts (Berg, Dickhaut and McCabe 1995; Carlin and Love 2013; Zheng, Cao, Wang, Yang, Chen and Song 2021). In both international and lab contexts (Lindskold 1978), Osgood (1962)’s development of the GRIT strategy and using repeated acts of reciprocity were successful in establishing trust among adversaries and preventing conflict. We still; however, lack a sufficient micro-level account of the relationship between reciprocity, trust, and cooperation in international relations.

Examining trust alongside policy support allows me to ask not only whether people are willing to cooperate with another country under specific conditions, but whether their underlying evaluation of that country as a trustworthy or reliable partner shifts as well. These two outcomes could theoretically diverge in interesting ways; as an example, perhaps group membership buffers countries from drops in trust when they violate reciprocity but does not shield them from a decrease in support for a specific cooperative agreement. Tracking both outcomes thus provides a rich test of how reciprocity and group identity interact in shaping public attitudes toward cooperation with other countries.

A Psychological Approach to Group-Based Reciprocity in IR

The theoretical framework I employ to link groups and reciprocity comes from the theory of Bounded Generalized Reciprocity (BGR) in psychology (Yamagishi, Jin and Kiyonari 1999; Yamagishi and Kiyonari 2000). This micro-level approach accounts for why group membership should matter for reciprocity judgments, and I generate specific predictions about asymmetries in the effect of reciprocity conditional on groupness in the context of international cooperation.

BGR begins from the observation that people rely on mental shortcuts or heuristics to guide

their cooperative decisions. Crucially, people pay close attention to and shape their expectations and behaviors of others based on reciprocity; as a rule of thumb we return favors and slights to others and punish or ostracize those who do not (even when these reciprocity failures do not involve us). The second key feature of BGR is that the reciprocity heuristic is context-sensitive. People do not use the reciprocity heuristic identically in every situation, and the strength and direction of reciprocity effects depend on the features of the social context that shape how people interpret another party's behavior. Specifically, reciprocity is conditioned by groupness: we form group identity and membership through categorization, and focus our reciprocity behaviors and expectations on those in our ingroup.

Both of these features are rooted in human evolution. Evolutionarily, we know that taking part in systems of direct and indirect reciprocity and relying on your group are efficient strategies for survival. Thus, we have developed several psychological processes that encourage us to identify with groups and apply the rules of reciprocity. BGR primarily focuses on a reputational mechanism, where reciprocity in groups is driven by a desire to maintain a reputation as a strong cooperator but also an enforcer of reciprocity norms, which should lead to receiving cooperative benefits from other group members over time. Other mechanisms also can facilitate the use of group-based reciprocity, and span from automatic (e.g. affect) to semi or fully deliberative (e.g. moral judgments) (Fehr and Fischbacher 2004; Fehr, Fischbacher and Gächter 2002; Romano, Balliet and Wu 2017). At the individual-level, this is why reciprocity often comes across to us as *felt*, instead of calculated. When someone does you a favor and you return it, or when someone wrongs you and you respond in kind, our motivations are often driven by a sense of what is right, what feels fair, or an intuition about how others see us. BGR formalizes this intuition and extends it to the domain of generalized reciprocity, where situations are characterized by social exchange, interdependent outcomes, opportunities for both direct and indirect reciprocity, and the ability to monitor and communicate about others behavior (Balliet, Wu and De Dreu 2014).

In international relations, I argue that reciprocity operates not just as a strategic logic at the state level, but as a psychological heuristic at the individual level: reciprocity is a cognitive shortcut that ordinary people use to make quick judgments about whether a foreign country deserves cooperation,

trust, and general goodwill. Crucially, I also argue that this heuristic is not applied uniformly across all countries. Instead, how people respond to reciprocity depends on who is doing the reciprocating.

Group Identity and Expectations

BGR holds that people extend higher baseline levels of trust and cooperation to ingroup members than to outgroup members or strangers (Yamagishi, Jin and Kiyonari 1999). This is not simply a matter of preferring people who are similar, or in the case of SIT, categorizing ourselves into group out a motive for self-esteem. Rather, BGR grounds ingroup favoritism in an instrumental logic rooted in the structure of social interaction: we interact more frequently with ingroup members, rely on them more heavily for cooperative gains, and assume that they share our norms, values, and goals (Brewer and Caporael 2013). Because the expected benefits of ingroup cooperation are higher, so too are the expected obligations.

These differences in expectations of ingroup members, and conversely, lower expectations of outgroup members is the theoretical driver of my moderation argument. In the context of reciprocity violations or negative reciprocity, behaviors like these are more than isolated cooperative failures, they threaten the entire system of benefits that come from belonging to a bounded group. Cooperative partners who take advantage of others, or free ride, by not following or enforcing reciprocity threaten the “generalized” trust that ingroup members have for one another that is essential for cooperation (Rathbun 2011; Uslaner 2002). This is the theorized reason for why we see evidence of mechanisms like reputational concerns, emotional reactions, moral judgment, and motivations to correct bad behavior driving reciprocity behavior (Fehr and Fischbacher 2004; Fehr, Fischbacher and Gächter 2002; Romano, Balliet and Wu 2017); these psychological processes facilitate the use of the reciprocity heuristic, which helps maintain the social benefits of a group. Furthermore, research in psychology consistently finds that people react more negatively to expectation-violating behavior than to behavior that is consistent with prior expectations, even when the material consequences are identical (Mellers, Schwartz and Ritov 1999). We often observe these conditional reciprocity behaviors in action: studies on third party punishment in economic games consistently find that people punish reciprocity violators from their own ingroup more harshly than those from an out-

group (Gummerum, Takezawa and Keller 2009; Shinada, Yamagishi and Ohmura 2004). However, this work has been conducted almost exclusively in laboratory settings with minimal groups, and has not been extended to international relations contexts where the stakes and the nature of group membership are considerably more complex. Applied to international cooperation, reciprocity violations and negative reciprocity by an ingroup country should produce a larger drop in cooperation support and trust than the same violation by a control or outgroup country, because only the ingroup violation is breaking with expectations.

The outgroup case follows a different logic entirely. Outgroup members are not assumed to share our norms or values, and we do not rely on them in the same way we rely on ingroup members. As a result, we do not extend the same baseline trust or cooperative expectations to outgroup countries, and thus are not surprised when they let us down. Because no expectation has been violated, BGR predicts that outgroup violations or negative behavior should not produce a stronger negative reaction than the same behavior by a control or ingroup country. There is even evidence that positive reciprocity behaviors by the outgroup can still lead to social rewards through the development of a trustworthy reputation (Romano, Balliet and Wu 2017). Whereas ingroup violations are experienced as betrayals, outgroup violations are experienced as confirmations of prior beliefs, and confirmations, however unpleasant, do not provoke the same type of negative response.

Hypotheses

These theoretical considerations generate a set of specific, testable predictions. I begin with the baseline prediction that reciprocity independently shapes policy support for cooperation and trust, regardless of groupness condition. This prediction follows directly from the BGR framework: when a country violates positive reciprocity (failing to return a favor) or uses negative reciprocity (returning a negative act), people should rate that country as less trustworthy, and should be less willing to cooperate with them than when they reciprocate positively (returns a favor). This baseline effect should hold across all groupness conditions, because the reciprocity heuristic is not uniquely tied to group membership. It operates generally, even if group membership conditions its strength.

Hypothesis 1: *Exposure to a reciprocity violation by Lurana will decrease individual support for cooperation with and trust in Lurana, relative to a positive reciprocity condition, regardless of groupness condition.*

Hypothesis 2: *Exposure to negative reciprocity by Lurana will decrease individual support for cooperation with and trust in Lurana, relative to a positive reciprocity condition, regardless of groupness condition.*

Before turning to the moderation hypotheses, it is also worth noting that I expect a significant, direct effect of group identity on cooperation support and trust, independent of reciprocity. While this goes beyond the scope of BGR, in line with prior research on group affiliation in international relations, I expect that holding all else constant people are more likely to support cooperation with and trust another country if that country is an ingroup, compared to a control country. I expect this group effect to work in the opposite direction for outgroups: people should be less trusting and supportive of cooperation with countries who are outgroups than control countries.

Hypothesis 3: *Exposure to ingroup information about Lurana will increase individual support for cooperation with and trust in Lurana, relative to the control condition, regardless of reciprocity condition.*

Hypothesis 4: *Exposure to outgroup information about Lurana will decrease individual support for cooperation with and trust in Lurana, relative to the control condition, regardless of reciprocity condition.*

The asymmetry between ingroup and outgroup reciprocity expectations is the basis for my theoretical predictions on the moderation of the reciprocity effect by group identity. It is worth being precise about what this asymmetry does and does not imply. BGR does not predict that outgroup countries are immune to reciprocity effects altogether. A country, whether ingroup, outgroup, or unknown, that violates reciprocity or acts negatively should still be evaluated less favorably than one that reciprocates positively. I predict specifically that the *moderation* of reciprocity effects by group identity is asymmetric: stronger in the ingroup direction, null in the outgroup direction.

The ingroup effect is driven by expectation violation; the absence of an outgroup effect is driven by expectation confirmation.

Hypothesis 5: *The negative effect of a reciprocity violation (and negative reciprocity) on support for cooperation with and trust in Lurana will be significantly larger when Lurana is framed as an ingroup member, compared to the control condition.*

Hypothesis 6: *The negative effect of a reciprocity violation (and negative reciprocity) on support for cooperation with and trust in Lurana will not be significantly larger when Lurana is framed as an outgroup member, compared to the control condition.*

Research Design

To test the hypotheses derived above, I employ a novel survey experiment administered to a national, politically and demographically representative sample of 1722 American adults recruited through Prolific.⁴ Survey experiments are well-suited to this research question for several reasons. First, random assignment to treatment conditions allows me to isolate the causal effect of reciprocity and group identity on cooperation support and trust, holding constant all other features of the scenario. The survey format also allows me to reach a large, diverse sample while maintaining experimental control over the information participants receive. This research approach is thus ideal for exploring the micro-foundational processes that underlie the direct and moderated effect of reciprocity on cooperation.

The experiment follows a 3×3 between-subjects factorial design, crossing three groupness conditions (control, ingroup, and outgroup) with three reciprocity conditions (positive reciprocity, violation of positive reciprocity, and negative reciprocity) for a total of nine experimental conditions. Participants are randomly assigned to one of these nine conditions and receive a single vignette about the

⁴ The study was pre-registered on the Open Science Framework prior to data collection at https://osf.io/wm6c8/overview?view_only=06bcd39c4bc41f49559126fc34d37bc.

US and a fictional country, called Lurana to avoid introducing prior beliefs about real countries. Group and reciprocity conditions are combined and randomly assigned within the single vignette. The design allows me to estimate both the independent main effects of groupness and reciprocity on the two outcomes and the interaction between them, which is the primary quantity of interest for testing the BGR moderation hypotheses.

Before being assigned to an experimental condition, all participants completed a pre-treatment block measuring standard demographic characteristics including age, education, ethnicity, and ideology, with other demographic information like gender, political party identification, and country of birth provided by the platform Prolific separately. I also measured foreign policy orientations (FPOs) using the scale developed by Rathbun et al. (2016), which serve as covariates in the main regression models.

Participants were instructed to read about Lurana, a fictional country created for the purposes of the study, and informed that they would be asked their opinions about a situation the United States could face in the future.

Issue Type

All participants then read a shared vignette preamble describing Lurana as a country far from the United States, and one that is requesting cooperation with the United States. Participants were randomly assigned to either a security or a trade version of the scenario. I alter the issue area of the vignette to assess whether these effects generalize across different domains of international cooperation. Issue type is not a formal dimension of the factorial design but rather serves as a built-in replication across cooperative domains, allowing me to examine whether the groupness and reciprocity effects are robust to the type of international agreement.

In the *security scenario*, Lurana has recently experienced a series of cyberattacks targeting its infrastructure and has asked the United States for a security agreement under which the US government would share technical information about how to detect and prevent cyberattacks. In the *trade scenario*, Lurana is experiencing economic challenges in its manufacturing sector and has asked

the United States for a trade agreement under which the US would allow more clothing products from Lurana to be sold in the American market. In both scenarios, participants are told that the agreement is not expected to help or hurt the United States directly, which holds constant the material stakes of cooperation across conditions and ensures that treatment effects on cooperation support reflect responses to groupness and reciprocity rather than to perceived national interest. The reciprocity and groupness treatments are identical in structure across both issue types, with the content of the prior interaction adapted to fit the relevant domain.

Groupness Treatments

The groupness treatment was embedded in the opening of the vignette, immediately after the introduction of Lurana. In the *control condition*, no information about Lurana's relationship to the United States was provided beyond what was necessary to understand the cooperation request. In the *ingroup condition*, participants read:

Many people consider Lurana a part of the same group or team as the United States. Both the United States and Lurana share democratic governments, free-market economies, high levels of freedom of speech, and shared membership in many international organizations. Lurana is also ethnically and culturally similar to the US. Americans often say that the US and Lurana have been interacting for a long time, and that they feel glad that the US and Lurana are on the same team.

In the *outgroup condition*, participants read:

Many people consider Lurana a part of a different group or team from the United States. The United States and Lurana have different political systems, different approaches to managing their economies, and do not often work together through international organizations. Lurana is also ethnically and culturally different from the US. Americans often say that the US and Lurana have not interacted very much, and that they do not see the US and Lurana as being on the same team.

The ingroup manipulation draws on multiple dimensions of group identification characteristics

that are referenced in the international relations and psychology literature, as well as my own pilot data on the psychological determinants of group identification in order to construct a robust ingroup and outgroup cue while avoiding confounders as much as possible.⁵ These include regime type, economic systems, international organization membership, cultural and ethnic similarity, and psychological factors like satisfaction and history of interactions (Leach, Van Zomeren, Zebel, Vliek, Pennekamp, Doosje, Ouwerkerk and Spears 2008; Rathbun, Parker and Pomeroy 2025; Renshon and Powers 2024; Tomz and Weeks 2021, 2013).⁶ The outgroup manipulation mirrors this structure symmetrically, describing Lurana as differing from the United States across the same dimensions.

Reciprocity Treatments

Following the groupness treatment, all participants received information about a prior interaction between the United States and Lurana on a separate but related issue, which constituted the reciprocity treatment. The content of this prior interaction varied by reciprocity condition, with the specific cooperative domain adapted to match the issue type scenario. For the purposes of demonstration, here are the reciprocity treatments fit to the security scenario:

Positive Reciprocity Condition: Recently, the United States and Lurana interacted on another issue. In that case, the United States provided Lurana with assistance in improving monitoring of its national airspace to help detect potential security threats. After receiving this benefit, Lurana returned the favor. When the United States requested similar assistance, Lurana agreed to share its own information.

Violation of Positive Reciprocity Condition: Recently, the United States and Lurana have interacted on another issue. In that case, the United States provided Lurana with assistance in improving monitoring of its national airspace to help detect potential security threats. However,

⁵ See appendix for more information on pilot data.

⁶ It is worth noting that because my primary focus is on whether group identification moderates the effect of reciprocity, I am less concerned with *how* I create ingroups and outgroups and what factors determine identification with these groups as I am with ensuring that participants do in fact significantly and substantially *identify* with these groups or not. Future research on the determinants and origins of groups in international relations may be able to shed more light on this topic.

after receiving this benefit, Lurana did not return the favor. When the United States requested similar assistance, Lurana denied this request and did not share its own information.

Negative Reciprocity Condition: Recently, the United States and Lurana have interacted on another issue. In that case, the United States restricted the information it shared with Lurana about monitoring of national airspace for security threats, which made it harder to detect potential security threats. In response, Lurana returned the favor. Lurana also restricted its own information shared with the United States.

In the trade scenario, the prior interaction involved market access for household appliances rather than airspace monitoring information, but followed the same three-condition structure: positive reciprocity (Lurana returned market access when asked), violation of positive reciprocity (Lurana refused to return market access after receiving it), and negative reciprocity (the United States restricted Lurana's market access first, and Lurana responded in kind).

The exact issues of the previous reciprocity interactions were intentionally made different than the issues Lurana is requesting cooperation from the United States on, while staying within the same general domain of security or trade. This was done to avoid respondents interpreting the reciprocity treatments, and thus the ensuing request for cooperation, as examples of a uniform issue-type or country-specific policy by the United States instead of the intended interpretation as an example of reciprocity.

Outcome Measures and Mediators

Once respondents had confirmed they had read the vignette carefully, their attitudes on the primary outcome measures were then recorded. The first primary outcome is individual support for the proposed cooperation agreement between the United States and Lurana, measured on an 11-point likert scale ranging from 0 (strongly oppose) to 10 (strongly support). The second primary outcome is trust in Lurana, measured with the item "How trustworthy do you think Lurana is?" on an 11-point scale ranging from 0 (not at all) to 10 (completely). As a supplementary outcome, I also measured participants' affective evaluation of Lurana using a standard feeling thermometer ranging

from 0 to 100, which was rescaled to a 0-10 scale for analysis, as well as perceptions of hostility and threat towards the United States on a similar 11-point scale.

To explore the mechanisms underlying the treatment effects, I measured a set of potential mediators following the primary outcomes. These include reputational concerns (“If the US does not agree to the deal with Lurana, other countries will doubt America’s willingness to cooperate on issues in the future”), moral obligation (“Accepting the agreement with Lurana is the ‘right’ thing to do”), perceptions of fairness (“Lurana acts fairly towards other countries”), the desire to correct bad behavior (“The United States should reward or withhold cooperation in order to influence how other countries act in the future”), and long-run beliefs about the rewards of cooperative behavior (“The US will be better off in the long run if they make the agreement with Lurana”).⁷

Attention and Manipulation Checks

Participants also completed a set of attention/comprehension and manipulation checks. Attention checks asked participants to recall basic features of the scenario, including what problem Lurana had recently experienced, whether the situation affected the United States directly, what Lurana had requested from the United States, and what impact the agreement was expected to have on the United States.⁸

To assess whether the groupness manipulation was successful, I primarily use two group identification measures. The first is the Inclusion of Other in the Self (IOS) scale (Aron, Aron and Smollan 1992), a well-established measure of ingroup identification which uses a series of 7 overlapping circle pairs to measure the perceived closeness of the relationship between the United States and Lurana, with more overlap indicating a closer relationship. Participants in the ingroup condition are expected to select pairs with greater overlap than participants in the control condition, and participants in the outgroup condition are expected to select pairs with less overlap. For analysis, I coded the circle pairs 1-7, with 1 indicating the circle pair with the farther distance between them

⁷ All mediator items are measured on an 11-point scale ranging from 0 to 10. Mediation analyses using these items are reported in the appendix.

⁸ Results were robust to the inclusion and exclusion of those who failed attention checks.

and 7 indicating the pair of circles with the most overlap. I also included a text-based measure of ingroup identification, adapted from Cameron (2004), asking respondents to what extent they agree or disagree on an 11-point likert scale with the statement “The United States sees itself as a part of a group with Lurana”. I also included a direct manipulation check asking participants which statement best matched the description they read from the choices of: the US and Lurana were described as being on the same team, a different team, or no team language was used.

To assess whether the reciprocity manipulation was successful, I asked participants to recall what the United States did in the prior interaction and how Lurana responded. I also measured participants’ beliefs about Lurana’s general propensity to reciprocate positively or negatively, using items adapted from Perugini, Gallucci, Presaghi and Ercolani (2003), with items assessing whether Lurana would accept political costs to help a country that helped it, return favors, take revenge for wrongs, and respond in kind to unfair treatment. These scales can both be used as a manipulation check and a potential mediator of the reciprocity treatments.

Empirical Strategy

I estimate the effects of the experimental treatments using ordinary least squares regression. All continuous covariates are normalized to run from 0 to 1 to facilitate interpretation of the coefficients. The baseline model takes the form:

$$Y_i = \alpha + \beta_1 \text{Security}_i + \beta_2 \text{Group}_i + \beta_3 \text{Reciprocity}_i + \mathbf{X}_i \boldsymbol{\gamma} + \varepsilon_i \quad (3.1)$$

To test the BGR moderation hypotheses, I estimate a full interaction model that allows the effect of reciprocity to vary by the group identity of the country:

$$\begin{aligned}
Y_i = & \alpha + \beta_1 \text{Security}_i + \beta_2 \text{Ingroup}_i + \beta_3 \text{Outgroup}_i \\
& + \beta_4 \text{Violation}_i + \beta_5 \text{NegRecip}_i \\
& + \beta_6 (\text{Ingroup} \times \text{Violation})_i \\
& + \beta_7 (\text{Ingroup} \times \text{NegRecip})_i \\
& + \beta_8 (\text{Outgroup} \times \text{Violation})_i \\
& + \beta_9 (\text{Outgroup} \times \text{NegRecip})_i \\
& + \mathbf{X}_i \boldsymbol{\gamma} + \varepsilon_i
\end{aligned} \tag{3.2}$$

Where for respondent i :

- Y_i is the outcome of interest: either *policy support* or *trust*.
- Security_i is a dummy variable coded 1 for the security scenario and 0 for the trade scenario.
- Ingroup_i and Outgroup_i are dummy variables for the groupness treatment, with the control condition as the reference category. Hypotheses 3 and 4 predict $\beta_2 > 0$ and $\beta_3 < 0$, respectively.
- Violation_i and NegRecip_i are dummy variables for the reciprocity treatment, with the positive reciprocity condition as the reference category. Hypotheses 1 and 2 predict $\beta_4 < 0$ and $\beta_5 < 0$, respectively.
- β_6 and β_7 capture whether the effect of a reciprocity violation and negative reciprocity is larger when Lurana is an ingroup member than a control country. Hypothesis 5 predicts $\beta_6 < 0$ and $\beta_7 < 0$.
- β_8 and β_9 capture whether the effect of a reciprocity violation and negative reciprocity is larger when Lurana is an outgroup member than a control country. Hypothesis 6 predicts that β_8 and β_9 will not be statistically distinguishable from zero.
- $\mathbf{X}_i$ is a vector of pre-treatment covariates including age, gender, party identification, ideology, and foreign policy orientations.

- ε_i is the error term.

Empirical Results

Manipulation Checks

Before turning to the main results, I assess whether the experimental treatments successfully manipulated participants' perceptions of groupness and reciprocity as intended.

The groupness manipulation performed well on both measures. A one-way ANOVA reveals highly significant differences across group conditions in perceived overlap between the United States and Lurana on the Inclusion of Other in the Self scale ($F(2, 1719) = 548.8, p < 0.001$) and in group identification ($F(2, 1719) = 826.2, p < 0.001$). Post-hoc Tukey tests confirm that all three group conditions are significantly different from one another on both measures (all $p < 0.001$). Respondents in the ingroup condition perceive Lurana as significantly closer to the United States than control respondents (IOS difference: +1.85; group identification difference: +3.75), while respondents in the outgroup condition perceive Lurana as significantly more distant than control respondents (IOS difference: -0.74; group identification difference: -1.68). The ingroup and outgroup conditions are also significantly different from one another on both measures, confirming that the manipulation successfully induced meaningfully different perceptions of group membership across all three conditions.

The reciprocity manipulation also performed as intended. A one-way ANOVA reveals significant differences across reciprocity conditions in both positive reciprocity expectations ($F(2, 1719) = 826.2, p < 0.001$) and negative reciprocity expectations ($F(2, 1719) = 132.0, p < 0.001$). Post-hoc Tukey tests confirm that all pairwise comparisons across conditions are statistically significant (all $p < 0.001$). Respondents in the positive reciprocity condition hold the highest expectations that Lurana will return favors and cooperate in the future, while respondents in the violation condition hold the lowest positive expectations, a difference of approximately 5 points on the expectation scale. Respondents in the negative reciprocity condition fall between these two extremes on positive

expectations, but hold the highest negative reciprocity expectations, consistent with the treatment content in which Lurana responds in kind to an initial American restriction. Taken together, these results confirm that the groupness and reciprocity manipulations successfully shifted perceptions in the intended directions, providing a solid foundation for interpreting the main results.

Main Effects of Reciprocity and Group Identity

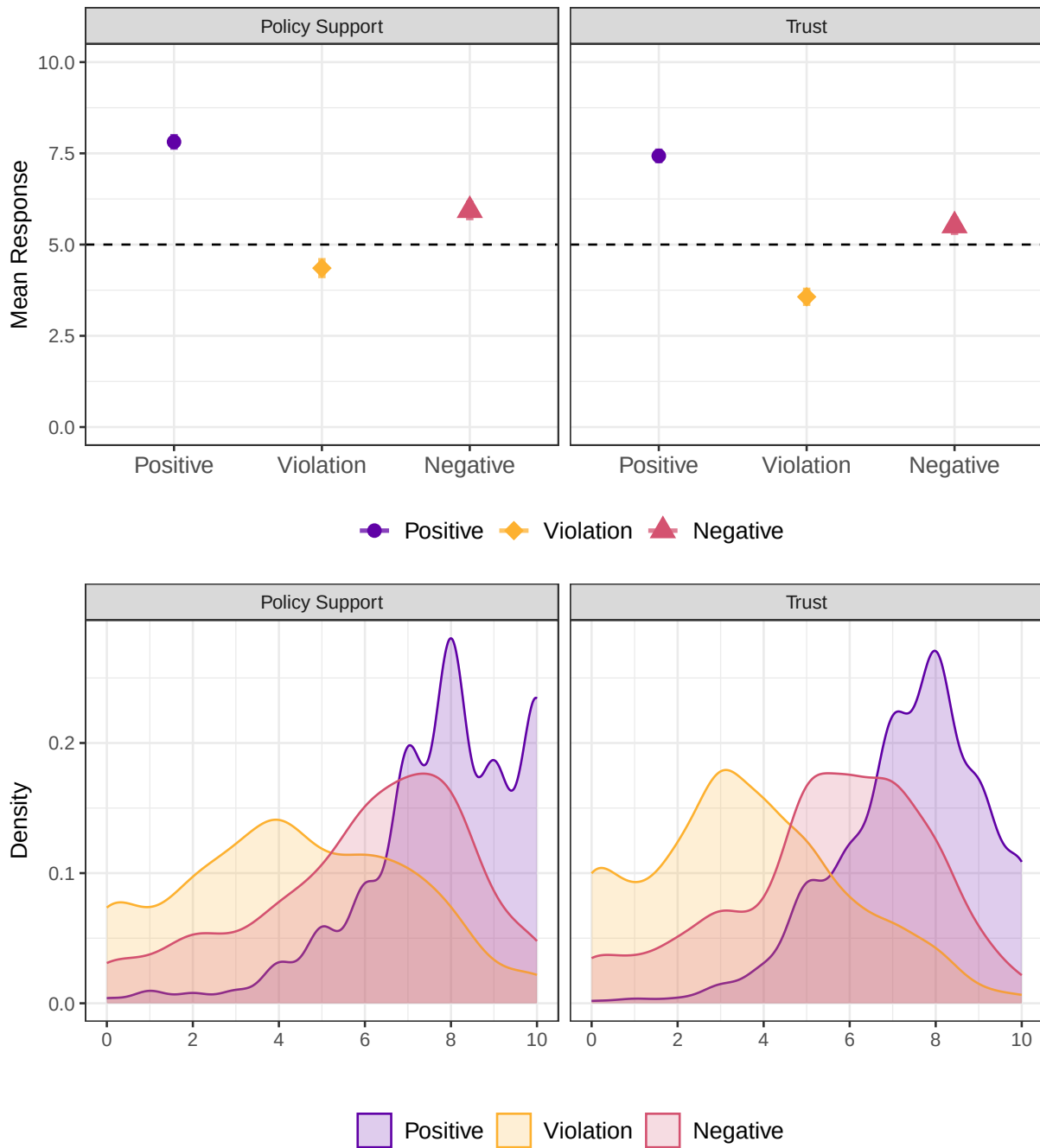
First, it should be noted that for ease of presentation and a more comprehensive discussion of group-based reciprocity's effect on cooperation, I pool across issue type and include a security scenario indicator as a covariate in all models. The pattern of results mirror each other in both scenarios, and the difference in mean policy support and trust between the security and trade scenarios is substantively negligible (approximately 0.22-0.27 points on a 0–10 scale, or less than a quarter of a point). Further, while the scenario difference is statistically significant for policy support ($t=2.04$, $p=0.04$), it is not for trust ($t=1.81$, $p=0.07$), indicating that it should be reasonable to treat issue type as a control rather than a theoretically meaningful dimension of the design.

I begin by examining the independent effects of reciprocity and group identity on policy support and trust. Figures 3.1 and 3.2 show the mean response and distribution of policy support and trust by reciprocity and group condition, respectively, and Figure 3.3 presents the estimated coefficients from the baseline models for both outcomes.⁹

The results offer strong support for Hypotheses 1 and 2. Exposure to a reciprocity violation by Lurana produces a large and highly significant decrease in both policy support and trust relative to the positive reciprocity condition. Respondents in the violation condition express approximately 3.5 points less support for the cooperation agreement ($p < 0.001$) and nearly 3.9 points less trust in Lurana ($p < 0.001$) than respondents in the positive reciprocity condition. The effect of negative reciprocity is similarly negative and significant, reducing policy support by approximately 1.9 points ($p < 0.001$) and trust comparably ($p < 0.001$). Notably, the violation condition produces a substantially larger drop in both outcomes than the negative reciprocity condition.

⁹ Results are consistent across all of the model specifications I ran.

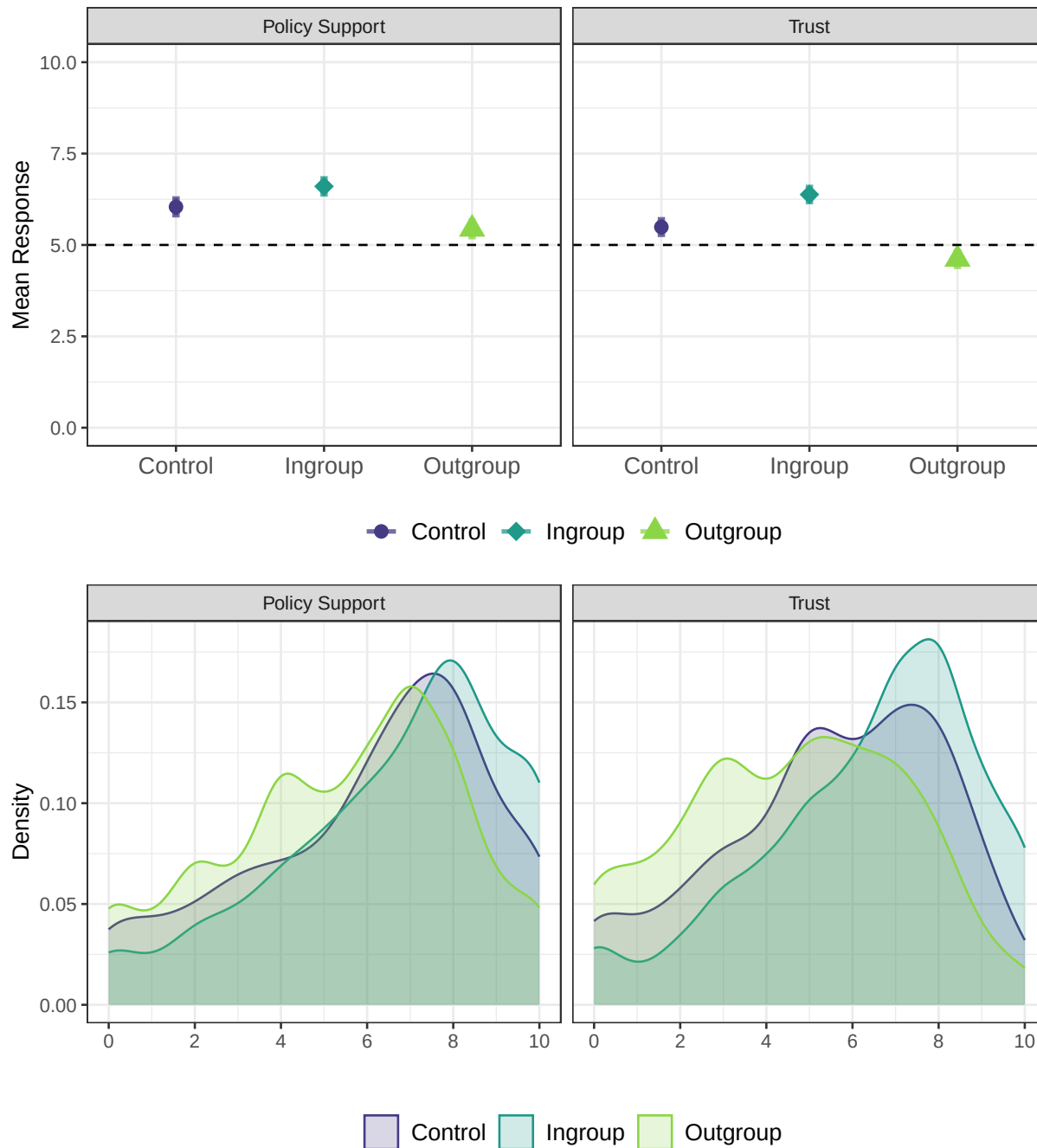
Figure 3.1: Policy Support and Trust by Reciprocity Treatment Group



Note: Top panels display condition means with 95% confidence intervals; the dashed line indicates the scale midpoint. Bottom panels display the distribution of responses by reciprocity condition.

Hypotheses 3 and 4 are also supported. Group identity independently shapes both outcomes in the predicted directions, although to a much lesser degree than reciprocity. Respondents in the ingroup condition express significantly higher support for cooperation with Lurana (approximately

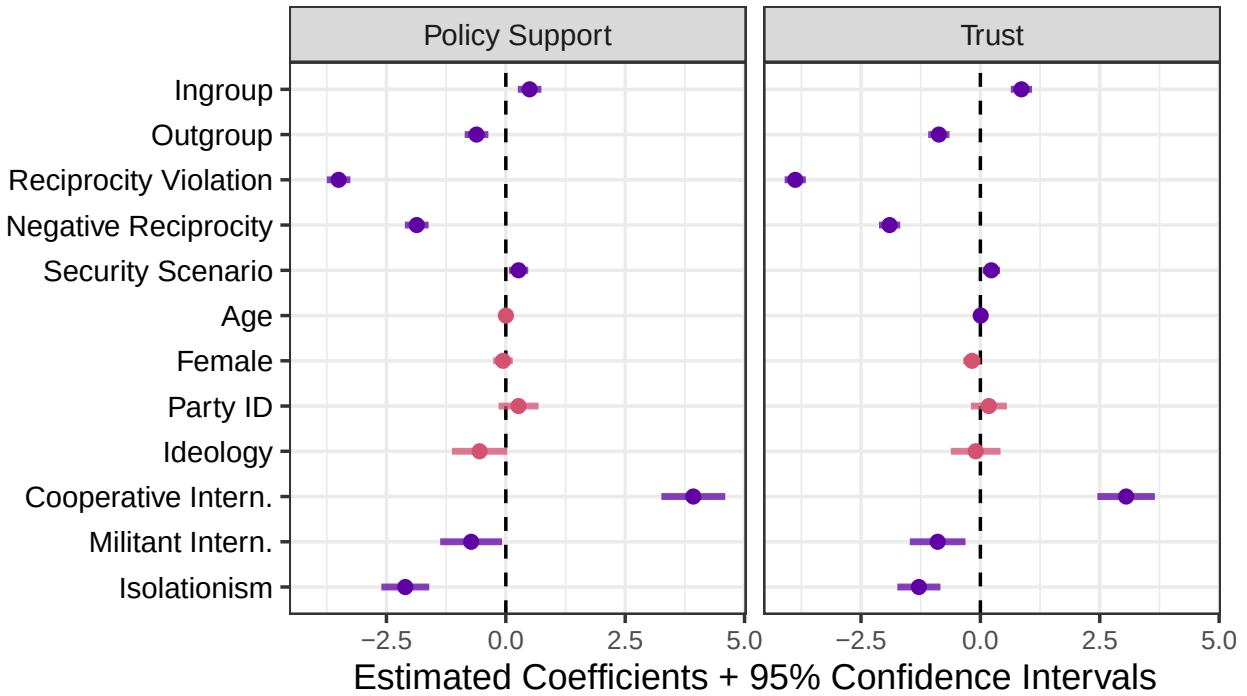
Figure 3.2: Policy Support and Trust by Group Treatment



Note: Top panels display condition means with 95% confidence intervals; the dashed line indicates the scale midpoint. Bottom panels display the distribution of responses by group condition.

0.5 points, $p < 0.001$) and significantly greater trust (approximately 0.86 points, $p < 0.001$) than respondents in the control condition. The outgroup effect runs in the opposite direction: respondents who learn that Lurana is an outgroup country express significantly lower support (approx-

Figure 3.3: The Effects of Reciprocity and Groups on Cooperation



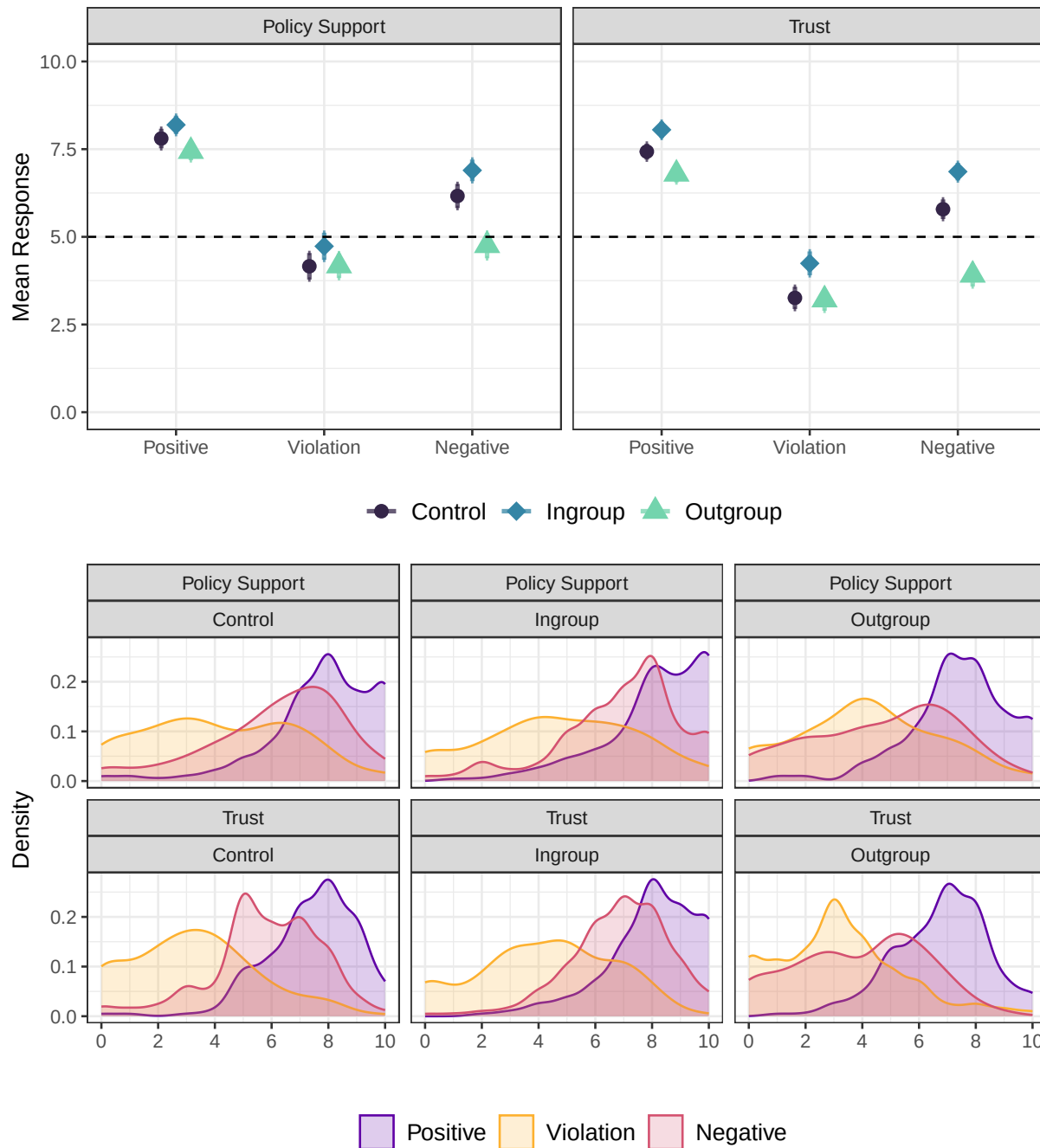
mately 0.61 points, $p < 0.001$) and lower trust (approximately 0.87 points, $p < 0.001$) than control respondents. These findings confirm that group identity shapes baseline evaluations of Lurana independently of its reciprocal behavior, consistent with the broader literature on ingroup favoritism in cooperative contexts.

Taken together, the main effects suggest that both who Lurana is and how Lurana behaves matter for American public support for cooperation and trust, but the question remains whether and how these two factors interact.

The Moderation of Reciprocity by Group Identity

Figure 3.4 presents the mean response of policy support and trust across the nine conditions of the 3×3 design, and Figure 3.5 displays the estimated coefficients from the interaction models for both outcomes. Contrary to the central predictions of BGR, I find that group identity does not significantly moderate the reciprocity effect, with one important exception.

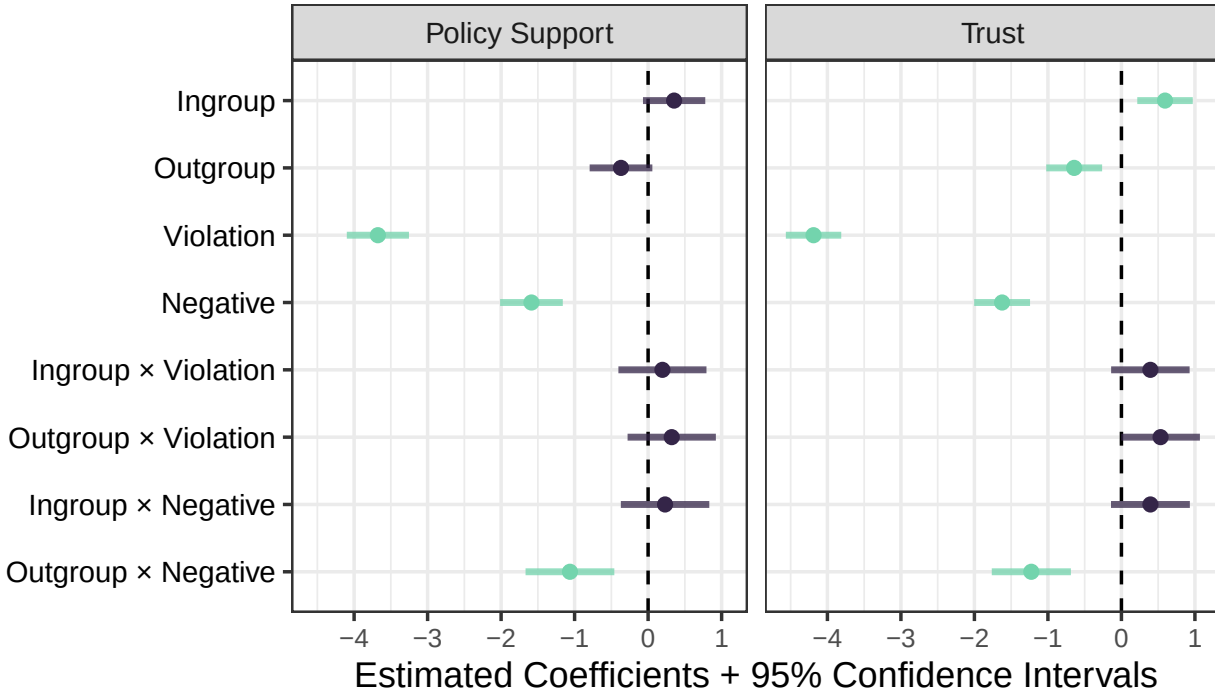
Figure 3.4: Policy Support and Trust by All Treatment Groups



Note: Top panels display condition means with 95% confidence intervals; the dashed line indicates the scale midpoint. Bottom panels display the distribution of responses by every treatment condition.

Hypothesis 5 predicted that ingroup countries would face a significantly larger penalty for reciprocity violations and negative reciprocity than control countries. The data does not support this prediction. The interaction terms for Ingroup $\times$ Violation and Ingroup $\times$ Negative Reciprocity are

Figure 3.5: Interaction Coefficients



substantively small and statistically indistinguishable from zero for both policy support (Ingroup $\times$ Violation: $\hat{\beta} = 0.19$, $p = 0.52$; Ingroup $\times$ Negative: $\hat{\beta} = 0.23$, $p = 0.45$) and trust (Ingroup $\times$ Violation: $\hat{\beta} = 0.39$, $p = 0.15$; Ingroup $\times$ Negative: $\hat{\beta} = 0.39$, $p = 0.15$). Ingroup countries that violate reciprocity or act negatively are not judged more harshly than control countries that behave identically. The elevated expectations that BGR predicts should attach to ingroup membership do not appear to translate into harsher evaluative responses when those expectations are violated.

Hypothesis 6 predicted that outgroup countries would not face a significantly larger penalty for reciprocity violations or negative reciprocity than control countries, a null prediction grounded in the logic that outgroup negative behavior is expectation-consistent and therefore absorbed rather than amplified. The data partially support this prediction, but with a striking reversal in the negative reciprocity condition. The interaction term for Outgroup $\times$ Violation is null for both policy support ($\hat{\beta} = 0.32$, $p = 0.29$) and trust ($\hat{\beta} = 0.53$, $p = 0.05$), consistent with Hypothesis 6. When an outgroup country merely fails to return a favor it received, respondents do not react more negatively than they would to the same behavior by an unknown control country.

However, the Outgroup $\times$ Negative Reciprocity interaction tells a different story entirely. When an outgroup Lurana responds to an initial American non-cooperative act with one of its own, engaging in negative reciprocity, respondents penalize Lurana significantly more harshly than the control in the same condition. This effect is large and highly significant for both policy support ($\hat{\beta} = -1.06$, $p < 0.001$) and trust ($\hat{\beta} = -1.23$, $p < 0.001$). Outgroup countries that engage in negative reciprocity face the steepest cooperative penalties of any group condition, suffering a larger drop in support and trust when moving from positive to negative reciprocity than either ingroup or control countries do. This finding directly contradicts my prediction that outgroup negative behavior should be absorbed rather than amplified based on reciprocity expectations.

The top panel of Figure 3.4 illustrates this pattern clearly. Across both outcomes, policy support and trust decline steeply as reciprocity moves from positive to violation to negative for all three group conditions. From the positive to the violation condition, the three group lines move roughly in parallel: the gap between ingroup, control, and outgroup conditions remains largely stable, indicating that the drop produced by a reciprocity violation is similar regardless of group membership. However, as reciprocity moves from the violation condition to the negative reciprocity condition, the outgroup line diverges sharply from the ingroup and control lines, falling substantially further than the other two conditions. This widening gap in the negative reciprocity condition is the visual confirmation of the outgroup amplification effect: outgroup countries are not simply evaluated more negatively in general, they are disproportionately penalized when they engage in negative reciprocity, over and above their baseline outgroup penalty.

Summary of Findings

Table 3.1 summarizes the hypothesis tests. The results indicate mixed support for BGR's ability to explain the micro-level process of how reciprocity works in international relations. On the one hand, the finding that reciprocity is a significant independent influence on cooperative outcomes is consistent with the BGR account. Trading negative acts towards one another, and even worse, breaking an implicit debt by not repaying a favor, should decrease policy support and trust towards offending countries. On the other hand, my central moderation prediction, that ingroup countries

will face harsher penalties for reciprocity violations because they break elevated expectations, is not supported by the data. Further, the outgroup null prediction is only partially supported; the effect of negative reciprocity is amplified by outgroup identification. Outgroups that engage in an explicitly negative, and in some views the most adversarial, form of reciprocity appear to face a uniquely harsh cooperative response.

Table 3.1: Summary of Hypothesis Tests

Hypothesis	Prediction	Supported?	Outcomes
H1	Reciprocity violation $\downarrow$ support/trust vs. positive	✓	Both
H2	Negative reciprocity $\downarrow$ support/trust vs. positive	✓	Both
H3	Ingroup $\uparrow$ support/trust vs. control	✓	Both
H4	Outgroup $\downarrow$ support/trust vs. control	✓	Both
H5	Ingroup $\times$ violation/negative $\downarrow$ more than control	$\times$	Both
H6	Outgroup $\times$ violation/negative not $\downarrow$ more than control	Partial	Both

Discussion

If BGR’s theoretical expectations can be summarized as “you can forgive a stranger more easily than a friend”, the results of this study imply that when it comes to reciprocity between countries, the more appropriate description would be that there is no grace for friends nor mercy for enemies. The strong support for Hypotheses 1 and 2, the lack of support for Hypothesis 5, and the exception to Hypothesis 6 are striking results and have important implications for our understanding of how reciprocity works in cooperative settings like interstate politics that are saturated with group dynamics.

Reciprocity Matters — A Lot

One takeaway from these results, in line with prior studies conducted at the state-level, is that reciprocity is a powerful determinant of cooperative behavior and attitudes. Across and within different contexts, reciprocity and violations of reciprocity have a sizable effect on the policy attitudes of the public. People want to return favors or slights, and oppose cooperating with countries who do not follow the rules regardless of the issue area or the country or an individual’s personal char-

acteristics. Fair is fair, and playing by the rules seems to be important to the public's judgments on other countries. The finding that most respondents reacted more negatively to a violation of reciprocity than tit-for-tat negative reciprocity particularly buttresses this perspective. One instance, while inherently unfavorable, still follows the rules of reciprocity, while the other is a break from an expected and ubiquitous rule of cooperative behavior.¹⁰

Coupled with the finding that reciprocity was a substantially larger influence on cooperative attitudes than groupness, further research should explore whether reciprocity cues drown out the “noise” of other information about state characteristics or intentions, whether they be group membership or something else. Take some of the real world examples mentioned in the introduction; one possible implication from these results could be that the harsh backlash friendly countries face from leaders and publics, as well as retaliatory policy responses, for not repaying favors may be less about group membership than they first appear. Instead, these instances might be telling stories about a selection effect, where people are reacting how they normally would towards any country that has let them down. It just so happens that the countries that let us down tend to be the ones we interact with the most or consider close.

Bounded or Unbounded Reciprocity

A central feature of the BGR theory is that people instinctively tend to categorize themselves into groups in social exchange settings. This derives from work in evolutionary and social psychology that indicates humans are wired to view interactions through a group “lens”, because groups are important to survival and receiving cooperative benefits. BGR's main predictions about ingroup bias, like favoring, monitoring, and sometimes punishing ingroup members more than others, stem from this group framework. In other words, reciprocity is “bounded” by the group. The evidence from this study suggests that the world of interstate affairs may be an exception, where people are less concerned about what groups countries belong to, and simply more about whether they follow

¹⁰ Supplementary mediation analysis in the Appendix attempts to partially address whether individuals use reciprocity as a heuristic by estimating whether the effect of reciprocity is mediated by the psychological mechanisms outlined in BGR and other theories of reciprocity. The results confirm that mechanisms associated with the heuristic use of reciprocity significantly mediate the effect of reciprocity in this study.

the rules of reciprocity or not.

How can we explain this finding? Romano, Balliet and Wu (2017) develop an extension of BGR that explains the conditions under which reciprocity may become “unbounded”, or extend reciprocity expectations and behaviors to those beyond the ingroup. In this account of reciprocity, people become “good” reciprocators and cooperative partners when they are made aware of the possibility that outgroup members, and not just ingroup members, can affect whether they receive indirect benefits or not. This works through a reputational logic: if Person A’s reciprocity behavior towards Person B is being observed by Person C, and Person A’s behavior can be communicated to others (e.g. gossip), Person A is more likely to follow the rules of reciprocity towards Person B regardless of the group membership of Person B or C. Violating reciprocity expectations decreases a person’s chances of receiving indirect benefits from cooperation from third parties, like others “paying it forward” by granting a favor unprompted or rewarding someone for helping another out. From the BGR perspective, these indirect and reputational benefits are confined to the group, because you are unlikely to interact with the outgroup and your fellow ingroup members are unlikely to care about how you behave when you do, because they are also not relying on outgroup members. The theory of Unbounded Indirect Reciprocity explains what happens when this is not the case: the moderating effect of groupness disappears.

For reciprocity to become “unbounded”, people need informational cues that tell them their behavior is being monitored or communicated to future interaction partners, regardless of group membership. Theoretically this can take a few forms, such as information about gossiping, public monitoring, or social networks (and thus the ability to gossip). Firstly, in today’s world of interconnectedness countries can expect to interact with non-ingroup members on a regular basis, and even regularly receive third party benefits or aggressions from them. One needs to look no farther than relations between the US, China, and Russia to see how direct and indirect reciprocity can occur (Goldstein 1991; Goldstein and Freeman 1990, 1991). In terms of communicating or observing reciprocity behavior, these types of informational cues are numerous and likely common knowledge to both leaders and members of the public. Countries, regardless of group affiliation, can share information about other countries with one another in a variety of ways. Diplomatic visits, co-

operation agreements, sharing intelligence, collaboration on national projects and other behaviors are all common and well-known interactions that could spread a country's past behavior to others. These interactions are not restricted to exchanges between friendly countries either, where a country or its fellow ingroup members networks overlap with outgroup countries. Reciprocity can also often be directly observable: many countries (and private actors) have expansive surveillance and intelligence apparatus designed to monitor other states' behavior, and international organizations or multilateral agreements are formal opportunities to monitor what other countries do.

Thus, one potential alternative explanation for the lack of interaction between ingroupness and reciprocity is people perceive the possibility that outgroups can affect the indirect cooperative benefits the United States can receive. The knowledge that the US may have potential future interactions with other outgroups, or that other ingroup countries' may interact with outgroups, may change the incentive to focus on ingroup reciprocity to a focus on *anyone's* reciprocity. Prior research examining the individual-level interaction effect between country similarity and reciprocity on trust in the context of COVID-19 aid found similarly null results (Roskos-Ewoldsen, Ellithorpe and Kinne 2026). Future research should more explicitly control for or manipulate these types of informational cues to test whether that attenuates the relationship between group membership and reciprocity.

Still, anecdotal examples where people use explicitly group and reciprocity-relevant logics for justifying feelings of betrayal or harsh policy responses, like the US reaction to the Iraq War and war with Iran, indicate that there may be other factors that are missing from this study's picture of group-based reciprocity. Why were countries like France and Germany the most targeted by US leaders and the public when other countries considered "allies" or "partners" who have been assisted by the US in the past did not suffer nearly the same degree of backlash? Perhaps in international contexts, considerations like power, status, or the deepness of past interactions further complicate the boundary conditions of reciprocity and groupness.

The Outgroup $\times$ Negative Reciprocity Amplification Effect

Perhaps the most interesting finding about group membership moderating the effect of reciprocity is respondents' drop in support and trust for Lurana when Lurana is framed as an outgroup. While I predicted that outgroup, negative reciprocity behavior to be absorbed rather than amplified due to already low expectations, the data tells the opposite story. Outgroups that engage in negative reciprocity face the steepest cooperative penalties of any condition in the design.

I consider several possible explanations accounts below, though exploring this question goes beyond the scope of the present study and represents an important direction for future research.

First, one possibility is that the negative outgroup amplification effect is actually reflecting a story in reverse, and a product of how I presented the results. In other words, the data is telling us that outgroups that reciprocate a negative act are in fact in line with people's expectations and are predictably low, but when outgroups cut against expectations and surprise us with positive reciprocity, they experience the greatest gains in support and trust.

Since my design involved 3 reciprocity conditions ranging from positive, positive violation, and negative reciprocity, for analysis and interpretation of results I had to pick a condition as the reference category. I chose positive reciprocity as the reference category partially because the real world puzzles that motivated this study are based on what happens when friendly countries do not follow the rules of reciprocity or act negatively. This decision may obscure a finding that may be in line with BGR's logic about what happens when actors break from the reciprocity expectations people have of them. If respondents are pleasantly surprised when an outgroup country behaves cooperatively in the positive reciprocity condition, the subsequent shift to negative reciprocity may represent a larger psychological distance than the same shift for ingroup or control countries, precisely because the positive baseline was unexpected. In other words, if negative reciprocity was instead the reference category, the outgroup amplification effect is not about responding more sharply to a negative act but about the contrast between an unexpected cooperative baseline and a negative outcome. Because BGR is generally focused on ingroup bias as opposed to outgroup derogation, and generally assumes a theoretical outgroup neglect, there are to my knowledge no

precise predictions about outgroup membership moderating the effect of reciprocity (Balliet, Wu and De Dreu 2014). This may be a fruitful avenue of future research, exploring whether group moderation of reciprocity is specific to group characteristics or a reflection of a more general process when expectations are broken.

Another possibility is that the outgroup amplification effect is driven by something related to threat perception rather than reciprocity expectations. When an outgroup country, and in this case a specific outgroup that has significant differences across conflict-adjacent dimensions like political regime type and ethnicity, responds to an American non-cooperative act with one of its own, this behavior may be interpreted not merely as tit-for-tat reciprocity but as a signal of hostility and adversarial intent. Outgroup countries that engage in negative reciprocity may be perceived as confirming and escalating their outgroup status, crossing a threshold from merely an “other” to active threat. This interpretation is consistent with research on intergroup conflict and threat perception in Social Identity Theory, which finds that outgroup members who behave in ways that confirm negative stereotypes or signal hostility elicit stronger aversive responses and derogation than those who behave neutrally (Tajfel 1974). This may also align with theories in international relations on threat spirals, where uncertainty about a country’s type or intentions can lead to an increasingly negative spiral of aggressions (Jervis 1976; Kydd 1997). Indeed, supplementary analysis using measured threat as the main outcome shows that outgroupness, violations of reciprocity and negative reciprocity significantly and substantially increase perceptions of threat, although there is no interaction between groupness and reciprocity in any condition.¹¹ Interestingly, the pattern of violations of reciprocity eliciting a larger response than a negative act still holds in the threat outcome, which indicates that respondents were still attuned to reciprocity considerations. In this interpretation, the outgroup negative reciprocity effect is not about expectation violation at all, but rather it is about perceived threat escalation, which produces a qualitatively different and theoretically underexplored manifestation of reciprocity.

A third possibility could be psychological retribution, either within the framework of reciprocity or through a different process. When an outgroup country engages in negative reciprocity, respon-

¹¹ Results included in the appendix

dents may interpret this not as understandable tit-for-tat behavior but as an act of defiance that warrants punishment. This interpretation would predict that the outgroup amplification effect is moderated by respondents' prior attitudes toward outgroup countries, a hypothesis that future research could test directly. From the perspective of reciprocity, some research has shown that positive and negative reciprocity in social exchanges is asymmetric; separate from the perceived "value" of the reciprocal act itself, people also assign "value" to the act of doing something positive or negative, and in particular negative acts are assigned greater values (Keysar, Converse, Wang and Epley 2008). What may objectively appear to be roughly equivalent behaviors can thus become escalatory. This is theorized to encourage prosocial behavior, where cooperative acts beget more generous positive acts, while non-cooperative acts are stamped out through escalation. Similarly, McDermott, Lopez and Hatemi (2017) argues that people may seek revenge, which goes beyond a reciprocal desire to "get even", as a pathway to deterrence or conflict resolution. Feelings like hatred or anger and the urge to achieve revenge at any cost may be enough to prevent adversaries from engaging in an undesired action. Similarly, people's desire to moralize and seek justice for themselves and others may lead to steeper retribution (Schroeder, Steel, Woodell and Bembenek 2003).

Whatever the underlying mechanism, the outgroup negative reciprocity finding has important implications for BGR as a theory of group-based cooperation in international relations. The result suggests that the relationship between group identity and reciprocity is more complex than BGR can currently explain, particularly in regards to outgroups. Rather than group identity simply raising or leaving unchanged the sensitivity to reciprocity violations, it may condition reciprocity effects differently depending on the type and direction of the reciprocal act. Disentangling these mechanisms is an important task for future work, and suggests that a micro-level theory of group-based reciprocity in international relations will need further development and precision in its predictions.

Conclusion

When Canada felt the sting of American tariffs, when NATO allies hesitated over the war with Iran, and when France refused to join the Iraq War, the reactions from American leaders and the public

shared a common thread: a sense that these countries had violated an implicit rule of cooperative behavior, and that these violations hurt more because they came from a country close to them. Examples like these are striking, as they highlighted that people were just as much or even more concerned with how these countries had behaved compared to the objective costs to their country. This paper examines that logic more closely, and asks whether group membership conditions how people respond to reciprocity failures by friendly and unfriendly countries alike.

The results tell a clear and unexpected story. Reciprocity is a powerful and robust influence on public support for international cooperation and trust in other countries. Across issue areas, group conditions, and individual attributes, violations of reciprocity and negative reciprocal acts produce large drops in both policy support and trust. This finding is consistent with the BGR framework and with the broader literature on the importance of reciprocity to sustained cooperation; people follow the rule of reciprocity, and they expect others to do the same. But contrary to one of BGR's central theoretical predictions, this reciprocity effect generally does not depend on whether the offending country is perceived as a friend, a stranger, or an enemy. There is no grace for friends and no mercy for enemies. Ingroup countries that violate reciprocity or engage in tit-for-tate negative reciprocity are not judged more harshly than control countries that do the same, suggesting that the elevated expectations BGR predicts for ingroup members does not translate into stronger penalties when those expectations are violated in the context of interstate politics.

This null finding is theoretically important. It suggests that in international relations contexts, the reciprocity heuristic may operate in an "unbounded" fashion (Romano, Balliet and Wu 2017), extending across group lines rather than being confined to ingroup interactions. The interconnected nature of international politics, where countries routinely interact across group boundaries - where reputations travel through diplomatic, economic, and informational networks, and where the indirect consequences of non-cooperation can be felt regardless of group membership - may create the conditions under which the group-based boundaries of reciprocity disappear. People may simply care whether countries follow the rules, irrespective of what team they are on.

The one exception to this pattern is the finding that outgroup countries face a disproportionate penalty when they engage in negative reciprocity. This outgroup amplification effect is not pre-

dicted by BGR and runs in the opposite direction of the theory's expectation that outgroup negative behavior should be absorbed rather than amplified due to already low cooperative expectations. Several possible explanations are plausible: the results may actually reflect a steep positive jump for outgroups that defy expectations and positively reciprocate; outgroup negative reciprocity may function through mechanisms related to threat and outgroups; or it may trigger retributive motivations that are stronger toward outgroups. Adjudicating between these accounts will require further research, but the finding points toward the need for a more nuanced theory of group-based reciprocity in international relations.

This paper makes several contributions to the study of international cooperation, public opinion, and the micro-foundations of reciprocity. First, it demonstrates that reciprocity is not just a feature of state behavior but a meaningful determinant of individual-level foreign policy attitudes, operating through psychological mechanisms like emotion, fairness, moral obligation, and reputational concern. Second, to my knowledge, this is one of the first experimental tests of whether group identity moderates reciprocity effects in international cooperation contexts. Third, my findings indicate that the group-based conditionality predicted by BGR may not cleanly generalize from laboratory settings to the domain of international politics. Finally, the outgroup amplification finding opens a new empirical puzzle about when and why group membership amplifies rather than dampens responses to reciprocal acts by outgroups, with implications for how publics and leaders respond to cooperative and non-cooperative behavior from countries perceived as fundamentally different from themselves.

Several limitations of the present study should be noted. The sample is limited to the American public, and it is possible that the lack of group moderation reflects features specific to how Americans perceive international group membership, or to the particular salience of reciprocity norms in American political culture. Cross-national replications, including samples from countries with different relationships to international institutions and different histories of ingroup-outgroup conflict, would help strengthen the generalizability of these findings. In terms of design, the use of a fictional country avoids confounding from prior attitudes toward real countries, but also limits the ecological validity of the results; it is possible that the group moderation effects predicted by

BGR would emerge more clearly when the ingroup and outgroup countries are real and deeply meaningful to respondents. Important factors relevant to intergroup bias, like informational cues about gossip, were also not controlled for or manipulated in this study. Finally, the cooperative outcomes measured in this study (support for a cooperation agreement and trust) may not translate directly into actual policy preferences or political behavior, and future work might make use of more behavioral outcomes or elite-level samples.

Despite these limitations, the results of this study have important implications for how we think about the role of reciprocity and group identity in international relations. The finding that reciprocity operates powerfully and without meaningful group moderation suggests that countries cannot rely on the goodwill generated by ingroup status to absorb the costs of non-cooperative behavior. Friends who fail to return favors, or who act negatively, face the same reputational and cooperative consequences as strangers who do the same. Adversaries, on the other hand, may have the most to gain or the most to lose through reciprocity; the gap between cooperative reciprocity and non-cooperative is extremely steep. In a world where countries are increasingly interconnected across traditional group boundaries, the rules of reciprocity appear to apply broadly, and the public, at least, seems to be keeping score.

Appendix

Full Survey Instrumentation

The survey was administered in three blocks: a pre-treatment block collecting demographic data and covariates, a vignette and treatment block, and a post-treatment block collecting outcome measures, mediators, and manipulation checks. The full text of all items is reproduced below.

Block 1: Demographics and Covariates

Citizenship and Background

Are you a US citizen or permanent resident?

US citizen Permanent resident Temporary visa Something else (please specify)

Were you born in the US?

Yes No

[If not born in the US] What is your nationality? *[drop-down menu of countries]*

What is your age (in numeric years)?

Are you: Male Female Nonbinary Prefer own terminology (please enter) Prefer not to say

What is your highest level of education achieved?

Some high school High school degree/GED Some college Associate's degree or other 2-year degree Trade degree Bachelor's degree or other 4-year degree Master's degree PhD, MD, or other professional degree

What is your race/ethnicity (select all that apply)?

Hispanic or Latino/a White/Caucasian Black/African American Indigenous, Native American, or Alaskan Native Asian or Pacific Islander Arab or Middle Eastern
Other (please enter)

What political orientation do you identify with?

Very Liberal Liberal Somewhat Liberal Neither Liberal nor Conservative Somewhat Conservative Conservative Very Conservative

What political party in the US do you most align with?

Strong Democrat Lean Democrat Independent Lean Republican Strong Republican Some Other Party

[If answered Independent or Some Other Party] If you had to choose, do you think of yourself as closer to the Democratic Party or the Republican Party?

Closer to Democratic Party Closer to Republican Party Neither

Foreign Policy Orientations (Rathbun et al. 2016) The following items were presented in randomized order. Respondents indicated their level of agreement or disagreement on a 7-point scale ranging from strongly disagree to strongly agree.

Cooperative Internationalism ($\alpha = 0.88$)

1. The United States needs to cooperate more with the United Nations.
2. It is essential for the United States to work with other countries to solve problems such as over-population, hunger, and pollution.
3. Promoting and defending human rights in other countries is of utmost importance.
4. Helping to improve the standard of living in less developed countries is of utmost importance.

Militant Internationalism ($\alpha = 0.84$)

1. The United States should take all steps including the use of force to prevent aggression by any expansionist power.
2. Rather than simply countering our opponents' thrusts, it is necessary to strike at the heart of an opponent's power.
3. Going to war is unfortunate but sometimes the only solution to international problems.
4. American military strength is not the best way to ensure world peace. (*Reverse-coded*)

Isolationism ($\alpha = 0.78$)

1. The U.S. should mind its own business internationally and let other countries get along the best they can on their own.
2. We should not think so much in international terms but concentrate more on our own national problems.
3. The U.S. needs to play an active role in solving conflicts around the world. (*Reverse-coded*)
4. Our allies are perfectly capable of defending themselves and they can afford it. The United States should focus on internal rather than external threats to its well-being.

Block 2: Vignettes and Treatments

Instructions

Please remember to read closely and pay attention. During this study you will be asked questions to check your memory and comprehension.

There's a lot of talk these days about relations between the United States and other countries in the world. We'd like to get your thoughts about a situation the US could face in the future.

The following text describes a scenario about another country called "Lurana" — a fictional country created for the purposes of this study — interacting with the US. After describing the situation, we will ask for your opinions and beliefs about the issue.

Security Scenario (Cybersecurity) *Common Preamble:*

Lurana is a country far from the United States. *[Groupness treatment inserted here.]*

Lurana has recently experienced several cyberattacks targeting its infrastructure. These attacks do not affect the US or involve the US directly. The government of Lurana is taking steps to improve its cybersecurity. Here are some important points about the situation:

Unknown groups have launched cyberattacks against Lurana's electrical grid and transportation systems, causing temporary disruptions. Although Lurana has restored these systems, officials are concerned about future attacks.

Lurana has asked the United States for a security agreement. Lurana wants the US government to share technical information about how to detect and prevent cyberattacks. If the US agrees, security officials in Lurana would use this information to strengthen their systems against future threats. From the US perspective, sharing this information with Lurana is not expected to help or hurt the United States.

[Reciprocity treatment inserted here.]

Trade Scenario *Common Preamble:*

Lurana is a country far from the United States. *[Groupness treatment inserted here.]*

Lurana is experiencing economic challenges in its manufacturing sector. These challenges do not affect the US or involve the US directly. The government of Lurana is trying to support local industries. Here are some important points about the situation:

Factories in Lurana that produce textiles have recently faced declining demand, leading to job losses and financial strain in the sector.

Lurana has asked the United States for a trade agreement. Lurana wants the US to allow more clothing products from Lurana to be sold in the US. Textile exports are an important part of Lurana's economy. If the United States agrees, Lurana hopes increased exports will help support jobs and stabilize the industry. From the US perspective, the trade agreement with Lurana is not expected to help or hurt the United States.

[Reciprocity treatment inserted here.]

Groupness Treatments *Control Condition:* No additional information about Lurana's relationship to the United States was provided.

Ingroup Condition:

Many people consider Lurana a part of the same group or team as the United States. Both the United States and Lurana share democratic governments, free-market economies, high levels of freedom of speech, and shared membership in many international organizations. Lurana is also ethnically and culturally similar to the US. Americans often say that the US and Lurana have been interacting for a long time, and that they feel glad that the US and Lurana are on the same team.

Outgroup Condition:

Many people consider Lurana a part of a different group or team from the United States. The United States and Lurana have different political systems, different approaches to managing their economies, and do not often work together through international organizations. Lurana is also ethnically and culturally different from the US. Americans often say that the US and Lurana have not interacted very much, and that they do not see the US and Lurana as being on the same team.

Reciprocity Treatments — Security Scenario *Positive Reciprocity Condition:*

Recently, the United States and Lurana interacted on another issue. In that case, the United States provided Lurana with assistance in improving monitoring of its national airspace to help detect potential security threats. After receiving this benefit, Lurana returned the favor. When the United States requested similar assistance, Lurana agreed to share its own information.

Violation of Positive Reciprocity Condition:

Recently, the United States and Lurana have interacted on another issue. In that case, the United States provided Lurana with assistance in improving monitoring of its national airspace

to help detect potential security threats. However, after receiving this benefit, Lurana did not return the favor. When the United States requested similar assistance, Lurana denied this request and did not share its own information.

Negative Reciprocity Condition:

Recently, the United States and Lurana have interacted on another issue. In that case, the United States restricted the information it shared with Lurana about monitoring of national airspace for security threats, which made it harder to detect potential security threats. In response, Lurana returned the favor. Lurana also restricted its own information shared with the United States.

Reciprocity Treatments — Trade Scenario *Positive Reciprocity Condition:*

Recently, the United States and Lurana have interacted on another issue. In that case, the United States allowed more household appliances, such as refrigerators and washing machines, from Lurana to be sold in the US market. After receiving this benefit, Lurana returned the favor. When the United States requested similar assistance, Lurana allowed more household appliances from the United States to be sold in its market.

Violation of Positive Reciprocity Condition:

Recently, the United States and Lurana have interacted on another issue. In that case, the United States allowed more household appliances, such as refrigerators and washing machines, from Lurana to be sold in the US market. However, after receiving this benefit, Lurana did not return the favor. When the United States requested similar assistance, Lurana denied this request and did not allow more household appliances from the United States to be sold in its market.

Negative Reciprocity Condition:

Recently, the United States and Lurana have interacted on another issue. In that case, the United States restricted the number of household appliances, such as refrigerators and washing

machines, from Lurana that could be sold in the US market, which made it harder for Lurana's companies to sell their products abroad. In response, Lurana returned the favor. Lurana also restricted the number of household appliances from the United States that could be sold in its market.

Block 3: Outcome Measures, Mediators, and Manipulation Checks

Primary Outcome Measures

We would like you to rate your feelings about Lurana using something we call the feeling thermometer. Ratings between 50 and 100 degrees mean that you feel favorable and warm toward Lurana. Ratings between 0 and 50 degrees mean that you don't feel favorable toward Lurana. You would rate Lurana at 50 degrees if you don't feel particularly warm or cold toward the country. *How would you rate Lurana?* [0–100]

Do you support or oppose the proposed agreement with Lurana? [0 strongly oppose — 10 strongly support]

How trustworthy do you think Lurana is? [0 not at all — 10 completely]

Supplementary Outcome Measures

Lurana is hostile toward the U.S. [0–10]

Lurana is a threat to American interests. [0–10]

Mediators All mediator items were measured on an 11-point scale from 0 (not at all) to 10 (extremely).

If the U.S. does not agree to the deal with Lurana, other countries will doubt America's willingness to cooperate on issues in the future.

Accepting the agreement with Lurana is the “right” thing to do.

Lurana acts fairly towards other countries.

The United States should reward or withhold cooperation in order to influence how other countries act in the future.

The U.S. will be better off in the long run if they make the agreement with Lurana.

Groupness Manipulation Checks

Which statement appears closest to the description you read?

- (A) The US and Lurana were described as being on the same team
- (B) The US and Lurana were described as being on a different team
- (C) No team language was used

[Control and outgroup conditions] Please indicate your agreement or disagreement with the following statement (−5 strongly disagree to +5 strongly agree):

The United States sees itself as a part of a group with Lurana.

[Ingroup condition only] Please indicate your agreement or disagreement with the following statements (−5 strongly disagree to +5 strongly agree):

The U.S. does not have a lot in common with Lurana and countries in its group.

Overall, being on the same team as Lurana is important to the U.S.’s self-image.

Generally, the U.S. feels good about being in the same group as Lurana.

The United States identifies with countries in the United States and Lurana group.

Is there anything that is important to you when deciding whether the US sees itself as part of a group with Lurana? If so, can you explain? *[Yes (open text) / No / Not sure]*

[Outgroup condition only]

Is there anything that is important to you when deciding whether the US sees itself as part of a different group than Lurana? If so, can you explain? *[Yes (open text) / No / Not sure]*

Inclusion of Other in the Self (IOS) Scale (Aron, Aron and Smollan 1992; ?): Respondents were shown seven pairs of overlapping circles and asked to choose the pair that best represented the relationship between Lurana and the United States, with more overlap indicating a closer relationship. Responses were coded 1–7, with 1 indicating the most distant pair and 7 indicating the most overlapping pair.

Reciprocity Manipulation Checks

In the past interaction, what did the United States do?

- (A) Provided assistance to Lurana
- (B) Imposed restrictions on Lurana
- (C) A previous interaction with Lurana was not mentioned

In the past interaction, how did Lurana respond to the United States?

- (A) Returned assistance to the United States
- (B) Refused to return assistance
- (C) Imposed restrictions on the US
- (D) No response was described

Reciprocity Propensity Items (Perugini et al. 2003): The following items were presented on a scale from –5 (strongly disagree) to +5 (strongly agree).

Lurana would accept political costs in order to help a country who helped it before.

If a country does a favor for Lurana then Lurana would be ready to return the favor.

If a country asks Lurana for information, Lurana would be happy to help.

If Lurana suffers a serious wrong, Lurana will take revenge as soon as possible, no matter what the costs.

Lurana would be willing to invest time and effort to reciprocate an unfair action by another country.

If a country puts Lurana in a difficult position, Lurana would do the same to that country.

Attention Checks *[Security scenario respondents]*

What problem has Lurana recently experienced?

- (A) **Cyberattacks targeting infrastructure** [correct]
- (B) Military conflict with a neighboring country
- (C) A public health crisis

Do the cyberattacks affect the United States directly?

- (A) **No** [correct]
- (B) Yes
- (C) This information was not provided

What type of support did Lurana request from the United States?

- (A) **Technical information on preventing cyberattacks** [correct]
- (B) Military troops
- (C) Financial aid

From the US perspective, what impact would sharing this information have?

- (A) **It is not expected to help or hurt the United States** [correct]

- (B) It would significantly help the United States
- (C) It would significantly hurt the United States

[Trade scenario respondents]

What economic issue is Lurana facing?

- (A) **Declining demand in its textile industry** [correct]
- (B) Rising oil prices
- (C) A shortage of skilled workers

Do these economic challenges affect the United States directly?

- (A) **No** [correct]
- (B) Yes
- (C) This information was not provided

What did Lurana request from the United States?

- (A) **Allow more clothing products to be sold in the US** [correct]
- (B) Direct financial assistance
- (C) Access to new technology

From the US perspective, what impact would the agreement have?

- (A) **It is not expected to help or hurt the United States** [correct]
- (B) It would significantly help the United States
- (C) It would significantly hurt the United States

[All respondents]

Did this scenario with Lurana bring any specific events to mind?

Yes No I don't know

[If yes] Would you like to explain what came to mind? *[Yes (open text) / No]*

Regression Tables

Tables 2-4 below depict the main, interaction, and supplementary models used for analysis.

Table 3.2: Main Models: Effect of Treatment on Policy Support and Trust

	<i>Dependent variable:</i>	
	Policy Support	Trust
	(1)	(2)
Ingroup	0.50*** (0.13)	0.86*** (0.11)
Outgroup	-0.61*** (0.13)	-0.87*** (0.11)
Violation	-3.50*** (0.13)	-3.88*** (0.11)
Negative	-1.87*** (0.13)	-1.90*** (0.11)
Security Scenario	0.27** (0.10)	0.23* (0.09)
Age	0.002 (0.003)	0.01* (0.003)
Female	-0.06 (0.10)	-0.18 (0.09)
Party ID	0.27 (0.21)	0.18 (0.19)
Ideology	-0.55 (0.30)	-0.10 (0.27)
Militant Intern.	-0.73* (0.33)	-0.89** (0.30)
Cooperative Intern.	3.93*** (0.34)	3.05*** (0.31)
Isolationism	-2.11*** (0.26)	-1.29*** (0.23)
Constant	7.24*** (0.36)	6.64*** (0.32)
Observations	1,722	1,722
R ²	0.41	0.48
Adjusted R ²	0.40	0.48
Residual Std. Error (df = 1709)	2.14	1.92
F Statistic (df = 12; 1709)	97.59***	133.63***

Note:

*p<0.05; **p<0.01; ***p<0.001

Table 3.3: Interaction Models: Policy Support and Trust

	<i>Dependent variable:</i>	
	Policy Support	Trust
	(1)	(2)
Ingroup	0.35 (0.22)	0.59** (0.19)
Outgroup	-0.37 (0.22)	-0.64*** (0.19)
Violation	-3.68*** (0.22)	-4.19*** (0.19)
Negative	-1.59*** (0.22)	-1.62*** (0.19)
Ingroup $\times$ Violation	0.19 (0.31)	0.39 (0.27)
Outgroup $\times$ Violation	0.32 (0.31)	0.53 (0.27)
Ingroup $\times$ Negative	0.23 (0.31)	0.39 (0.27)
Outgroup $\times$ Negative	-1.06*** (0.31)	-1.23*** (0.27)
Security Scenario	0.27** (0.10)	0.23* (0.09)
Age	0.001 (0.003)	0.005 (0.003)
Female	-0.03 (0.10)	-0.15 (0.09)
Party ID	0.25 (0.21)	0.16 (0.19)
Ideology	-0.53 (0.29)	-0.07 (0.26)
Militant Intern.	-0.70* (0.33)	-0.87** (0.29)
Cooperative Intern.	3.89*** (0.34)	3.00*** (0.30)
Isolationism	-2.12*** (0.25)	-1.31*** (0.23)
Constant	7.24*** (0.37)	6.71*** (0.33)
Observations	1,722	1,722
R ²	0.42	0.50
Adjusted R ²	0.41	0.50
Residual Std. Error (df = 1705)	2.12	1.89
F Statistic (df = 16; 1705)	76.47***	107.85***

Note:

*p<0.05; **p<0.01; ***p<0.001

Table 3.4: Effects of Treatment on Threat Perception (Supplementary)

	<i>Dependent variable:</i>
	Threat
Ingroup	−0.26* (0.12)
Outgroup	0.69*** (0.12)
Violation	2.20*** (0.12)
Negative	1.49*** (0.12)
Security Scenario	0.13 (0.09)
Age	−0.01* (0.003)
Female	0.26** (0.10)
Party ID	−0.49* (0.20)
Ideology	1.06*** (0.27)
Militant Intern.	1.77*** (0.30)
Cooperative Intern.	−0.49 (0.31)
Isolationism	1.29*** (0.23)
Constant	−0.38 (0.33)
Observations	1,722
R ²	0.26
Adjusted R ²	0.25
Residual Std. Error	1.96 (df = 1709)
F Statistic	49.64*** (df = 12; 1709)
<i>Note:</i>	*p<0.05; **p<0.01; ***p<0.001

Mediation Analysis Results

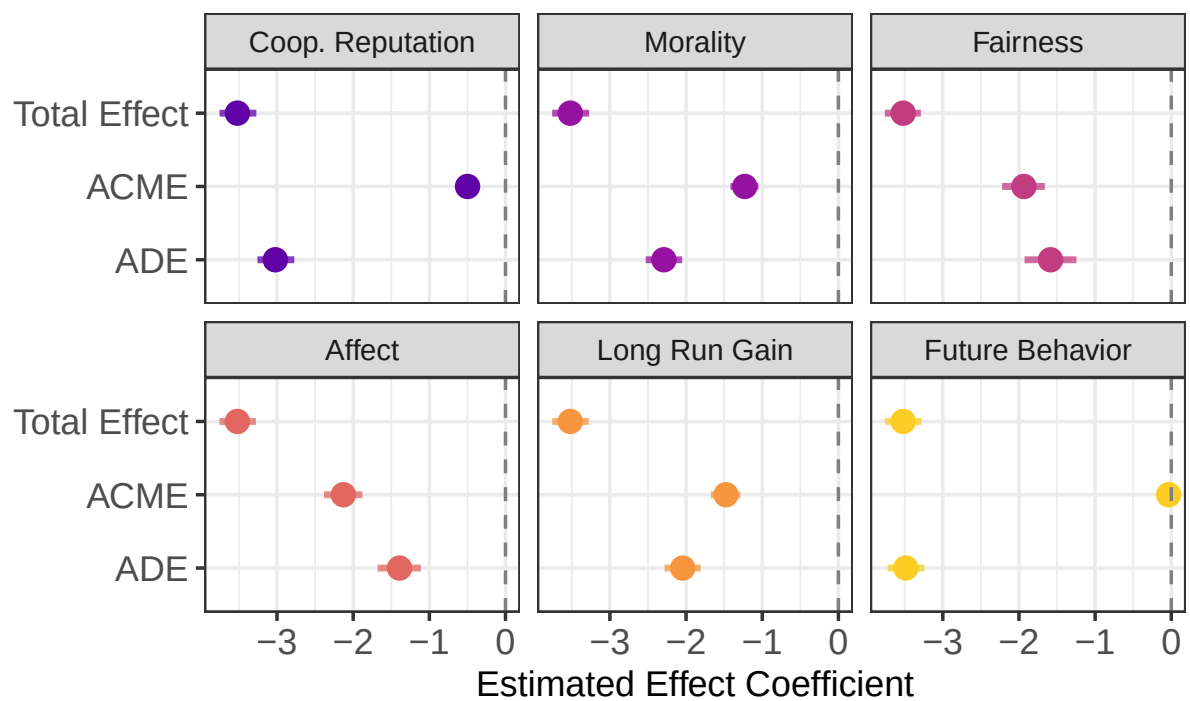
Regardless of support for BGR's predictions about group moderation, it appears people are at least relying on reciprocity when making judgments about international cooperation. But are people really using reciprocity as a mental shortcut, or shorthand strategy as BGR and other psychological approaches to reciprocity suggest? Or are they thinking about reciprocity in a more deliberative, instrumental and strategic manner? The nature of the design of this study limits our ability to really unpack these questions fully, but mediation analysis can give us an idea of what measured considerations are driving the reciprocity effect.

Mediation models (Tables 5:10) lend support for the argument that not only is reciprocity important for making cooperative decisions in international settings, but that reciprocity operates through many of the psychological mechanisms that theoretically drive the use of reciprocity as a heuristic: emotion, reputation, fairness, and moral judgments all significantly mediate the reciprocity effect across both outcomes.¹² Figure 4.10 presents mediation analysis results for the effect of a reciprocity violation on policy support. The total effect of the reciprocity violation on policy support is large and significant ($\hat{\tau} = -3.52, p < 0.001$), consistent with the main results. All six mediators show significant average causal mediation effects (ACME), though they vary considerably in the proportion of the total effect they account for. Affect (respondents' general feelings of warmth or coldness toward Lurana) is the strongest mediator, accounting for approximately 60% of the total effect ($\hat{\delta} = -2.13, p < 0.001$), followed closely by perceptions of fairness, which account for approximately 55% of the total effect ($\hat{\delta} = -1.94, p < 0.001$). Moral obligation accounts for approximately 35% of the total effect ($\hat{\delta} = -1.23, p < 0.001$), and long run instrumental gain accounts for approximately 42% ($\hat{\delta} = -1.48, p < 0.001$). Cooperative reputation concerns mediate a smaller but still significant share of the effect, accounting for approximately 14% ($\hat{\delta} = -0.50, p < 0.001$). Future behavior expectations account for the smallest share of mediation, with a statistically significant but substantively negligible ACME of approximately 1% of the total effect ($\hat{\delta} = -0.03, p = 0.008$). Taken together, these results suggest that the negative effect of a reciprocity violation on policy support operates largely through affective and normative channels, or how respondents

¹² Mediation models were estimated using nonparametric bootstrapping with 10,000 simulations.

feel about Lurana and whether they perceive its behavior as fair and morally appropriate.

Figure 3.6: Mediators of the Violation Reciprocity Effect on Policy Support



Note: Because each mediator is estimated independently rather than simultaneously, the proportions mediated sum to well over 100%. This is standard practice for this methodological approach, but reflects the fact that the mediators are correlated with one another.

Table 3.5: Reputation

Effect	Estimate	CI_Lower	CI_Upper	p_value
ACME	-0.50	-0.64	-0.37	0.00
ADE	-3.02	-3.26	-2.77	0.00
Total Effect	-3.52	-3.76	-3.27	0.00
Prop. Mediated	0.14	0.11	0.18	0.00

Table 3.6: Morality

Effect	Estimate	CI_Lower	CI_Upper	p_value
ACME	-1.23	-1.42	-1.05	0.00
ADE	-2.29	-2.53	-2.05	0.00
Total Effect	-3.52	-3.76	-3.27	0.00
Prop. Mediated	0.35	0.30	0.40	0.00

Table 3.7: Fairness

Effect	Estimate	CI_Lower	CI_Upper	p_value
ACME	-1.93	-2.22	-1.66	0.00
ADE	-1.59	-1.93	-1.24	0.00
Total Effect	-3.52	-3.76	-3.29	0.00
Prop. Mediated	0.55	0.47	0.64	0.00

Table 3.8: Affect

Effect	Estimate	CI_Lower	CI_Upper	p-value
ACME	-2.13	-2.38	-1.88	0.00
ADE	-1.39	-1.68	-1.11	0.00
Total Effect	-3.52	-3.76	-3.28	0.00
Prop. Mediated	0.60	0.54	0.68	0.00

Table 3.9: Long Run Gain

Effect	Estimate	CI_Lower	CI_Upper	p-value
ACME	-1.48	-1.67	-1.29	0.00
ADE	-2.04	-2.28	-1.81	0.00
Total Effect	-3.52	-3.76	-3.28	0.00
Prop. Mediated	0.42	0.37	0.47	0.00

Table 3.10: Future Behavior

Effect	Estimate	CI_Lower	CI_Upper	p-value
ACME	-0.03	-0.08	-0.01	0.01
ADE	-3.49	-3.72	-3.24	0.00
Total Effect	-3.52	-3.76	-3.28	0.00
Prop. Mediated	0.01	0.00	0.02	0.01

Bibliography

- Aldrich, John H, Christopher Gelpi, Peter Feaver, Jason Reifler and Kristin Thompson Sharp. 2006. "Foreign policy and the electoral connection." *Annu. Rev. Polit. Sci.* 9:477–502.
- Alexander, Richard. 2017. *The biology of moral systems*. Routledge.
- Aron, Arthur, Elaine N Aron and Danny Smollan. 1992. "Inclusion of other in the self scale and the structure of interpersonal closeness." *Journal of personality and social psychology* 63(4):596.
- Axelrod, Robert. 1980. "Effective choice in the prisoner's dilemma." *Journal of conflict resolution* 24(1):3–25.
- Axelrod, Robert. 2006. *The evolution of cooperation*. Revised ed. New York: Basic Books.
- Balliet, Daniel, Junhui Wu and Carsten KW De Dreu. 2014. "Ingroup favoritism in cooperation: a meta-analysis." *Psychological bulletin* 140(6):1556.
- Baum, Matthew A and Philip BK Potter. 2008. "The relationships between mass media, public opinion, and foreign policy: Toward a theoretical synthesis." *Annu. Rev. Polit. Sci.* 11:39–65.
- Berg, Joyce, John Dickhaut and Kevin McCabe. 1995. "Trust, reciprocity, and social history." *Games and economic behavior* 10(1):122–142.
- Bergé, Laurent R., Kyle Butts and Grant McDermott. 2026. "Fast and user-friendly econometrics estimations: The R package fixest." *arXiv preprint arXiv:2601.21749*.
- Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and

- Michael Ward. 2015. "ICEWS Coded Event Data." .
URL: <https://doi.org/10.7910/DVN/28075>
- Boyd, Robert and Peter J Richerson. 1988. "The evolution of reciprocity in sizable groups." *Journal of theoretical Biology* 132(3):337–356.
- Boyd, Robert and Peter J Richerson. 1989. "The evolution of indirect reciprocity." *Social networks* 11(3):213–236.
- Brandt, Hannelore, Hisashi Ohtsuki, Yoh Iwasa and Karl Sigmund. 2007. A survey of indirect reciprocity. In *Mathematics for ecology and environmental sciences*. Springer pp. 21–49.
- Brewer, Marilynn B and Linnda R Caporael. 2013. An evolutionary perspective on social identity: Revisiting groups. In *Evolution and social psychology*. Psychology Press pp. 143–161.
- Cameron, James E. 2004. "A three-factor model of social identity." *Self and identity* 3(3):239–262.
- Caporael, Linnda R. 1997. "The evolution of truly social cognition: The core configurations model." *Personality and Social Psychology Review* 1(4):276–298.
- Carlin, Ryan E and Gregory J Love. 2013. "The politics of interpersonal trust and reciprocity: An experimental approach." *Political behavior* 35(1):43–63.
- Carson, Austin. 2016. "Facing off and saving face: covert intervention and escalation management in the Korean War." *International Organization* 70(1):103–131.
- Carson, Austin and Keren Yarhi-Milo. 2017. "Covert communication: The intelligibility and credibility of signaling in secret." *Security Studies* 26(1):124–156.
- Chiba, Daina, Jesse C Johnson and Brett Ashley Leeds. 2015. "Careful commitments: Democratic states and alliance design." *The Journal of Politics* 77(4):968–982.
- Crescenzi, Mark JC, Rebecca H Best and Bo Ram Kwon. 2010. Reciprocity in international studies. In *Oxford Research Encyclopedia of International Studies*.
- Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes." *American political science review* 88(3):577–592.

- Fearon, James D. 1997a. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1):68–90.
- Fearon, James D. 1997b. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1):68–90.
- Fearon, James D. 1998. "Bargaining, enforcement, and international cooperation." *International organization* 52(2):269–305.
- Fehr, Ernst and Simon Gächter. 2002. "Altruistic punishment in humans." *Nature* 415(6868):137–140.
- Fehr, Ernst and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and human behavior* 25(2):63–87.
- Fehr, Ernst, Urs Fischbacher and Simon Gächter. 2002. "Strong reciprocity, human cooperation, and the enforcement of social norms." *Human nature* 13:1–25.
- Frank, Robert H. 2011. "The strategic role of the emotions." *Emotion Review* 3(3):252–254.
- Gartzke, Erik. 1998. "Kant we all just get along? Opportunity, willingness, and the origins of the democratic peace." *American Journal of Political Science* pp. 1–27.
- Gartzke, Erik and Jon R Lindsay. 2015. "Weaving tangled webs: offense, defense, and deception in cyberspace." *Security Studies* 24(2):316–348.
- Gartzke, Erik and Kristian Skrede Gleditsch. 2004. "Why democracies may actually be less reliable allies." *American Journal of Political Science* 48(4):775–795.
- Gartzke, Erik and Kristian Skrede Gleditsch. 2006. "Identity and conflict: Ties that bind and differences that divide." *European Journal of International Relations* 12(1):53–87.
- Gibbs, David N. 1995. "Secrecy and international relations." *Journal of Peace Research* 32(2):213–228.
- Goldstein, Joshua S. 1991. "Reciprocity in superpower relations: An empirical analysis." *International Studies Quarterly* 35(2):195–209.

- Goldstein, Joshua S and John R Freeman. 1990. *Three-way street: Strategic reciprocity in world politics*. Chicago: University of Chicago Press.
- Goldstein, Joshua S and John R Freeman. 1991. "US-Soviet-Chinese relations: Routine, reciprocity, or rational expectations?" *American Political Science Review* 85(1):17–35.
- Goldstein, Joshua S and Jon C Pevehouse. 1997. "Reciprocity, bullying, and international cooperation: Time-series analysis of the Bosnia conflict." *American Political Science Review* 91(3):515–529.
- Gray, Kurt, Adrian F Ward and Michael I Norton. 2014. "Paying it forward: generalized reciprocity and the limits of generosity." *Journal of experimental psychology: General* 143(1):247.
- Grieco, Joseph M. 1988. "Anarchy and the limits of cooperation: a realist critique of the newest liberal institutionalism." *International organization* 42(3):485–507.
- Gummerum, Michaela, Masanori Takezawa and Monika Keller. 2009. "The influence of social category and reciprocity on adults' and children's altruistic behavior." *Evolutionary Psychology* 7(2):147470490900700212.
- Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath et al. 2001. "Cooperation, reciprocity and punishment in fifteen small-scale societies." *American Economic Review* 91(2):73–78.
- Holsti, Ole R. 2004. *Public opinion and American foreign policy*. University of Michigan Press.
- Huth, Paul K. 1988. "Extended deterrence and the outbreak of war." *American Political Science Review* 82(2):423–443.
- Huth, Paul K. 1999. "Deterrence and international conflict: Empirical findings and theoretical debates." *Annual Review of Political Science* 2(1):25–48.
- Hyde, Susan D. 2015. "Experiments in international relations: Lab, survey, and field." *Annual Review of Political Science* 18:403–424.
- Imai, Kosuke and In Song Kim. 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis* 29(3):405–415.

- Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1):51 – 71.
URL: <https://doi.org/10.1214/10-STS321>
- Jervis, Robert. 1976. *Perception and misperception in international politics: New edition*. Princeton University Press.
- Jervis, Robert, Keren Yarhi-Milo and Don Casler. 2021. "Redefining the debate over reputation and credibility in international security: Promises and Limits of New Scholarship." *World Politics* 73(1):167–203.
- Keohane, Robert O. 1986. "Reciprocity in international relations." *International organization* 40(1):1–27.
- Kertzer, Joshua D. 2022. "Re-assessing elite-public gaps in political behavior." *American Journal of Political Science* 66(3):539–553.
- Keysar, Boaz, Benjamin A Converse, Jiunwen Wang and Nicholas Epley. 2008. "Reciprocity is not give and take: Asymmetric reciprocity to positive and negative acts." *Psychological Science* 19(12):1280–1286.
- Kydd, Andrew. 1997. "Game theory and the spiral model." *World Politics* 49(3):371–400.
- Kydd, Andrew. 2000. "Trust, reassurance, and cooperation." *International Organization* 54(2):325–357.
- Kydd, Andrew H. 2007. *Trust and mistrust in international relations*. Princeton, NJ: Princeton University Press.
- Larson, Deborah Welch. 1988. "The psychology of reciprocity in international relations." *Negot. J.* 4:281.
- Leach, Colin Wayne, Martijn Van Zomeren, Sven Zebel, Michael LW Vliek, Sjoerd F Pennekamp, Bertjan Doosje, Jaap W Ouwerkerk and Russell Spears. 2008. "Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification." *Journal of personality and social psychology* 95(1):144.

- Leng, Russell J. 1993a. *Interstate Crisis Behavior, 1816-1980*. Cambridge University Press.
- Leng, Russell J. 1993b. "Reciprocating influence strategies in interstate crisis bargaining." *Journal of Conflict Resolution* 37(1):3-41.
- Lindskold, Svenn. 1978. "Trust development, the GRIT proposal, and the effects of conciliatory acts on conflict and cooperation." *Psychological Bulletin* 85(4):772.
- Marcus, George E, Michael MacKuen, Jennifer Wolak and Luke Keele. 2006. The measure and mismeasure of emotion. In *Feeling politics: Emotion in political information processing*. Springer pp. 31-45.
- McDermott, Rose, Anthony C Lopez and Peter K Hatemi. 2017. "'Blunt Not the Heart, Enrage It': The Psychology of Revenge and Deterrence (November 2017)." *Texas National Security Review*.
- Mellers, Barbara, Alan Schwartz and Ilana Ritov. 1999. "Emotion-based choice." *Journal of experimental psychology: General* 128(3):332.
- Miller, Steven V. 2022. "peacesciencer: An R Package for Quantitative Peace Science Research." *Conflict Management and Peace Science* 39(6):755-779.
- Nowak, Martin A and Karl Sigmund. 2005. "Evolution of indirect reciprocity." *Nature* 437(7063):1291-1298.
- Nowak, Martin A and Sébastien Roch. 2007. "Upstream reciprocity and the evolution of gratitude." *Proceedings of the royal society B: Biological Sciences* 274(1610):605-610.
- Oneal, John R. and Bruce Russett. 2001. "Clear and Clean: The Fixed Effects of the Liberal Peace." *International Organization* 55(2):469-485.
- Osgood, Charles E. 1962. "An alternative to war or surrender."
- Oskamp, Stuart. 1971. "Effects of programmed strategies on cooperation in the Prisoner's Dilemma and other mixed-motive games." *Journal of Conflict Resolution* 15(2):225-259.
- Perugini, Marco, Marcello Gallucci, Fabio Presaghi and Anna Paola Ercolani. 2003. "The personal norm of reciprocity." *European Journal of Personality* 17(4):251-283.

- Poznansky, Michael and Evan Perkoski. 2018. "Rethinking secrecy in cyberspace: the politics of voluntary attribution." *Journal of Global Security Studies* 3(4):402–416.
- Rabin, Matthew et al. 1992. *Incorporating fairness into game theory and economics*. University of California at Berkeley, Department of Economics.
- Rajmaira, Sheen and Michael D Ward. 1990. "Evolving foreign policy norms: Reciprocity in the superpower triad." *International Studies Quarterly* 34(4):457–475.
- Rathbun, Brian C. 2011. *Trust in international cooperation: International security institutions, domestic politics and American multilateralism*. Vol. 121 Cambridge University Press.
- Rathbun, Brian C, Christopher Sebastian Parker and Caleb Pomeroy. 2025. "Separate but unequal: Ethnocentrism and racialization explain the "democratic" peace in public opinion." *American Political Science Review* 119(2):621–636.
- Rathbun, Brian C, Joshua D Kertzer, Jason Reifler, Paul Goren and Thomas J Scotto. 2016. "Taking foreign policy personally: Personal values and foreign policy attitudes." *International Studies Quarterly* 60(1):124–137.
- Reiter, Dan and Allan C Stam. 2002. *Democracies at war*. Princeton University Press.
- Renshon, Jonathan and Ryan Powers. 2024. "Identity and the Social Construction of Reputation in World Politics."
- Rhodes, Carolyn. 1989. "Reciprocity in trade: The utility of a bargaining strategy." *International Organization* 43(2):273–299.
- Romano, Angelo, Daniel Balliet and Junhui Wu. 2017. "Unbounded indirect reciprocity: Is reputation-based cooperation bounded by group membership?" *Journal of Experimental Social Psychology* 71:59–67.
- Roskos-Ewoldsen, Andrew, Morgan Ellithorpe and Brandon J Kinne. 2026. "The Effect of Reciprocity on Trust: International Cooperation and COVID Aid." *Journal of Conflict Resolution* 70(1):115–139.

- Russett, Bruce M and John R Oneal. 2001. "Triangulating peace: Democracy, interdependence, and international organizations." (*No Title*) .
- Sanger, David E. 2012. *Confront and conceal: Obama's secret wars and surprising use of American power*. Crown.
- Schelling, Thomas C. 2008. Arms and influence. In *Strategic Studies*. Routledge pp. 96–114.
- Schroeder, David A, Julie E Steel, Andria J Woodell and Alicia F Bembenek. 2003. "Justice within social dilemmas." *Personality and Social Psychology Review* 7(4):374–387.
- Shinada, Mizuho, Toshio Yamagishi and Yu Ohmura. 2004. "False friends are worse than bitter enemies: "Altruistic" punishment of in-group members." *Evolution and Human Behavior* 25(6):379–393.
- Spence, Michael. 1978. Job market signaling. In *Uncertainty in economics*. Elsevier pp. 281–306.
- Tajfel, Henri. 1970. "Experiments in intergroup discrimination." *Scientific american* 223(5):96–103.
- Tajfel, Henri. 1974. "Social identity and intergroup behaviour." *Social science information* 13(2):65–93.
- Tomz, Michael and Jessica LP Weeks. 2021. "Military alliances and public support for war." *International Studies Quarterly* 65(3):811–824.
- Tomz, Michael, Jessica LP Weeks and Keren Yarhi-Milo. 2020. "Public opinion and decisions about military force in democracies." *International Organization* 74(1):119–143.
- Tomz, Michael R and Jessica LP Weeks. 2013. "Public opinion and the democratic peace." *American political science review* 107(4):849–865.
- Uslaner, Eric M. 2002. *The Moral Foundations of Trust*. Cambridge University Press.
- Ward, Michael D and Sheen Rajmaira. 1992. "Reciprocity and norms in US-Soviet foreign policy." *Journal of Conflict Resolution* 36(2):342–368.
- Ward, Michael Don. 1982. "Cooperation and conflict in foreign policy behavior: Reaction and memory." *International Studies Quarterly* 26(1):87–126.

- Weisiger, Alex and Keren Yarhi-Milo. 2015. "Revisiting reputation: How past actions matter in international politics." *International Organization* 69(2):473–495.
- Wilson, Warner. 1971. "Reciprocation and other techniques for inducing cooperation in the Prisoner's Dilemma game." *Journal of Conflict Resolution* 15(2):167–195.
- Wu, Junhui, Daniel Balliet and Paul AM Van Lange. 2016. "Reputation, gossip, and human cooperation." *Social and Personality Psychology Compass* 10(6):350–364.
- Yamagishi, Toshio, Nobuhito Jin and Toko Kiyonari. 1999. "Bounded generalized reciprocity: ingroup boasting and ingroup favoritism. Adv." *Group Process* 16:161–197.
- Yamagishi, Toshio and Toko Kiyonari. 2000. "The group as the container of generalized reciprocity." *Social psychology quarterly* pp. 116–132.
- Yarhi-Milo, Keren, Joshua D Kertzer and Jonathan Renshon. 2018. "Tying hands, sinking costs, and leader attributes." *Journal of Conflict Resolution* 62(10):2150–2179.
- Zheng, Wenqian, Shengli Cao, Yangyang Wang, Ke Yang, Yunyun Chen and Guoping Song. 2021. "The impact of social value orientation, game context and trust on cooperative behavior after cooperative video game play." *Psychological Reports* 124(3):1353–1369.

Chapter 4

The Price of Retaliation: Understanding Costly Reciprocity in International Relations

The Price of Retaliation: Understanding Costly Reciprocity in International Relations

Andrew Roskos-Ewoldsen

Department of Political Science, University of California, Davis

Abstract

Countries often go out of their way to return favors or slights, even sometimes paying steep costs to do so. But when and why do individuals follow this logic of reciprocity? Returning insults or injuries in particular can be dangerous in international relations, especially in security situations, where the threat of “an eye for an eye leaves the whole world blind” looms large. This study explores the mechanisms and motivations behind the individual-level use of reciprocity, with a specific focus on reputation, costs, and audiences. I argue that people’s support for reciprocity is driven by a heuristic mental process, where we use reciprocity as a short hand strategy when interacting with other countries. I posit that people support the use of reciprocity for moral and emotional reasons as well as strategic ones such as reputational benefits; it feels good to “get even” because we feel it is the “right” or “fair” thing to do. We thus may disregard costs to reciprocate a negative act against our country, and tailor our behavior depending on the audiences that will see our reciprocal act. Utilizing a novel survey experiment on a representative sample of the US public, I find mixed evidence for my argument. The public more strongly opposes retaliatory reciprocity when the costs to do so are higher, but the type of reciprocity they support and the strength of this support depends on whether the counterpart country acted visibly to the rest of the world or not. These findings add nuance to our understanding of how reputational concerns drive reciprocity, and imply that other considerations like morality or emotion may also play a part how and when we use reciprocity in international security contexts.

Introduction

Why do countries seem to go out of their way to return a favor or slight? Why would people want to follow the logic of reciprocity, even in security situations, where the threat of “an eye for an eye leaves the whole world blind” looms large?

History tells us that there is precedent for reciprocating in security contexts, even if the costs are high. In 2015, Iranian General Qassem Suleimani met Vladimir Putin and pleaded with him to have Russia intervene in the ongoing Syrian civil war. Russia obliged, and in doing so helped stabilize a struggling Syrian regime that was one of Iran’s closest allies. In 2022, as Russia began suffering its own loss of troops, territory, and resources during the war with Ukraine, Iran repaid the favor and began supplying Russia with military supplies such as drones to help refuel their war efforts. This came with significant international and domestic costs, as siding with an increasingly isolated Russia only served to intensify an already contentious relationship with a world superpower in the United States, and led to 35 Iranian diplomats calling for Iran to reverse their decision and declare neutrality.¹ Sometimes reciprocating negative or aggressive acts leads to conflict; leading up to the Spanish-American War in 1898 that claimed over 2,000 lives, the US sent the battleship USS Maine to Cuba as a veiled threat against the Spanish. In return, Spain reciprocated by sending their own battleship *Viczaya* to New York, and ultimately the accidental sinking of the USS Maine triggered the start of the war.²

This study investigates whether individuals supports the use of reciprocity in security contexts, even when it is perceived to be costly to themselves or their country to do so, or when there is no global audience to see these reciprocal exchanges. This is informed by widely accepted approaches to reciprocity in international relations and psychology. Reciprocity, or the returning of a behavior, is prominent in international relations research on both cooperation and conflict (Crescenzi, Best and Kwon 2010; Goldstein and Pevehouse 1997; Keohane 1986), where reputation is a key mechanism for interactions between countries with long time horizons or when there are multiple countries

¹ See “Iran Arms Russia in the War in Ukraine” *The New Yorker*, 5 November 2022.

² See “Sailors, Soldiers, and Marines of the Spanish-American War” *Prologue Magazine*, Spring 1998, Vol. 30, No. 1.

involved. Additionally, costly signaling is one of the most prominent explanations for why states may pay high costs on foreign policy decisions, and the general logic is that by engaging in a costly act such as returning a favor or an aggression, you reveal credible information about your intentions (Fearon 1994, 1997*a*; Schelling 2008). In psychology, reputation also plays a key role in explaining why individuals use reciprocity across many different societal contexts; the theory of Bounded Generalized Reciprocity (BGR) for instance posits that people use reciprocity as a mental rule of thumb when interacting with others to strengthen our reputations as reliable cooperation partners, and thus gain cooperative benefits in the long term (Yamagishi, Jin and Kiyonari 1999).

Despite this considerable work on reciprocity, we lack micro-level explanations of how reciprocity works in international security contexts. To fill this gap, I build on this psychological approach to reciprocity, and synthesize the BGR approach with other insights from social and evolutionary psychology to argue for a more nuanced explanation of how reciprocity works at the individual level in international security situations. Prior research in evolutionary psychology theorizes that humans have evolved to use reciprocity behaviors as a basis for making cooperative decisions, and we do so because reciprocity is one of the most efficient strategies for overcoming barriers to cooperation and ultimately human survival (Boyd and Richerson 1988, 1989; Fehr and Fischbacher 2004; Fehr, Fischbacher and Gächter 2002). Essentially, our brains have several ways of encouraging us to use reciprocity in strategic interactions with other actors, and reciprocity functions as a sort of rule of thumb that guides our behavior. We have developed these mental mechanisms that drive reciprocity, like feeling positive emotions when we do the “right” thing by returning favors, because in the long run following this rule brings self-interest benefits from cooperation. Put differently, the logic of this synthesized BGR approach is simple and intuitive: returning a kind gesture with an equally kind gesture or returning an offense with an equal offense generally leaves us better off. In line with this approach, I argue that in international security contexts, people use reciprocity as a way to build a reputation as a strong cooperative partner, and thus will support using reciprocity even when in situations that are costly if it potentially improves our reputation with other countries. My theoretical expectations are thus that the public will support returning a negative reciprocal act against another country, even when it is costly to do so, but this support depends on whether other countries and audiences are able to witness the exchange and thus reap the reputational benefits

of using reciprocity.

I employ a novel experimental survey on a nationally representative sample of 1000 US adult citizens to test whether support for reciprocating an act to a counterpart country is influenced by the costliness or visibility of the reciprocity exchange. Due to research constraints, I restrict my analysis to negative reciprocity, or the exchanging of negative, aggressive security acts. Participants in the survey are given vignettes describing a security scenario based on real world events, where a counterpart country commits a negative, aggressive act against the United States. The sample is randomly assigned to “visibility” condition groups, where half of the participants are assigned to a “public” or “private” condition and are exposed to information detailing whether the counterpart country’s negative act was known to a wider global audience or covertly conveyed to the US. Respondents are then asked their level of support for three types of reciprocal responses: the US reciprocating the same type of negative act against the counterpart country either publicly, privately, or anonymously. Respondents are also at this stage randomly assigned to a “cost” condition group, either “high cost” or “low cost”, where I manipulate the costs of the reciprocal act.

The results indicate mixed support for my argument. I find that, contrary to my hypotheses, the public is sensitive to costs; people are more opposed to each type of negative reciprocity act when costs are high. However; I find that support for public, private, or anonymous negative reciprocity is at times dependent on the visibility of the reciprocity exchange: if the counterpart country committed the first negative act publicly versus privately, this increased individual support for a corresponding public act of negative reciprocity. This is complicated by the finding that the public/private visibility treatment was not a significant influence on support for the other two types of reciprocal responses. Finally, I use mediation analysis and find empirical support for my synthesized psychological approach to reciprocity; mechanisms like reputational concerns, considerations about long run gains, fairness, and moral obligation drive the effect of cost and visibility on public support for negative reciprocity. I take these findings as an indication that the way reciprocity, and the role of reputations, works is far more complicated at the individual-level than we have generally accounted for in international relations research. The public appears to

understand the logic of reciprocity and use it; they recognize that taking an “eye for an eye” is the “right” thing to do and can improve your reputation or leave you better off in the long run, but are still unwilling to pay costs to do so. Puzzlingly, people also seem only conditionally concerned about who is watching and the reputational consequences of reciprocity, indicating that further research should explore these boundary conditions.

In the following sections of the paper, I first review the literature on reciprocity and cooperation. I follow this by briefly defining and discussing indirect reciprocity for the purposes of this study. I then lay out my hypotheses and theoretical expectations for the effect of indirect reciprocity on public support for negative reciprocity. I then explain my research strategy and the design of my experimental survey. In the results section, I describe the main findings of my survey experiment. I then discuss and interpret these results, and provide a brief supplementary analysis on potential mechanisms. Finally, I conclude with a discussion of limitations and thoughts on the implications and future directions of my research.

Literature Review

What factors make the difference between taking an “eye for an eye”, or “turning the other cheek” in international relations? Returning a negative security act from another country with a negative security act of your own is a common occurrence in world politics, and reciprocity represents a key feature of international relations research on both cooperation and conflict (Goldstein and Pevehouse 1997; Keohane 1986). Prior work most commonly focuses on reciprocity as an instrumental strategy for countries to gain the benefits of cooperation; tit-for-tat, or players returning cooperation with cooperation and defection with defection (Axelrod 1980). In both formal and experimental approaches, reciprocity successfully induces cooperation from players in prisoner’s-dilemma-style settings (Axelrod 2006; Oskamp 1971; Wilson 1971). Goldstein and Pevehouse (1997) also highlights how reciprocity can be used as a strategy in instances of conflict, where the threat of conflict is tied to the prospects of cooperation with other actors.

Relatedly, costly signaling is one of the most prominent explanations for why states may return a

negative security act even when it is costly, and the general logic is that by engaging in a costly act such as returning a favor or an aggression, you reveal credible information about your intentions (Fearon 1994, 1997*a*; Schelling 2008). Deterrence theory in international relations has benefited substantially from this type of game theoretic, bargaining approach to conflict, where actions by a state reveal information about that state's willingness to fight (Fearon 1994, 1997*a*; Schelling 2008). Following the logic of costly signaling, states may commit an act that escalates the potential for conflict and would be politically costly to back down from in order to convince opponents that an attack would lead to a sub-optimal war, and thus deter opponents. Costly signals can thus be interpreted as a form of rational reciprocity; when a challenger country commits a threatening act as a signal of intentions, a defending state may engage in tit-for-tat returning of behaviors in order to deter their opponent (Leng 1993*b*). Observational studies on interstate deterrence show that costly signals and reciprocity often do deter opponents, although overly cautious or aggressive signaling can still lead to conflict (Huth 1988, 1999; Leng 1993*a*).

Likewise, in cooperation contexts and amidst the anarchy of international relations, talk is considered cheap. Political leaders may publicly profess an interest in cooperation while harboring exploitative intentions (Grieco 1988). From this rationalist perspective, players must acquire credible information about another's intentions in order to build trust and cooperate. Costly signaling is by far the most commonly studied means of conveying such information (Fearon 1997*b*; Spence 1978). In the context of international cooperation, costly signals are actions or policies that inflict nontrivial costs on the sender, such as yielding control of an important resource or making substantial concessions in a treaty negotiation (Fearon 1998; Kydd 2000). Such actions reveal information about the sender's trustworthiness because states that prefer exploitation should be unwilling to incur the costs of sending the cooperative signal. As players grow more confident in their assessments of one another's intentions, prospects for cooperation improve.

These approaches all rely heavily on reputations driving reciprocity at the state level; states behave in ways regarding security affairs, including returning favors or offenses, that are aimed at bolstering their reputation, either for cooperation or resolve. Yet, despite the insights offered by these rational approaches to deterrence and conflict, research thus far has not empirically disentangled

the individual-level motivations behind the use of reciprocity in interstate security contexts.

While state-level behavior is obviously important, it is worth understanding how real human beings use fundamental concepts like reciprocity by testing micro-level explanations. Representative studies on reciprocity often focus on the interactions between major powers, or at most the “triangular” interactions of three actors (Goldstein 1991; Goldstein and Freeman 1990, 1991; Rajmaira and Ward 1990; Rhodes 1989; Ward and Rajmaira 1992; Ward 1982). These studies find that almost all actors use reciprocity at rates higher than expected, but they tell us very little about why states, or people, actually reciprocate. These approaches to reciprocity also do not adequately account for seemingly irrational reciprocity-related behaviors, such as indirect/upstream reciprocity, or “paying it forward”, as opposed to behaviors towards a direct interaction partner in direct reciprocity (Gray, Ward and Norton 2014; Nowak and Sigmund 2005; Nowak and Roch 2007).

Focusing on the individual-level also offers the benefit of building on previous research on potential micro-level channels for how reciprocity affects state security decisions: first, public opinion has been shown under many conditions to be an important influence on elite and state foreign policy decision-making (Aldrich et al. 2006; Chiba, Johnson and Leeds 2015; Gartzke and Gleditsch 2004; Tomz, Weeks and Yarhi-Milo 2020); and second, recent research shows that while elites and the public often have very different attitudes and traits, the decision-making processes between elites and the public are oftentimes strikingly similar (Kertzer 2022; Yarhi-Milo, Kertzer and Renshon 2018). While I am agnostic about how and when individual decision-making ultimately influences state-level decisions, further research beyond the scope of this study may further investigate how these reciprocity-specific processes might occur.

A Psychological Approach to Reciprocity

My theoretical explanation of reciprocity in international relations is mainly derived from the theory of Bounded Generalized Reciprocity (BGR) in psychology (Yamagishi, Jin and Kiyonari 1999), which maps out the conditions under which people use reciprocity generally. I synthesize this approach with research from social and evolutionary psychology positing that humans have

evolved to use reciprocity behaviors as a basis for making cooperative decisions, and we do so because reciprocity is one of the most efficient strategies for overcoming barriers to cooperation and ultimately human survival (Boyd and Richerson 1988, 1989; Fehr and Fischbacher 2004; Fehr, Fischbacher and Gächter 2002). Essentially, these bodies of work tell us that our brains have several ways of encouraging us to use reciprocity in strategic interactions with other actors, and reciprocity functions as a sort of rule of thumb that guides our behavior. We have developed these mental shortcuts to reciprocity because in the long run following this rule brings self-interest benefits from cooperation. The strategic motivation behind this psychological approach is simple and intuitive: returning kindness for another kindness or returning an insult with another insult generally leaves us better off. We are mentally wired to trust in the logic that “what goes around, comes around”, either through direct tit-for-tat or through our reputation when interacting with others, and following the rules of reciprocity will be a net benefit over the course of our interactions with other actors.

So how do we gain self-interest benefits through reciprocity? In addition to the effectiveness of tit-for-tat reciprocity between two actors (Axelrod 1980), studies have shown that potential interaction with other actors is sufficient to create incentives for “downstream”, indirect reciprocity, or returning favors or slights on behalf of others. According to research in evolutionary psychology, we need this system of reciprocity for survival. As humans, mutual cooperation in social and economic exchanges that involve large groups is essential, but difficult because of the ubiquity of material incentives to cheat our exchange partners (Shinada, Yamagishi and Ohmura 2004). Scholars argue that indirect reciprocity, or third parties rewarding (or punishing) good or (bad behavior), leads to almost universal cooperation in situations where selfish, egoist behavior would cause a cooperation breakdown and where direct reciprocity alone is insufficient to maintain cooperation (Balliet, Wu and De Dreu 2014; Boyd and Richerson 1988; Fehr and Fischbacher 2004; Romano, Balliet and Wu 2017). People use “indirect” reciprocity as third parties, or when they are not directly involved in an interaction, and “punish” or “reward” actors who violate or follow reciprocity. Thus, people have an incentive to follow the rules of reciprocity by returning favors or aggressions. In the realm of world politics and international security, this means that individuals may also support following the rules of reciprocity given that their country faces many potential cooperative or conflictual

partners over long time horizons. At the state level, countries indeed routinely match and return the behaviors of other countries, and these reciprocal acts have had far reaching implications for not only the states involved but also for other countries witnessing the reciprocity. In the Falklands War between Great Britain and Argentina over the Falkland Islands, for instance, Britain reciprocated Argentina's invasion of the islands with their own military retaliation, which not only answered questions about whether Britain would return the act in kind to Argentina directly but also acted as an indirect warning to other status-quo challenging countries in the region. Likewise, the constant trading of threats and provocations between North and South Korea has had a range of effects on how other countries like the United States, Russia, and China act towards both countries.

A crucial aspect of this theoretical approach is that the use of reciprocity is based on heuristics or automatic psychological processes as opposed to strictly rational evaluations of payoffs, and thus at the individual level studies consistently find that people use reciprocity in ways and situations that would not be expected or explained sufficiently by formal or rational actor models. Research using BGR for instance has demonstrated that people are more likely to behave cooperatively (i.e., be kind, allocate more rewards) towards ingroup members than outgroup members or strangers, even when it is not welfare-improving to do so (Balliet, Wu and De Dreu 2014; Tajfel 1974; Yamagishi and Kiyonari 2000). Other studies show that people enforce reciprocity as third party punishers even when it is costly to themselves and there is no welfare-improving reward for doing so (Fehr and Gächter 2002; Fehr and Fischbacher 2004; Fehr, Fischbacher and Gächter 2002). These studies find that there is an indirect reciprocity effect on behavior for both positive and negative actions, and across a wide variety of issue domains, but as of yet no research investigates whether this approach also applies to international security contexts.

Even if reciprocity is a strong strategy for attaining welfare-improving benefits and survival, what mental processes actually encourage the use of reciprocity? Based on this theoretical approach, I offer four potential mechanisms for how indirect reciprocity might facilitate the effect of indirect reciprocity on individual support for cooperation with another country: reputational concerns, correcting bad behavior, fairness/morality concerns, and emotion.

The first mechanism is based on the idea that reputations drive reciprocity behavior in human beings (Romano, Balliet and Wu 2017; Yamagishi, Jin and Kiyonari 1999). Whether in direct reciprocity, by returning favors or slights, or indirect reciprocity, by punishing or rewarding others as a third party, we are signaling to other potential cooperative partners that we are ideal candidates for cooperation. We follow the rules of reciprocity and thus others can be sure that they can rely on us to return favors or cooperate without cheating or exploitation, and we also credibly convey that we will retaliate if someone tries to cheat us or others, which is crucial to maintaining a system of cooperative benefits without a higher authority. Punishing or retaliating towards others who have attacked us thus shows a commitment to reciprocity. In experimental and formal games, merely making individuals aware that their behavior is being monitored by others is enough to trigger reputational concerns and higher use of third party, indirect reciprocity (Brandt et al. 2007; Romano, Balliet and Wu 2017; Wu, Balliet and Van Lange 2016). People are even willing to punish as third parties under these conditions when there is little to no prospect of future interactions with their observers, and are willing to incur costs to themselves in order to punish or reward others' behavior (Henrich et al. 2001). Put another way, when faced with complex strategic dilemmas with other actors, as human beings we have a tendency to simplify the interaction by using reciprocity as a heuristic about reputation: following the rule of reciprocity conveys our reputation to others and thus leaves us better off.

A second mechanism is the desire to correct bad behavior from potential future partners. By retaliating with a negative act against another country, for instance, we might incentivize that country and other observers to avoid violations in the future, out of fear that they may also receive retaliations for negative acts. Thus, we are more likely to benefit from cooperation down the road. Romano, Balliet and Wu (2017) demonstrate this empirically in experimental economic games, showing that people punish as a third party when there is an expectancy of future interaction with the two other parties. This is another way reciprocity may function as a strategic heuristic, telling us that our cooperative behavior can help us down the line, albeit without an exhaustive calculation of how and when.

Concerns about morality and fairness might also induce us to use reciprocity; we often feel morally

compelled to take an "eye for an eye" or "follow the golden rule", thinking that it is the "right" or "fair" thing to do. Fehr, Fischbacher and Gächter (2002) find in their study on reciprocity that people cooperate or not based on how "fairly" they are being treated, and are willing to punish as a third party on behalf of others who are treated unfairly. This fairness motive is strong enough that people even show a willingness to sacrifice resources to do this; they are willing to pay costs to punish unfairness or accept that doing so will bring no material rewards. Prior work has also demonstrated how fairness and morality concerns can be rooted in a self-serving purpose consistent with the human goal of accruing cooperation benefits: doing the "right" thing or adhering to moral rules may lead others to give more than they otherwise would, for instance, or eliminate cheaters of reciprocity by outing them as immoral (Alexander 2017). We thus have theoretically integrated considerations of fairness and morality into our use of reciprocity, and there are formal models that demonstrate how actors can incorporate fairness into their behavior at equilibrium when interacting with others (Rabin et al. 1992).

Finally, we may have automatic, emotional reactions that may trigger us to use reciprocity without full, conscious deliberation. We have all likely experienced an angry need to "get even" to someone who has harmed us first, or felt the satisfaction and joy that comes when we "pay it forward" to someone else. As human beings, emotions have the ability to induce us into certain behaviors. Fehr and Fischbacher (2004) find in their study that in one-shot prisoner's dilemma and dictator games, third party observers whose economic payoffs were unaffected by the games consistently punished those who violated reciprocity, even if it was costly to themselves. They attribute these results to self-reported emotions and personal values, arguing that the results are consistent with emotional and moral motives for the use of reciprocity. This conclusion is based on work in evolutionary game theory and psychology, where emotions like anger, hate, or joy can benefit human survival and self-interest benefits by inducing us to behave in ways that a conscious, selfish, rational actor would not (Frank 2011). Behavior based on emotion can theoretically lend credibility to commitments or threats that would otherwise not appear legitimate, like revenge in international relations (McDermott, Lopez and Hatemi 2017). In the case of reciprocity in security contexts, this might make our commitment to reciprocity and returning favors and offenses credible, even when it might not appear rational to do so, and thus increase the chances of cooperation with others. It also

might make our threats to retaliate or punish as a third party more credible.

These are familiar concepts in international relations, where states and leaders commonly use the language of these mechanisms in justifying various foreign policy decisions. Countries are constantly concerned about their reputation with other countries, for instance, and often cite reputation as a motive for their actions (Jervis, Yarhi-Milo and Casler 2021; Weisiger and Yarhi-Milo 2015). Former Prime Minister of Great Britain Winston Churchill once famously said: “We do not covet anything from any nation, except their respect.”³ Countries also commonly abandon opportunities for cooperation in the name of correcting bad behavior of other states, like with the United States’ sanctioning of Iran or North Korea for their nuclear arms ambitions. States also often emphasize morality and fairness in their actions, claiming that doing the “right” thing guides foreign policy as much as interests do. Fittingly, former US president Barack Obama once declared during a speech at the United Nations: “I believe that America is exceptional – in part because we have shown a willingness, through the sacrifice of blood and treasure, to stand up not only for our own narrow self-interest but for the interests of all.”⁴ Emotion is also apparent in world affairs, like in instances of negative reciprocity, or taking an “eye for an eye”. In the aftermath of a terrorist attack on India in 2019, for instance, Indian Prime Minister Narendra Modi expressed a deep-seated emotional response: “The amount of anger that is there in the country, people’s blood is boiling. I can very well understand this. I want to tell the terror outfits and their patrons that they have committed a big mistake. They have to pay a heavy price for this.”⁵ Research on public opinion has also produced evidence that some of these mechanisms affect public support for foreign policies, like going to war. Tomz, Weeks and Yarhi-Milo (2020) find that reputational concerns and personal commitments to fairness mediated public support for going to war on behalf of military allies.

Based on this psychologically-informed, heuristic approach to reciprocity, I expect that individuals favor returning a security-related behavior to other countries, regardless of cost to themselves or their country. I argue that this is because reciprocity, as a heuristic guide to behavior, is theoretically at least partially an automatic, emotional process as well as a moral one, and thus we

³ “Vive la France! SLEEP TO GATHER STRENGTH FOR THE MORNING.” 19, October, 1940, Retrieved from *International Churchill Society*.

⁴ See: “Obama tells the world: America is exceptional” *USA Today*, 24 September 2013.

⁵ See: “India’s blood is boiling, says Modi” *Telegraph India*, 15 February 2019.

tend to ignore information and considerations about costs.

Hypothesis 1: *Individual support for a reciprocal act will not be significantly different when the costs of that act are higher.*

I also consider how different types of security acts can potentially reveal the motivations behind the use of reciprocity at the individual level. Specifically, I examine how the visibility of security related behaviors can potentially affect how individuals use reciprocity. This is based on the tradition of manipulating the visibility of reciprocal acts in laboratory experimental studies on reciprocity, where the presence of an audience can influence reciprocal behavior (Balliet, Wu and De Dreu 2014; Fehr and Gächter 2002). States have also long engaged in covert and clandestine security actions like limited or secret wars, and more recently through cyber warfare. Of course, states also publicly and privately claim credit for security acts, whether they were originally secret or not. The reasons states do so are well documented, such as for gaining tactical advantages, avoiding the escalation of conflicts, or sidestepping negative reactions from domestic and international observers for security decisions (Carson 2016; Carson and Yarhi-Milo 2017; Gartzke and Lindsay 2015; Gibbs 1995; Poznansky and Perkoski 2018). For the purposes of this study, the visibility of a security act has important implications for parsing whether people reciprocate for reputational reasons, as expected in canonical theories of reciprocity in IR and in some psychological approaches to reciprocity like BGR. In particular, people may want to visibly retaliate against another country when the other country's offense was highly visible to strengthen their reputation for following reciprocity, but also may want to avoid being the first to reveal responsibility for a negative, reciprocal act out of fear that they will be seen as the aggressor/violator of reciprocity by other observers and damage this reputation. In order to show others that you are willing to do the "right" or "fair" thing, or that you are a "good" practitioner of reciprocity, you need to be aware of your audience. Additionally, since returning a security act anonymously would not affect the reputation of an individual's country to the other country or other actors, individuals should be ambivalent about support for this type of reciprocity. I thus expect that because reciprocity is theoretically at least partially tied to these reputational concerns, individuals should be sensitive to whether others can witness the original and returning acts in a reciprocity interaction. I thus develop the following

hypotheses:

Hypothesis 2: *Individual support for a public reciprocal act will be higher for individuals exposed to a public, initial act by another country compared to individuals exposed to a private negative act.*

Hypothesis 3: *Individual support for a private reciprocal act will be lower for individuals exposed to a private, initial act by another country compared to individuals exposed to a public, initial act.*

Hypothesis 4: *Individual support for an anonymous reciprocal act will not be different for individuals exposed to a private, initial act by another country compared to individuals exposed to a public, initial act.*

Research Design

To test my hypotheses, I employ a novel survey experiment measuring individual-level support for public, private, and anonymous reciprocal acts towards another country in a hypothetical scenario. Experiments based on the public have several limitations, such as differences between ordinary citizens and leaders/state decision-makers or incoherent attitudes on foreign policy (Baum and Potter 2008; Hyde 2015). However, as an initial test of this micro-level approach to reciprocity, experiments offer comparative advantages to other methods for exploring the relationship between the main variables of interest. By manipulating precisely individual exposure to the visibility and cost of reciprocity, and measuring individual support for retaliatory reciprocity, I can cleanly observe a causal effect should it exist as theorized.

Experiment and Treatments

The survey experiment was conducted in the summer of 2024 on a nationally representative sample of 1000 US adults, recruited through the survey company Prolific.⁶ The experiment follows a 2 x 2 factorial, between-subjects design. Before treatment, I collected participant information on several demographic and attitudinal variables (gender, age, education, party identification, political orientation, foreign policy orientation,) that might interact with my variables of interest or directly affect support for reciprocity.⁷

Following this, in a news article vignette, I give respondents a hypothetical international relations scenario based on a real-world example of a negative security act that can be reciprocated.⁸ In these scenarios, a counterpart country (Country A) is revealed to commit a negative, aggressive act towards the respondents' country in the form of a cyberattack.

The United States and Country A have recently been in a disagreement over a security issue, and tensions have grown between them.

Recently, the United States discovered a cyberattack on its military computers. This attack, called a Distributed Denial of Service (DDoS), involved many compromised computers working together to overload the US military computers with information requests, causing them to shut down for several hours. While the servers have now been secured again, military personnel were temporarily unable to access important US databases.

The respondents are also shown information about whether Country A claimed credit for the cyberattack publicly or privately, with half of the sample randomly assigned to “public” or “private” visibility treatment condition groups.

Public Visibility Treatment:

⁶ An a priori power analysis, based on results from a pilot study of 415 respondents, determined that a sample size of at least 320 would be sufficient to detect an effect at the standard power benchmark of .80.

⁷ Foreign policy orientation questions were taken from (Rathbun et al. 2016).

⁸ Based on cyberattacks by Syrian Electronic Army on the US military. See <https://www.forbes.com/sites/katevinton/2015/06/08/syrian-electronic-army-claims-responsibility-for-hacking-army-website/?sh=7476ba55197f> for link.

The cyberattack has become public knowledge to the rest of the world. Country A has publicly revealed they were responsible for the attack at a press conference.

Private Visibility Treatment:

Countries sometimes act in secret, and do not reveal their actions to the rest of the world. Due to the private nature of cyberattacks, the attack is still a secret from other countries. However, Country A has privately revealed to the United States they were responsible for the attack through diplomatic back channels.

Respondents are then shown three types of reciprocal responses the US government is considering in random order: the US reciprocating the same type of negative act against the counterpart country either publicly, privately, or anonymously. Respondents are also at this stage randomly assigned to a “cost” condition group, either “high cost” or “low cost”, where I manipulate the costs of the reciprocal act.

Low Cost: National security experts agree that returning a cyberattack to Country A will not be very costly to the US. The US has the resources and technology to return a cyberattack with relative ease.

High Cost: Experts agree that returning a cyberattack to Country A will be very costly to the US. The US government spends \$13 billion a year on cybersecurity. Once a cyberattack is used, the target learns how the attack works, and the attack cannot be used again. The US will lose the chance to use this attack in the future. Country A will also be better prepared for another attack, and the US will need to spend more money and develop new technologies to successfully launch cyber attacks in the future.

I manipulate visibility and cost for two main reasons. First, manipulating the costliness and visibility of a reciprocal act allows me to directly test the various theoretical mechanisms of reciprocity that might be responsible for reciprocity behaviors. More specifically, costliness and visibility speak directly to whether reputation is the driving mechanism behind the use of reciprocity in individuals. As discussed above, if reputation is driving the use of reciprocity at the individual-level, support for certain types of reciprocity should be influenced by the visibility of the original act. Similarly, according to the costly signaling approach, individuals should only prefer costly reciprocal acts

when these acts can credibly convey a reputation for resolve or reciprocity to others. Since talk is cheap, a private, costless reciprocity act should not be very appealing. Being able to look at cost and visibility should provide insights into the motivations behind the use of reciprocity at the individual level. Secondly, by manipulating the cyberattacks in this way, I can use more natural scenarios that map onto real-world events. Poznansky and Perkoski (2018) discuss the incentives states and non-state actors have for publically and privately claiming credit for cyber attacks, and there is notable evidence tracing why states consider private credit claiming of cyberattacks, such as the Stuxnet operation conducted by the United States and Israel to attack Iran nuclear facilities through cyberattacks. Sanger (2012) documents how US government officials considering privately informing the Iranians of their role in the attack in order to achieve concessions.

Measures, Outcomes, and Empirical Strategy

After reading the vignette, participants responded to the primary outcome measures of support for a reciprocal act towards Country A, as well as supplementary and manipulation check questions to assess whether respondents recognized and understood the content of the treatments. Participants were then debriefed at the end of the study about the purpose of the study and the nature of the manipulations.

The main outcomes of interest, support for public, private, and anonymous reciprocity towards Country A, were measured using an 11-point likert scale where participants rated how much they support or oppose the reciprocal act. The item order was randomized for all respondents. The scale ranged from 0 (strongly oppose) to 10 (strongly support), and was based on previous research on public support (Tomz, Weeks and Yarhi-Milo 2020).

Please indicate how much you support or oppose the responses the US is considering:

- The US returns a similar cyberattack to Country A and publicly takes credit through a press conference.
- The US returns a similar cyberattack to Country A and privately takes credit through a diplomatic communication with Country A.

- The US returns a similar cyberattack to Country A but does not take credit for the attack.
The US uses tactics that are difficult to trace and remains anonymous.

For supplementary analysis purposes, I also asked respondents questions that I could use to test whether the theoretical mechanisms specified by my psychological approach mediate the effect of cost or visibility on support for reciprocity towards Country A.⁹ For reputation, I ask respondents to what extent they agree that if the U.S. does not return a cyberattack to Country A, other countries will doubt the U.S.'s willingness to respond to challenges. For correcting bad behavior, participants rate how much they agree with the statement that "The U.S. should use a cyberattack to Country A to discourage other countries from challenging the U.S. in the future." Moral and fairness concerns were measured by asking respondents to rate how much they agreed that the US had a moral obligation to act to return a cyberattack to Country A or whether Country A "played fair" with other countries. Emotions are notoriously difficult to measure in a survey setting, but I include feeling thermometers (0-100 point scale, with 0 indicating "coldest" feelings and 100 being "warmest" towards Country A) as a basic measure, since they are a common method of approximating positive and negative emotion in political science and other fields like psychology (Marcus et al. 2006).

I use multiple linear regression to estimate the effect of the treatments on support for reciprocity, and also account for any variable imbalances in the condition groups. Covariates were normalized to range from 0 to 1 for ease of interpretation and comparison.

For supplementary analysis about what mechanisms may drive any cost or visibility treatment effects, I use mediation analysis (Imai, Keele and Yamamoto 2010) to estimate the Average Causal Mediation Effects (ACME) of each potential mechanism on the main outcome variable.¹⁰

⁹ Exact measurement procedures and survey instrumentation can be found in the appendix.

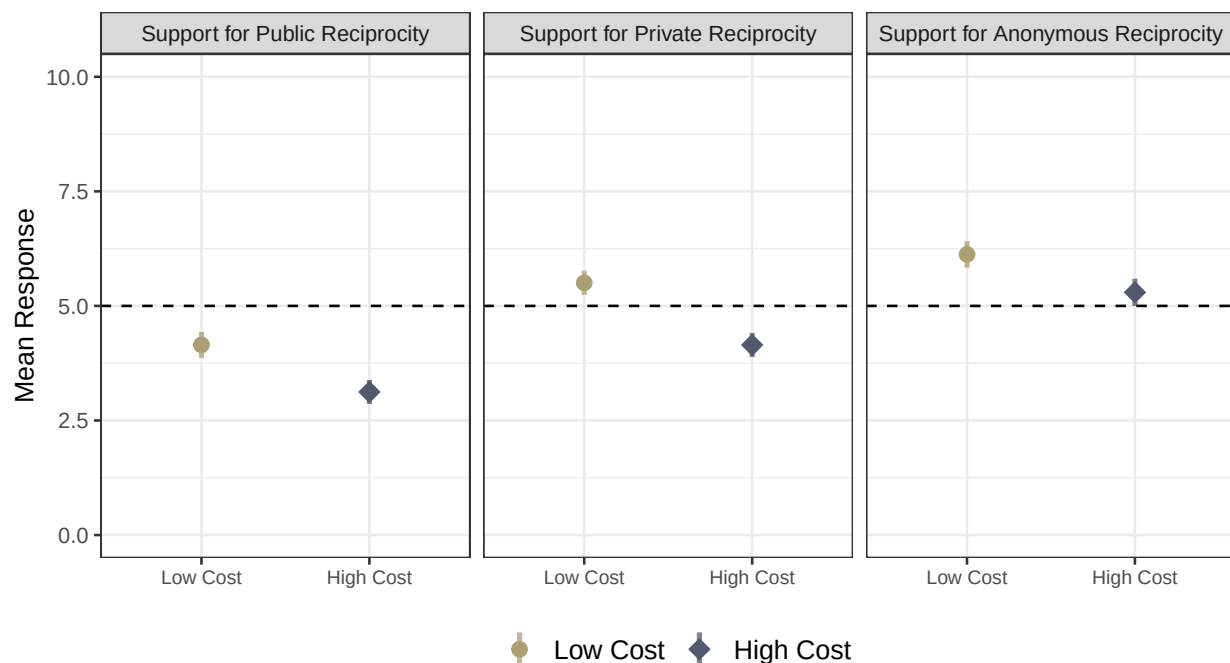
¹⁰ I use the Imai et al. procedure for mediation, where the effect of some X variable on Y is mediated by variable M through three steps. First, the effect of X on Y is estimated using a regression model. Second, the effect of X on M is estimated using another regression model. Finally, the outcome Y is regressed on both X and M. To test for significance, the data is resampled many times to create a multitude of simulations, otherwise known as bootstrapping. If there is an observed effect across these simulations, we can be confident that the effect is not by chance. I follow this process for each survey item related to these potential mechanisms.

Empirical Results

Effect of Costliness

A simple visual representation of my findings by cost treatment group are presented below in Figure 4.1, and offers a preliminary picture of what the cost effect might look like. Across the three dependent variable types, support was significantly lower in the high cost condition (4.2) than the low cost condition (5.3) on the 0-10 likert scale.

Figure 4.1: Support for Reciprocity by Cost Treatment Group + 95% Confidence Intervals

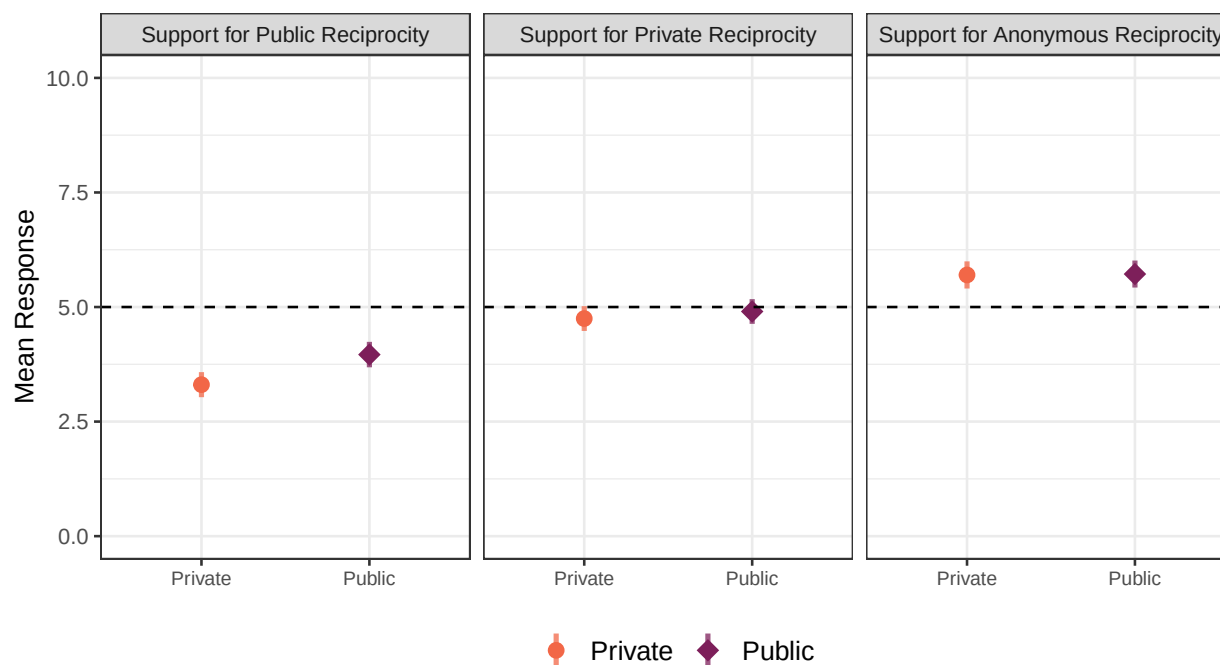


Note: The dashed line indicates the scale midpoint.

For the visibility treatment, Figure 4.2 displays the average level of support for each type of reciprocity response towards Country A. Only the publicly claimed US cyberattack retaliation had significantly different support based on whether an individual was in the private or public treatment group, with average support of 3.3 for respondents in the private condition and 4.0 in the public condition. The averages in the private and public conditions for support of the private and anonymous reciprocity options were not significantly different than zero. Across all three combined types of reciprocity, support was 4.6 and 4.9.

It is also worth noting that the averages for each type of support were significantly different according to t-tests, and that support for reciprocity rose as the US took less credit for the cyberattack. The average support for the publicly claimed cyberattack response was 3.6, the private reciprocity response to Country A was 4.8, and the anonymous reciprocity response was 5.7.

Figure 4.2: Support for Reciprocity by Group Treatment Group + 95% Confidence Intervals



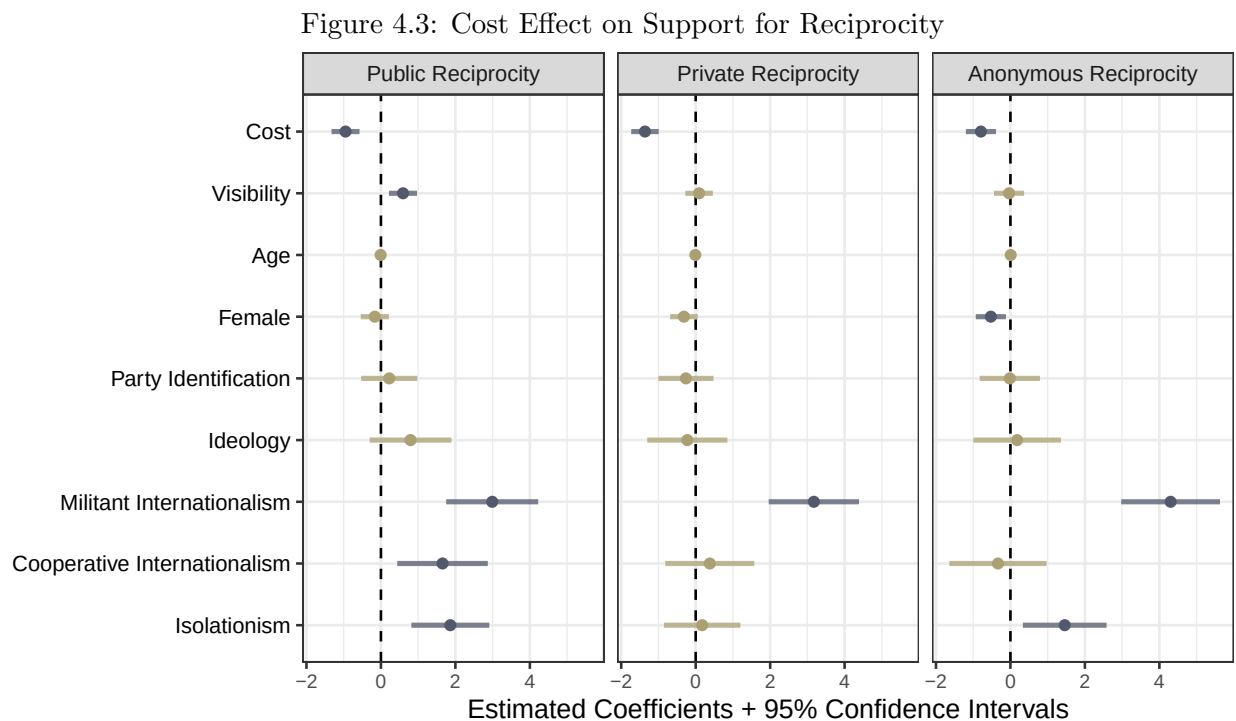
Note: The dashed line indicates the scale midpoint.

The Effects of Cost and Visibility on Reciprocity

Hypothesis 1 predicts that individuals will be no less supportive of a reciprocal cyberattack on Country A when they are exposed to information about the high costs of reciprocity as opposed to low costs. The findings strongly reject Hypothesis 1. Respondents in the high cost condition were significantly less likely to support a cyberattack retaliation against Country A than respondents in the low condition. Across regression models for all three types of support for reciprocity, exposure to information about the high cost of reciprocity significantly predicted a decrease in support for a cyberattack on Country A by over at least 0.8 points, and over a point on average on the 0-10 scale.

For visibility, hypotheses 2, 3, 4 predict that visibility will have a significant effect on support for the publicly and privately claimed reciprocity options towards Country A, but not on the anonymously delivered cyberattack option. I find support for hypotheses 2 and 4, but evidence to reject hypotheses 3. As expected, the regression model for support of the publicly claimed cyberattack against Country A predicts that individuals are 0.6 more likely to support this act when in the public condition group as opposed to the private. For the anonymous option, respondent support is not significantly influenced by visibility treatment group. Contrary to expectation; however, respondent support for the private reciprocity option is also not significantly influenced by the visibility treatment.

Figure 4.3 shows the estimated treatment effect on the three different measures of support for reciprocity towards Country A, as well as estimates for the covariates used in these regression models.¹¹



¹¹ These results remained consistent across all model specifications.

Conclusion

There are several substantive takeaways from these results. The first and most important is that the study demonstrates that people *do* care about costs when it comes to reciprocity. While not necessarily surprising, this is a substantial finding because there is a substantial body of work across different fields repeatedly demonstrating that people tend to ignore costs in order to use reciprocity. Relatedly, the mediation analysis indicates that costs in international security contexts may affect the key theoretical mechanisms that facilitate the use of reciprocity at the individual level, like reputational motivations, the desire to correct bad behavior, and moral/fairness concerns. Taking the results together, it is interesting from a costly signaling and reciprocity perspective that the public was generally more supportive of lower cost, smaller audience types of reciprocity, even after suffering an attack first, which comes closer to theoretical cheap talk than to credible signaling.

Second, evidence from this study indicates that the relationship between reciprocity and reputation is very complicated at the individual level and requires more careful theorizing and exploration. The evidence suggests that the reputational motivations for using reciprocity are significant but not straightforward, and that other considerations like morality are also playing a role. The public exhibits stronger support for having a public audience for negative reciprocity if they received a public attack first, although this does not appear to be driven by reputational concerns. It is also not clear why there is not a visibility effect on support for a private reciprocity response.

I hope this study represents an initial step in making reciprocity and its use in international relations more intuitive and easier to understand, and the theoretical approach and evidence presented here can help bridge an important gap between the analytical utility of canonical approaches to reciprocity in international relations and the plethora of evidence and perspectives from other fields showing that actors use reciprocity in strategic, but at times seemingly irrational ways. Reciprocity according to this theoretical approach is still strategic in its goal to accrue cooperation benefits and improve our chances of survival, but we have developed mental mechanisms beyond exhaustive calculation that facilitate this end. Using reciprocity as a heuristic might be more efficient in some instances, and under some conditions lead to unexpected behavior, but other times may lead us to

the same behaviors as predicted by formal or rational models.

This is also a potentially important finding when tied to the research that supports the idea that the public is important to a country's foreign policy choices. Leaders may be potentially influenced by how the public thinks (in addition to the possibility that leaders might also think in similar ways to ordinary citizens), or have incentives to "win over" or convince the public to back them, or even think and behave like the public themselves (Aldrich et al. 2006; Chiba, Johnson and Leeds 2015; Gartzke and Gleditsch 2004; Kertzer 2022; Tomz, Weeks and Yarhi-Milo 2020). Thus, knowing that citizens care about whether other countries follow the "golden rule" could potentially help us understand when and how cooperation occurs at the state-level too.

Finally, a substantive lesson for world politics may be that "cheap talk" is more appealing to the public than some types of wisdom might expect. Countries are often expected to take an "eye for an eye", but their people are not as willing to "leave the world blind" to do so as some might suggest, even if there are reputational or mental incentives to do so.

This study is not without limitations. First, as there are with all hypothetical experiments in international relations, there is the question of whether the research design accurately captures the most crucial aspects of the public opinion process. If people fundamentally think differently when using hypothetical scenarios with generic country names than when they are faced with these issues in the real world, the lessons from these types of studies are potentially flawed. While I tried to maximize the mundane realism of the experiment as much as possible, this is still a threat to the generalizability of the findings. Another limitation is the sparsity of the design. Due to resource constraints, many potential factors that might directly affect the outcomes, or interact with or mediate a cost or visibility effect were not included. I was not able to include examples of positive, security-related reciprocity scenarios in my design, for example. Future research should explore these possibilities, and more rigorously test the relationships between these variables.

Appendix

Foreign Policy Ideology as a Moderator

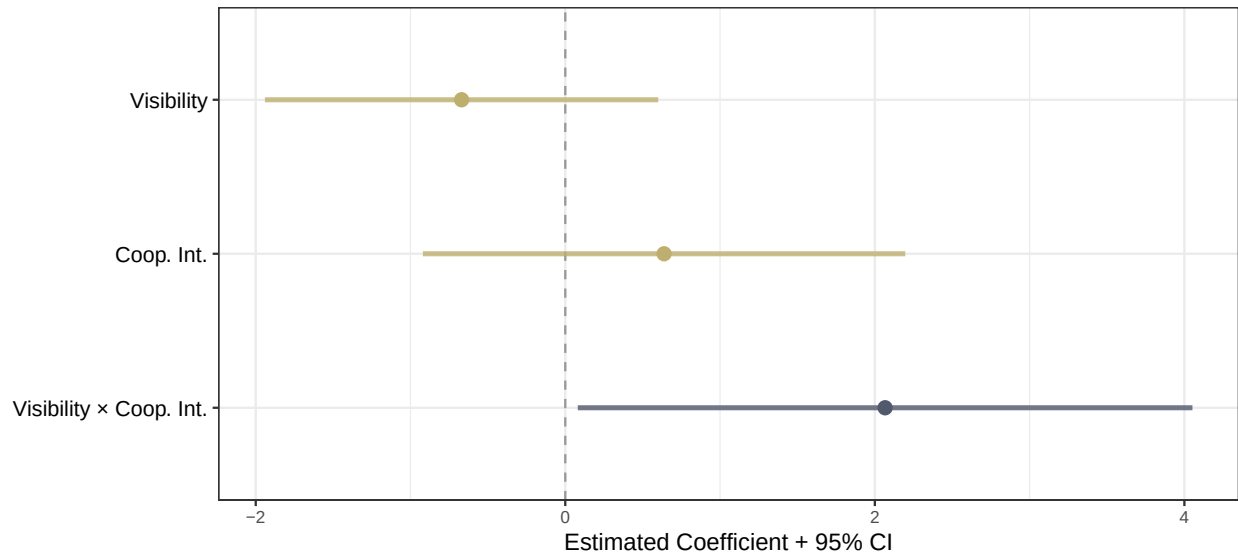
To assess whether pre-existing foreign policy orientations condition the treatment effects of visibility and costliness, I estimate a series of linear regression models that include interaction terms between each treatment and the three FPI factor scores: Militant Internationalism (hawkishness), Cooperative Internationalism (dovishness), and Isolationism. Each FPI score is normalized to range from 0 to 1. The subsections below discuss significant or near-significant interactions, organized by treatment variable. Full regression tables for all models with significant interactions are presented in Tables 4.1 and 4.2 at the end of this section. Each coefficient plot in this section displays only the treatment variable, the FPI moderator, and their interaction term (the variables of primary theoretical interest) with 95% confidence intervals. Darker points indicate statistically significant estimates ($p < .05$); lighter points are not significant.

FPI Moderating the Visibility Effect

For the visibility treatment, FPI is generally not a significant moderator, with a few interesting exceptions. First, for support for *public* reciprocity, dovishness (Cooperative Internationalism) significantly moderates the visibility effect: the more dovish a respondent, the more likely they were to support a publicly reciprocal act when Country A acted publicly. Figure 4.4 shows the estimated coefficients for this interaction. My initial interpretation is that doves, who value cooperation with others, are potentially sensitive to reputational concerns and appearing “tough but fair.”

For support for *private* reciprocity, both hawkishness and isolationism significantly moderate the visibility effect. The more hawkish or isolationist a respondent, the more supportive they were of private reciprocity when Country A acted publicly (Figures 4.5 and 4.6). This pattern does not hold for the other dependent variables, which makes it a somewhat puzzling finding. One possible interpretation is that hawks and isolationists are more willing to support any type of retaliatory response when the US is attacked, but are specifically more supportive of a lower-profile, private

Figure 4.4: Dovishness Moderates the Visibility Effect on Public Reciprocity
DV: Public Reciprocity



response when the attack itself was made publicly—perhaps because they view a private response as less diplomatically costly. Hawkishness also significantly moderates the visibility effect for support for *general* reciprocity (an average of the three types), as shown in Figure 4.7, though isolationism does not in this case.

Figure 4.5: Hawkishness Moderates the Visibility Effect on Private Reciprocity
DV: Private Reciprocity

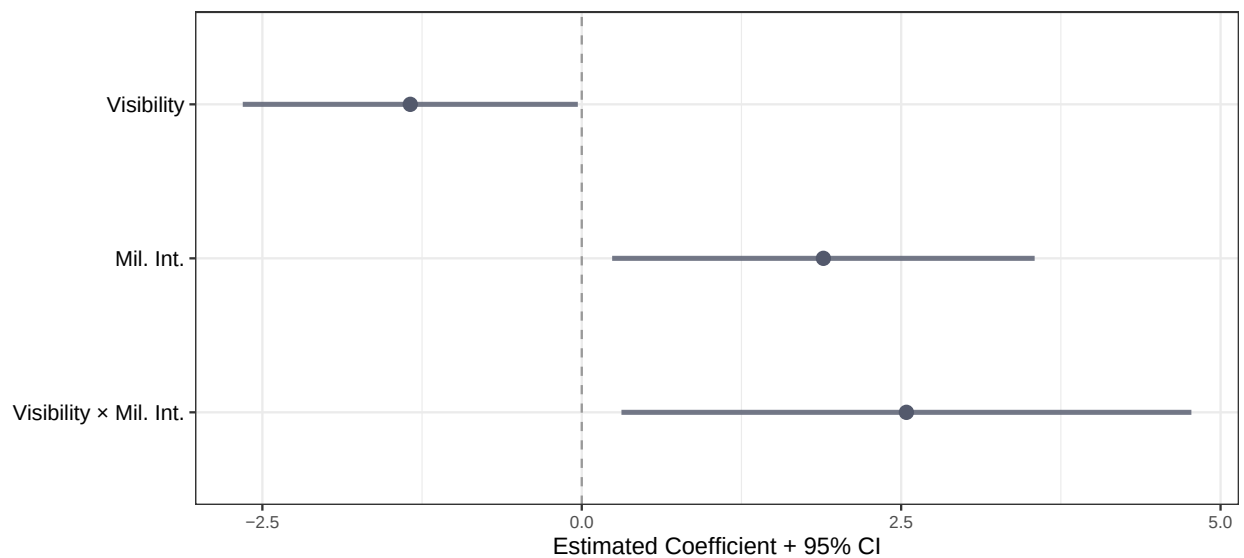


Figure 4.6: Isolationism Moderates the Visibility Effect on Private Reciprocity

DV: Private Reciprocity

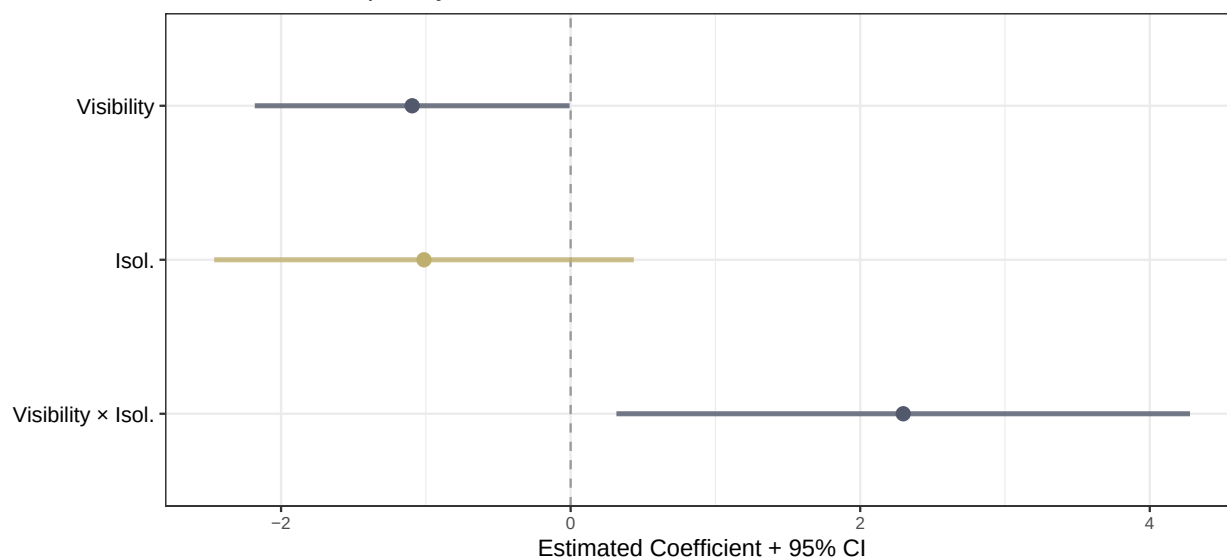
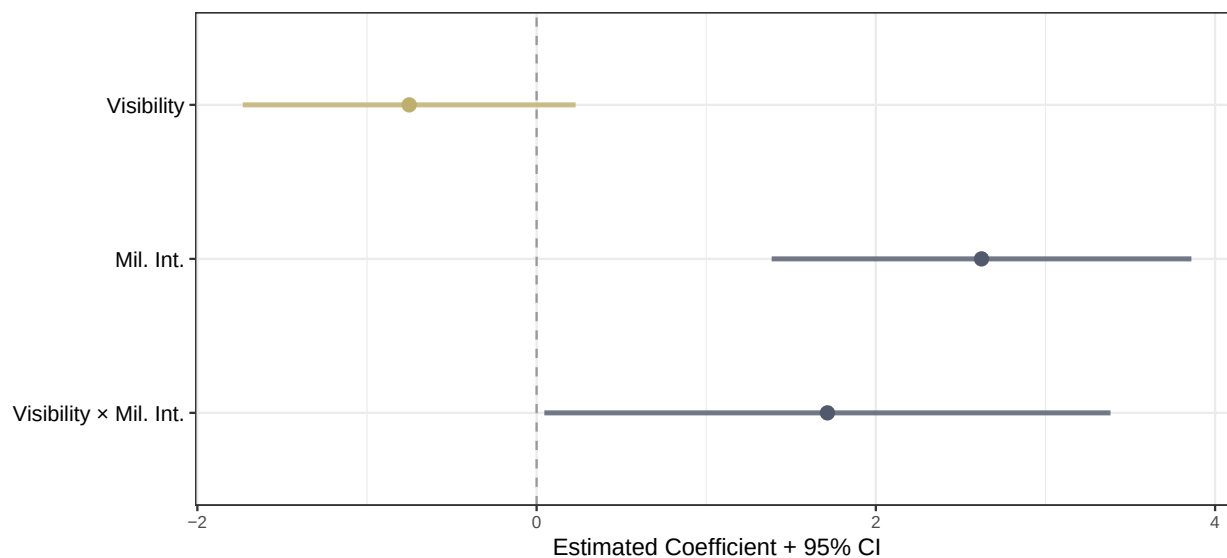


Figure 4.7: Hawkishness Moderates the Visibility Effect on General Reciprocity

DV: General Reciprocity



FPI Moderating the Costliness Effect

For the cost treatment, hawkishness (Militant Internationalism) is the most consistent moderator. For support for *private* and *general* reciprocity, hawkishness significantly exacerbates the decreasing effect of costliness on support for reciprocity, and is nearly significant for public reciprocity ($p =$

0.053). The combined coefficient plot in Figure 4.8 displays these three models together. Hawks being more sensitive to costs is somewhat surprising—one might expect hawkish respondents to be more willing to absorb costs in order to retaliate—and warrants further investigation.

Isolationism is a nearly significant moderator ($p = 0.098$) of the cost effect on support for *private* reciprocity, as shown in Figure 4.9. The pattern suggests that isolationism dampens the decreasing effect of costliness, such that the more isolationist a respondent, the more supportive they remain of using private reciprocity against Country A even as cost increases. This relationship does not hold for the other outcome variables and falls short of conventional significance, so it should be interpreted with caution.

Figure 4.8: Hawkishness Moderates the Costliness Effect on Reciprocity

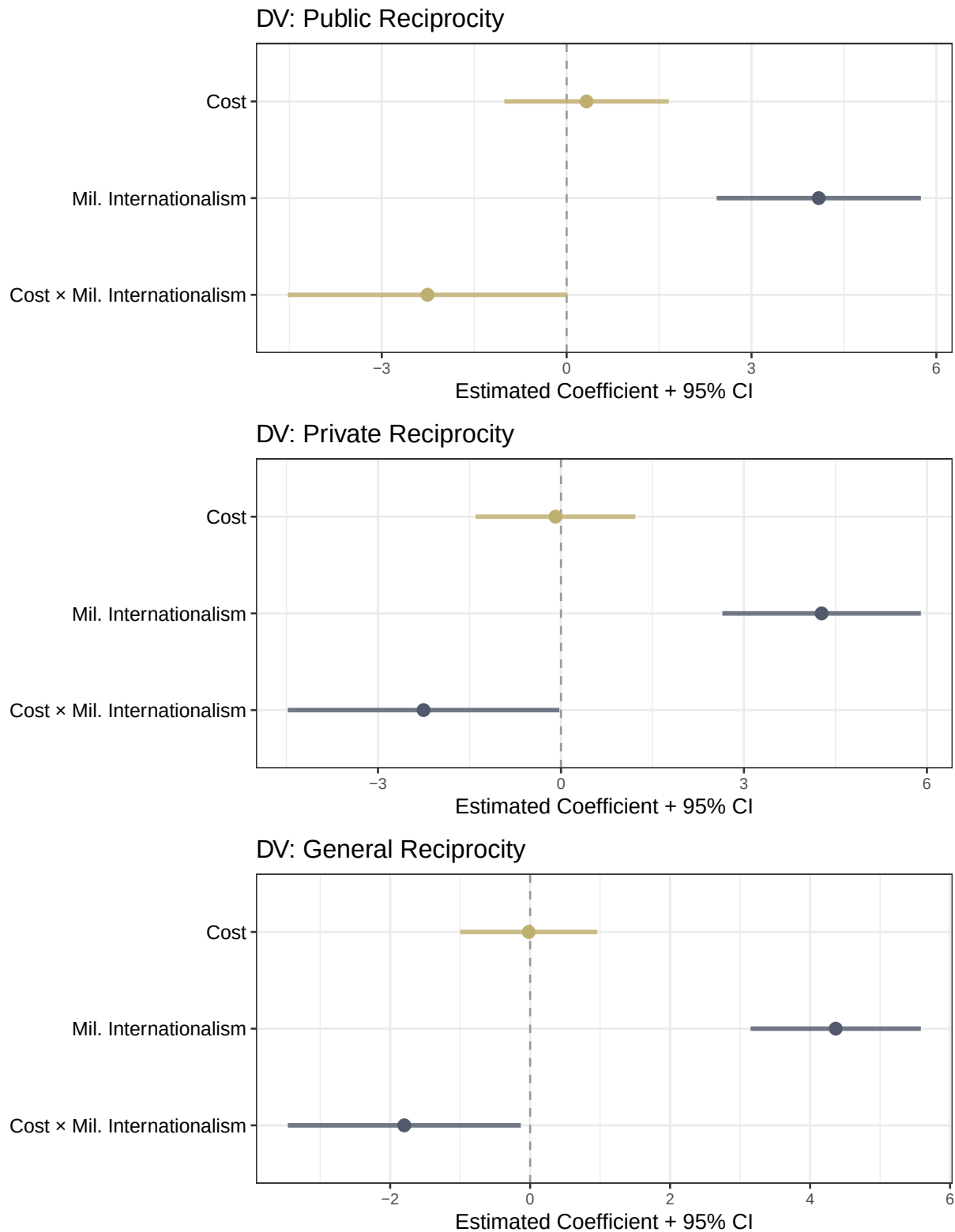
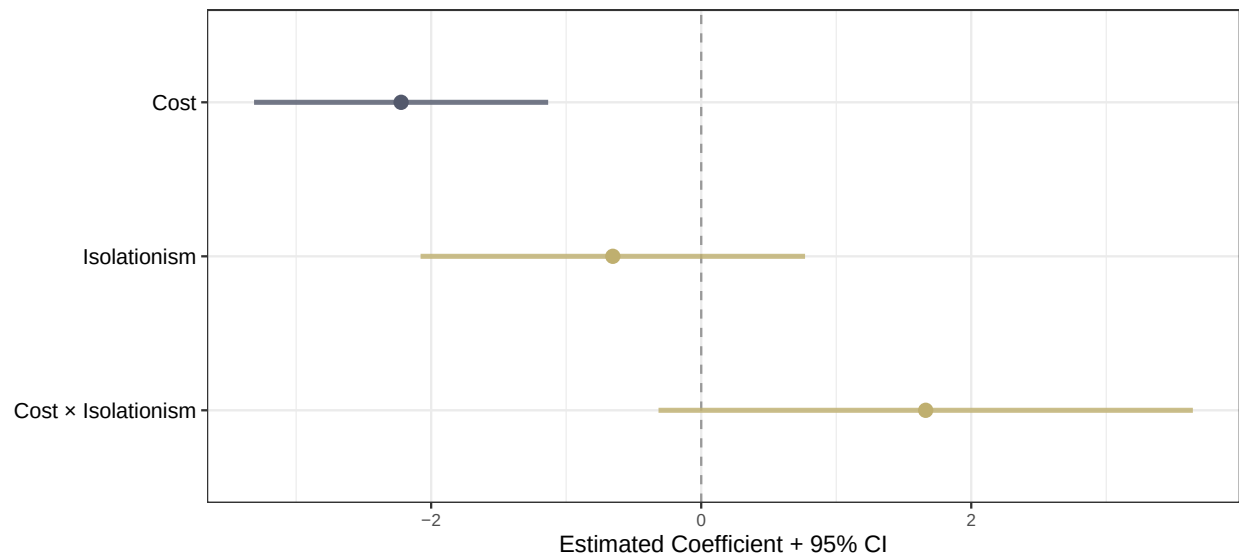


Figure 4.9: Isolationism Moderates the Costliness Effect on Private Reciprocity
DV: Private Reciprocity



Regression Tables

Table 4.1: FPI as Moderators of Visibility Effect

	<i>Dependent Variable:</i>			
	Public Reciprocity	Private Rec.	General Rec.	
	(1)	(2)	(3)	(4)
Visibility	−0.670 (0.648)	−1.342** (0.669)	−1.095** (0.555)	−0.751 (0.501)
Mil. Int.	3.083*** (0.631)	1.891** (0.844)	3.188*** (0.618)	2.623*** (0.631)
Coop. Int.	0.638 (0.795)	0.486 (0.610)	0.260 (0.610)	0.639 (0.456)
Isolat.	1.827*** (0.534)	0.170 (0.523)	−1.013 (0.739)	1.163*** (0.392)
Visibility × MI		2.541** (1.138)		1.715** (0.852)
Visibility × CI	2.066** (1.013)			
Visibility × Iso			2.297** (1.011)	
Cost	−0.962*** (0.192)	−1.365*** (0.188)	−1.373*** (0.188)	−1.037*** (0.141)
Constant	0.698 (0.769)	4.782*** (0.755)	4.852*** (0.765)	2.843*** (0.565)
Observations	1,000	1,000	1,000	1,000
R ²	0.095	0.086	0.086	0.135
Adjusted R ²	0.086	0.076	0.077	0.126
Residual Std. Error (df = 989)	3.013	2.957	2.956	2.213
F Statistic (df = 10; 989)	10.401***	9.262***	9.281***	15.425***

Note:

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 4.2: FPI as Moderators of Cost Effect

	<i>Dependent Variable:</i>			
	Public Reciprocity	Private Rec.		General Rec.
	(1)	(2)	(3)	(4)
Cost	0.305 (0.684)	-0.093 (0.668)	-2.230*** (0.555)	-0.027 (0.500)
Mil. Int.	4.077*** (0.849)	4.272*** (0.830)	3.105*** (0.620)	4.361*** (0.621)
Coop. Int.	1.775*** (0.622)	0.440 (0.608)	0.359 (0.608)	0.632 (0.455)
Isolationism	2.049*** (0.536)	0.259 (0.523)	-0.643 (0.726)	1.262*** (0.392)
Cost $\times$ MI	-2.251* (1.162)	-2.252** (1.135)		-1.796** (0.850)
Cost $\times$ Iso			1.672* (1.009)	
Constant	-0.377 (0.800)	3.484*** (0.781)	4.677*** (0.759)	1.936*** (0.585)
Observations	1,000	1,000	1,000	1,000
R ²	0.086	0.084	0.083	0.133
Adjusted R ²	0.078	0.076	0.075	0.125
Residual Std. Error (df = 990)	3.027	2.957	2.959	2.214
F Statistic (df = 9; 990)	10.360***	10.145***	10.001***	16.908***

*Note:** $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Supplementary Analysis of Mechanisms

While the results offer mixed evidence on how cost and visibility affect the individual use of reciprocity, they do not tell us specifically if these effects operate in line with a psychological approach to reciprocity. Put another way, were the participants less willing to support a reciprocal retaliation against Country A in the high cost condition because they used reciprocity as heuristic or mental shortcut to decision-making, as expected by this theoretical approach? While the main purpose of this study was to test whether cost and visibility affected individual use of reciprocity in security contexts, mediation analysis can at least offer some partial, potentially compelling insights into whether reciprocity really operates as a strategic mental shortcut in international contexts as my theoretical approach suggests.

This is an important question to explore for two reasons: first, understanding how exactly reciprocity works at the individual-level improves our understanding of the political world; and second, more specificity in the theorizing of important processes in international relations can lead to improved predictions and explanations. In this case, it is possible that the mental processes that drive reciprocity, cost and reputation considerations are more complicated than we thought, and requires more careful theorizing to adequately explain how individuals use reciprocity in IR.

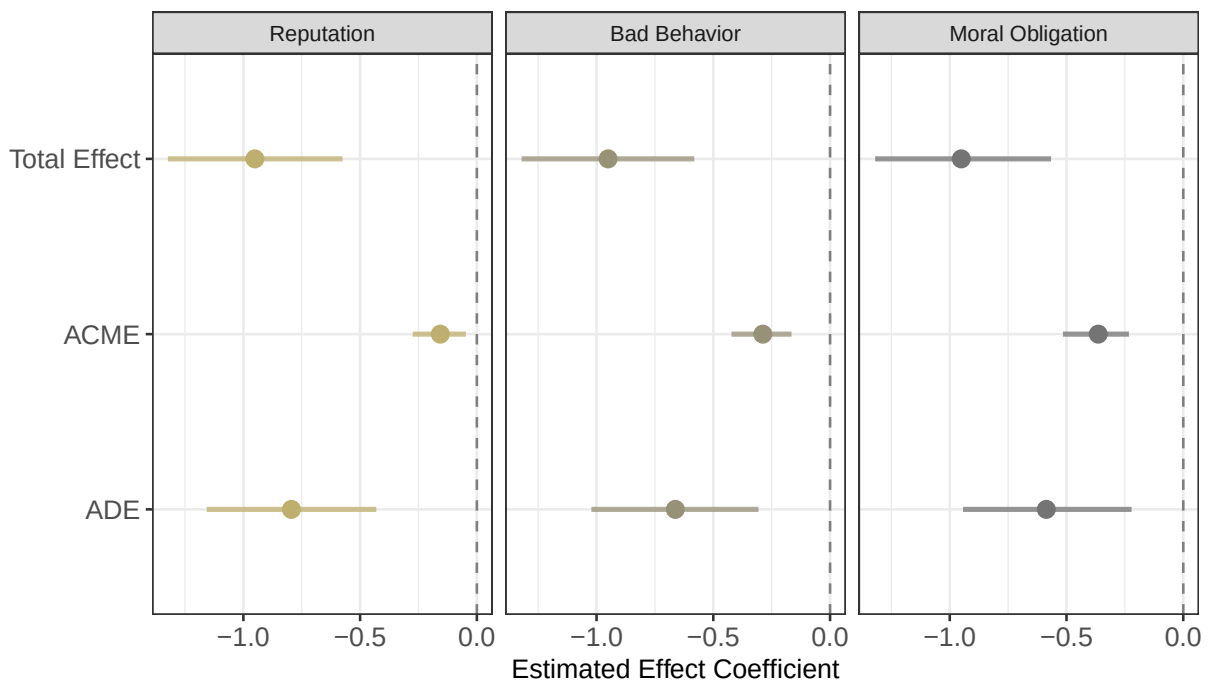
As discussed in the theory section of this paper, there are four mechanisms that could potentially drive the effect of reciprocity: reputational concerns, correcting bad behavior, fairness/morality concerns, and emotions. Leveraging the survey questions related to these mechanisms, I statistically test whether any of these mechanisms significantly mediate the observed effect of cost or visibility on support for negative reciprocity towards Country A. For purposes of clarity and brevity, I focus only on presenting findings related to the cost effect because there were no significant mediators for the visibility effect, and I only present findings on the support for the publicly claimed reciprocity dependent variable.¹² The results of the mediation analysis are presented below in figure 4.10, with significant mediators shown in each panel.¹³ For each mechanism's corresponding panel in

¹² Mediation results were consistent across all three dependent variables.

¹³ Main models can be found in the Appendix, and all results were robust to specifications that included different covariates.

the figure, the Total Effect of costly reciprocity on the main outcome (support for publicly claimed reciprocity towards Country A) is shown. It also shows the Averaged Causal Mediated Effect (ACME), or how much of this Total Effect was mediated by that specific mechanism. Finally, the Average Direct Effect (ADE) shows how much of the remaining Total Effect was not mediated by that mechanism. In terms of interpretation, the ACME and ADE coefficients are proportions of the Total Effect. In other words, as the ACME coefficient for a mediator gets closer to the Total Effect, the more that specific mediator is estimated to be responsible for that total effect. The farther away the ACME coefficient is from the Total Effect, the less that mechanism is responsible for the Total Effect of indirect reciprocity on support for cooperation with Vexa. The inverse is true of ADE.

Figure 4.10: Costliness Effect on Public Reciprocity Towards Country A



Two points are worth noting. First, three of the four mechanisms suggested by the psychological approach to reciprocity partially and significantly mediate the indirect reciprocity effect: reputation, correcting bad behavior, and moral obligation/fairness. How much participants show concerned for the US's reputation with others, agreed that US should use its cooperation to discourage other countries from challenging them, and thought that they had a moral obligation to retaliate against Country A all appear to partially drive the effect of cost on support for reciprocity. Emotions, mea-

sured through a feeling thermometer, did not mediate the costliness effect. This is an intriguing finding; the negative ACME for each mediator indicates that higher cost reduced concerns for reputation, bad behavior, and moral obligation, key theoretical heuristic mechanisms, and ultimately lowered support for reciprocity. An interpretation of this cost effect could be that consideration of costs play a more important role in the mental process of reciprocity than in other fields, and that our understanding of how individuals view the value of reputation and costly signals is more complicated than generally expected. Second, morality and fairness concerns seem to be particularly important drivers for the use of reciprocity, and are particularly affected by exposure to costs: moral obligation was estimated to mediate a large portion of the cost effect (38% of the Total Effect) on reciprocity. I interpret this as evidence that reciprocity, even in international security settings, can operate as a heuristic that goes beyond purely reputational concerns. Moral concerns, like feeling that retaliating against Country A is the right thing to do, has a positive effect on reciprocity in these models when cost is held constant.

Conclusions based on these results should be taken with some caution. There are obvious issues in trying to measure a strategic heuristic with survey questions, particularly in separating out how reputational concerns are being used heuristically rather than from a rational calculus. Both perspectives would expect that participants express concern about the US's reputation, for instance.

Nonetheless, these are fascinating findings, and lend support to the idea that individuals do likely use reciprocity through these mechanisms as theorized, even in the realm of international affairs. Perhaps more interestingly, the results indicate that the introduction of cost information in particular deeply affects this heuristic process. It also likely means that the mental processes that facilitate the use of reciprocity are very complicated, with multiple factors influencing this behavior simultaneously, and it is unclear to what extent these mechanisms interact with one another or complement/combat the others. In every day life, it is easy to imagine how these mechanisms can operate at the same time when exposed to reciprocity; we might be thinking about our reputation with others when we retaliate against someone who struck us first, but we may also simultaneously feel moral outrage to "get even" or make things "right". In particular, the mixed results of the visibility treatment and subsequent mediation analysis indicate that reputation as a motivation

for using reciprocity is not as straightforward at the individual level. People seem to conditionally care about who is watching their country's reciprocity acts, but not through the mechanisms that psychological approaches to reciprocity would predict. For example, if people indeed use reciprocity in a heuristic fashion that shapes how they interpret the costs of foreign policy or security decisions, this would represent a departure from both traditional IR approaches to reciprocity and the psychological literature on reciprocity. The use of cost information may not be as calculative as rational IR approaches to reciprocity and reputation have assumed on the one hand, but this information is much more salient than work on reciprocity in other fields where individuals are often insensitive to cost on the other.

Bibliography

- Aldrich, John H, Christopher Gelpi, Peter Feaver, Jason Reifler and Kristin Thompson Sharp. 2006. "Foreign policy and the electoral connection." *Annu. Rev. Polit. Sci.* 9:477–502.
- Alexander, Richard. 2017. *The biology of moral systems*. Routledge.
- Aron, Arthur, Elaine N Aron and Danny Smollan. 1992. "Inclusion of other in the self scale and the structure of interpersonal closeness." *Journal of personality and social psychology* 63(4):596.
- Axelrod, Robert. 1980. "Effective choice in the prisoner's dilemma." *Journal of conflict resolution* 24(1):3–25.
- Axelrod, Robert. 2006. *The evolution of cooperation*. Revised ed. New York: Basic Books.
- Balliet, Daniel, Junhui Wu and Carsten KW De Dreu. 2014. "Ingroup favoritism in cooperation: a meta-analysis." *Psychological bulletin* 140(6):1556.
- Baum, Matthew A and Philip BK Potter. 2008. "The relationships between mass media, public opinion, and foreign policy: Toward a theoretical synthesis." *Annu. Rev. Polit. Sci.* 11:39–65.
- Berg, Joyce, John Dickhaut and Kevin McCabe. 1995. "Trust, reciprocity, and social history." *Games and economic behavior* 10(1):122–142.
- Bergé, Laurent R., Kyle Butts and Grant McDermott. 2026. "Fast and user-friendly econometrics estimations: The R package fixest." *arXiv preprint arXiv:2601.21749*.
- Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and

- Michael Ward. 2015. "ICEWS Coded Event Data." .
URL: <https://doi.org/10.7910/DVN/28075>
- Boyd, Robert and Peter J Richerson. 1988. "The evolution of reciprocity in sizable groups." *Journal of theoretical Biology* 132(3):337–356.
- Boyd, Robert and Peter J Richerson. 1989. "The evolution of indirect reciprocity." *Social networks* 11(3):213–236.
- Brandt, Hannelore, Hisashi Ohtsuki, Yoh Iwasa and Karl Sigmund. 2007. A survey of indirect reciprocity. In *Mathematics for ecology and environmental sciences*. Springer pp. 21–49.
- Brewer, Marilynn B and Linnda R Caporael. 2013. An evolutionary perspective on social identity: Revisiting groups. In *Evolution and social psychology*. Psychology Press pp. 143–161.
- Cameron, James E. 2004. "A three-factor model of social identity." *Self and identity* 3(3):239–262.
- Caporael, Linnda R. 1997. "The evolution of truly social cognition: The core configurations model." *Personality and Social Psychology Review* 1(4):276–298.
- Carlin, Ryan E and Gregory J Love. 2013. "The politics of interpersonal trust and reciprocity: An experimental approach." *Political behavior* 35(1):43–63.
- Carson, Austin. 2016. "Facing off and saving face: covert intervention and escalation management in the Korean War." *International Organization* 70(1):103–131.
- Carson, Austin and Keren Yarhi-Milo. 2017. "Covert communication: The intelligibility and credibility of signaling in secret." *Security Studies* 26(1):124–156.
- Chiba, Daina, Jesse C Johnson and Brett Ashley Leeds. 2015. "Careful commitments: Democratic states and alliance design." *The Journal of Politics* 77(4):968–982.
- Crescenzi, Mark JC, Rebecca H Best and Bo Ram Kwon. 2010. Reciprocity in international studies. In *Oxford Research Encyclopedia of International Studies*.
- Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes." *American political science review* 88(3):577–592.

- Fearon, James D. 1997a. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1):68–90.
- Fearon, James D. 1997b. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1):68–90.
- Fearon, James D. 1998. "Bargaining, enforcement, and international cooperation." *International organization* 52(2):269–305.
- Fehr, Ernst and Simon Gächter. 2002. "Altruistic punishment in humans." *Nature* 415(6868):137–140.
- Fehr, Ernst and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and human behavior* 25(2):63–87.
- Fehr, Ernst, Urs Fischbacher and Simon Gächter. 2002. "Strong reciprocity, human cooperation, and the enforcement of social norms." *Human nature* 13:1–25.
- Frank, Robert H. 2011. "The strategic role of the emotions." *Emotion Review* 3(3):252–254.
- Gartzke, Erik. 1998. "Kant we all just get along? Opportunity, willingness, and the origins of the democratic peace." *American Journal of Political Science* pp. 1–27.
- Gartzke, Erik and Jon R Lindsay. 2015. "Weaving tangled webs: offense, defense, and deception in cyberspace." *Security Studies* 24(2):316–348.
- Gartzke, Erik and Kristian Skrede Gleditsch. 2004. "Why democracies may actually be less reliable allies." *American Journal of Political Science* 48(4):775–795.
- Gartzke, Erik and Kristian Skrede Gleditsch. 2006. "Identity and conflict: Ties that bind and differences that divide." *European Journal of International Relations* 12(1):53–87.
- Gibbs, David N. 1995. "Secrecy and international relations." *Journal of Peace Research* 32(2):213–228.
- Goldstein, Joshua S. 1991. "Reciprocity in superpower relations: An empirical analysis." *International Studies Quarterly* 35(2):195–209.

- Goldstein, Joshua S and John R Freeman. 1990. *Three-way street: Strategic reciprocity in world politics*. Chicago: University of Chicago Press.
- Goldstein, Joshua S and John R Freeman. 1991. "US-Soviet-Chinese relations: Routine, reciprocity, or rational expectations?" *American Political Science Review* 85(1):17–35.
- Goldstein, Joshua S and Jon C Pevehouse. 1997. "Reciprocity, bullying, and international cooperation: Time-series analysis of the Bosnia conflict." *American Political Science Review* 91(3):515–529.
- Gray, Kurt, Adrian F Ward and Michael I Norton. 2014. "Paying it forward: generalized reciprocity and the limits of generosity." *Journal of experimental psychology: General* 143(1):247.
- Grieco, Joseph M. 1988. "Anarchy and the limits of cooperation: a realist critique of the newest liberal institutionalism." *International organization* 42(3):485–507.
- Gummerum, Michaela, Masanori Takezawa and Monika Keller. 2009. "The influence of social category and reciprocity on adults' and children's altruistic behavior." *Evolutionary Psychology* 7(2):147470490900700212.
- Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath et al. 2001. "Cooperation, reciprocity and punishment in fifteen small-scale societies." *American Economic Review* 91(2):73–78.
- Holsti, Ole R. 2004. *Public opinion and American foreign policy*. University of Michigan Press.
- Huth, Paul K. 1988. "Extended deterrence and the outbreak of war." *American Political Science Review* 82(2):423–443.
- Huth, Paul K. 1999. "Deterrence and international conflict: Empirical findings and theoretical debates." *Annual Review of Political Science* 2(1):25–48.
- Hyde, Susan D. 2015. "Experiments in international relations: Lab, survey, and field." *Annual Review of Political Science* 18:403–424.
- Imai, Kosuke and In Song Kim. 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis* 29(3):405–415.

- Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1):51 – 71.
URL: <https://doi.org/10.1214/10-STS321>
- Jervis, Robert. 1976. *Perception and misperception in international politics: New edition*. Princeton University Press.
- Jervis, Robert, Keren Yarhi-Milo and Don Casler. 2021. "Redefining the debate over reputation and credibility in international security: Promises and Limits of New Scholarship." *World Politics* 73(1):167–203.
- Keohane, Robert O. 1986. "Reciprocity in international relations." *International organization* 40(1):1–27.
- Kertzer, Joshua D. 2022. "Re-assessing elite-public gaps in political behavior." *American Journal of Political Science* 66(3):539–553.
- Keysar, Boaz, Benjamin A Converse, Jiunwen Wang and Nicholas Epley. 2008. "Reciprocity is not give and take: Asymmetric reciprocity to positive and negative acts." *Psychological Science* 19(12):1280–1286.
- Kydd, Andrew. 1997. "Game theory and the spiral model." *World Politics* 49(3):371–400.
- Kydd, Andrew. 2000. "Trust, reassurance, and cooperation." *International Organization* 54(2):325–357.
- Kydd, Andrew H. 2007. *Trust and mistrust in international relations*. Princeton, NJ: Princeton University Press.
- Larson, Deborah Welch. 1988. "The psychology of reciprocity in international relations." *Negot. J.* 4:281.
- Leach, Colin Wayne, Martijn Van Zomeren, Sven Zebel, Michael LW Vliek, Sjoerd F Pennekamp, Bertjan Doosje, Jaap W Ouwerkerk and Russell Spears. 2008. "Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification." *Journal of personality and social psychology* 95(1):144.

- Leng, Russell J. 1993a. *Interstate Crisis Behavior, 1816-1980*. Cambridge University Press.
- Leng, Russell J. 1993b. "Reciprocating influence strategies in interstate crisis bargaining." *Journal of Conflict Resolution* 37(1):3-41.
- Lindskold, Svenn. 1978. "Trust development, the GRIT proposal, and the effects of conciliatory acts on conflict and cooperation." *Psychological Bulletin* 85(4):772.
- Marcus, George E, Michael MacKuen, Jennifer Wolak and Luke Keele. 2006. The measure and mismeasure of emotion. In *Feeling politics: Emotion in political information processing*. Springer pp. 31-45.
- McDermott, Rose, Anthony C Lopez and Peter K Hatemi. 2017. "'Blunt Not the Heart, Enrage It': The Psychology of Revenge and Deterrence (November 2017)." *Texas National Security Review*.
- Mellers, Barbara, Alan Schwartz and Ilana Ritov. 1999. "Emotion-based choice." *Journal of experimental psychology: General* 128(3):332.
- Miller, Steven V. 2022. "peacesciencer: An R Package for Quantitative Peace Science Research." *Conflict Management and Peace Science* 39(6):755-779.
- Nowak, Martin A and Karl Sigmund. 2005. "Evolution of indirect reciprocity." *Nature* 437(7063):1291-1298.
- Nowak, Martin A and Sébastien Roch. 2007. "Upstream reciprocity and the evolution of gratitude." *Proceedings of the royal society B: Biological Sciences* 274(1610):605-610.
- Oneal, John R. and Bruce Russett. 2001. "Clear and Clean: The Fixed Effects of the Liberal Peace." *International Organization* 55(2):469-485.
- Osgood, Charles E. 1962. "An alternative to war or surrender."
- Oskamp, Stuart. 1971. "Effects of programmed strategies on cooperation in the Prisoner's Dilemma and other mixed-motive games." *Journal of Conflict Resolution* 15(2):225-259.
- Perugini, Marco, Marcello Gallucci, Fabio Presaghi and Anna Paola Ercolani. 2003. "The personal norm of reciprocity." *European Journal of Personality* 17(4):251-283.

- Poznansky, Michael and Evan Perkoski. 2018. "Rethinking secrecy in cyberspace: the politics of voluntary attribution." *Journal of Global Security Studies* 3(4):402–416.
- Rabin, Matthew et al. 1992. *Incorporating fairness into game theory and economics*. University of California at Berkeley, Department of Economics.
- Rajmaira, Sheen and Michael D Ward. 1990. "Evolving foreign policy norms: Reciprocity in the superpower triad." *International Studies Quarterly* 34(4):457–475.
- Rathbun, Brian C. 2011. *Trust in international cooperation: International security institutions, domestic politics and American multilateralism*. Vol. 121 Cambridge University Press.
- Rathbun, Brian C, Christopher Sebastian Parker and Caleb Pomeroy. 2025. "Separate but unequal: Ethnocentrism and racialization explain the "democratic" peace in public opinion." *American Political Science Review* 119(2):621–636.
- Rathbun, Brian C, Joshua D Kertzer, Jason Reifler, Paul Goren and Thomas J Scotto. 2016. "Taking foreign policy personally: Personal values and foreign policy attitudes." *International Studies Quarterly* 60(1):124–137.
- Reiter, Dan and Allan C Stam. 2002. *Democracies at war*. Princeton University Press.
- Renshon, Jonathan and Ryan Powers. 2024. "Identity and the Social Construction of Reputation in World Politics."
- Rhodes, Carolyn. 1989. "Reciprocity in trade: The utility of a bargaining strategy." *International Organization* 43(2):273–299.
- Romano, Angelo, Daniel Balliet and Junhui Wu. 2017. "Unbounded indirect reciprocity: Is reputation-based cooperation bounded by group membership?" *Journal of Experimental Social Psychology* 71:59–67.
- Roskos-Ewoldsen, Andrew, Morgan Ellithorpe and Brandon J Kinne. 2026. "The Effect of Reciprocity on Trust: International Cooperation and COVID Aid." *Journal of Conflict Resolution* 70(1):115–139.

- Russett, Bruce M and John R Oneal. 2001. "Triangulating peace: Democracy, interdependence, and international organizations." (*No Title*) .
- Sanger, David E. 2012. *Confront and conceal: Obama's secret wars and surprising use of American power*. Crown.
- Schelling, Thomas C. 2008. Arms and influence. In *Strategic Studies*. Routledge pp. 96–114.
- Schroeder, David A, Julie E Steel, Andria J Woodell and Alicia F Bembenek. 2003. "Justice within social dilemmas." *Personality and Social Psychology Review* 7(4):374–387.
- Shinada, Mizuho, Toshio Yamagishi and Yu Ohmura. 2004. "False friends are worse than bitter enemies: "Altruistic" punishment of in-group members." *Evolution and Human Behavior* 25(6):379–393.
- Spence, Michael. 1978. Job market signaling. In *Uncertainty in economics*. Elsevier pp. 281–306.
- Tajfel, Henri. 1970. "Experiments in intergroup discrimination." *Scientific american* 223(5):96–103.
- Tajfel, Henri. 1974. "Social identity and intergroup behaviour." *Social science information* 13(2):65–93.
- Tomz, Michael and Jessica LP Weeks. 2021. "Military alliances and public support for war." *International Studies Quarterly* 65(3):811–824.
- Tomz, Michael, Jessica LP Weeks and Keren Yarhi-Milo. 2020. "Public opinion and decisions about military force in democracies." *International Organization* 74(1):119–143.
- Tomz, Michael R and Jessica LP Weeks. 2013. "Public opinion and the democratic peace." *American political science review* 107(4):849–865.
- Uslaner, Eric M. 2002. *The Moral Foundations of Trust*. Cambridge University Press.
- Ward, Michael D and Sheen Rajmaira. 1992. "Reciprocity and norms in US-Soviet foreign policy." *Journal of Conflict Resolution* 36(2):342–368.
- Ward, Michael Don. 1982. "Cooperation and conflict in foreign policy behavior: Reaction and memory." *International Studies Quarterly* 26(1):87–126.

- Weisiger, Alex and Keren Yarhi-Milo. 2015. "Revisiting reputation: How past actions matter in international politics." *International Organization* 69(2):473–495.
- Wilson, Warner. 1971. "Reciprocation and other techniques for inducing cooperation in the Prisoner's Dilemma game." *Journal of Conflict Resolution* 15(2):167–195.
- Wu, Junhui, Daniel Balliet and Paul AM Van Lange. 2016. "Reputation, gossip, and human cooperation." *Social and Personality Psychology Compass* 10(6):350–364.
- Yamagishi, Toshio, Nobuhito Jin and Toko Kiyonari. 1999. "Bounded generalized reciprocity: ingroup boasting and ingroup favoritism. Adv." *Group Process* 16:161–197.
- Yamagishi, Toshio and Toko Kiyonari. 2000. "The group as the container of generalized reciprocity." *Social psychology quarterly* pp. 116–132.
- Yarhi-Milo, Keren, Joshua D Kertzer and Jonathan Renshon. 2018. "Tying hands, sinking costs, and leader attributes." *Journal of Conflict Resolution* 62(10):2150–2179.
- Zheng, Wenqian, Shengli Cao, Yangyang Wang, Ke Yang, Yunyun Chen and Guoping Song. 2021. "The impact of social value orientation, game context and trust on cooperative behavior after cooperative video game play." *Psychological Reports* 124(3):1353–1369.

Chapter 5

Reciprocity in International Relations: A Synthesis

5.1 Overview

This dissertation set out to answer a deceptively simple question: why and how do individuals use reciprocity when making decisions in international relations contexts? The answer, it turns out, is both more nuanced and more interesting than existing accounts of reciprocity in either international relations or psychology would suggest.

Across three empirical studies (indirect reciprocity, group-based reciprocity, and costly reciprocity) I find consistent evidence that reciprocity is a powerful force in individual decision-making about international affairs. People penalize countries that fail to return favors, even when they were not involved in the original exchange. They respond differently to ingroups and outgroups, though not always in the ways theory predicts. And they use reciprocity even when it is costly, though cost sensitivity introduces important limits on the heuristic logic of the BGR framework. Taken together, these studies paint a picture of reciprocity as a fundamental but psychologically complex feature of how individuals understand and respond to international relations.

5.2 Summary of Findings

Chapter 2: Indirect Reciprocity

The first study produced the clearest and most theoretically consistent findings. When individuals learned that a potential partner country had failed to return a favor to a third, uninvolved country, their support for cooperating with that country dropped significantly, and this effect held across both security and trade contexts. Mediation analysis confirmed that the effect was driven by a combination of reputational, moral, and emotional mechanisms consistent with the BGR framework: people thought it was the right thing to avoid cooperation with the violating country, felt negatively toward it, and believed that how they handled the other country would affect their reputation. Observational evidence from ICEWS data provided partial corroboration at the state level, with countries that reciprocated material acts receiving more cooperative behavior from uninvolved third parties.

This study makes a significant contribution to our understanding of reciprocity in international relations by demonstrating that the logic of reciprocity extends beyond direct dyadic interactions. Countries are not evaluated solely on the basis of how they treat their direct partners; their broader reputation for following the rules of reciprocity with others also shapes whether third parties will cooperate with them. This finding has important practical implications for foreign policy: states that fail to return favors, even when they feel they should be forgiven for doing so, may face downstream cooperative costs with uninvolved parties.

Chapter 3: Group Reciprocity

The second study produced more mixed results, and in doing so revealed important limits and complications of the BGR framework in international contexts. Contrary to the prediction that ingroup violations of reciprocity would be penalized most harshly, group identity did not moderate the effect of reciprocity on cooperation support or trust. The penalty for failing to return a favor was substantial regardless of whether the offending country was an ingroup member, an outgroup

member, or an unknown country.

A striking finding was an amplification effect where BGR would not expect one to exist: out-group countries that used negative reciprocity were penalized most severely across all conditions. This suggests that the perceived gap between when an outgroup country behaves positively or aggressively in a reciprocal context is larger than it is for other countries.

These results suggest two important refinements to the BGR framework as applied to international contexts. First, reciprocity may actually be more “unbounded” than “bounded” by groupness, where reciprocity works the same for any country regardless of group because people expect their own country’s outcomes to be tied to more than the ingroup. The gossip, monitoring and broad social networks associated with international affairs may make the conditions for ingroup bias more difficult to achieve than theories derived from domestic social psychology would predict. Second, the interaction between group identity and negative reciprocity may be more important than the interaction between group identity and positive reciprocity or its violation, an asymmetry that warrants further exploration.

Chapter 4: Costly Reciprocity

The third study revealed perhaps the most theoretically challenging finding: contrary to the BGR prediction that individuals would use reciprocity as an automatic heuristic that downweights cost information, the public was sensitive to costs. People were more opposed to negative reciprocity, or retaliating against a country that had committed a cyberattack, when the costs of doing so were higher. This is not what a straightforward heuristic account would predict. Yet, because the effect of costs operated through moral and reputational mediators, and the effect of cost is moderated by foreign policy disposition, there is evidence to suggest that reciprocity in this context is not entirely an exhaustive, deliberative calculation.

The visibility of the original offense also produced conditional effects: when a counterpart country’s act was publicly visible, individuals expressed stronger support for a publicly visible retaliation, consistent with a reputational logic. But this effect did not generalize to other forms of reciprocity,

suggesting that the role of reputation building in reciprocity at the individual level is more complex than BGR anticipates.

5.3 Theoretical Implications

Taken together, the three studies suggest a more complex model of individual-level reciprocity in international relations than the standard BGR framework implies.

Reciprocity is pervasive but not automatic. Across all three studies, reciprocity was a powerful predictor of cooperative support, trust, or policy preferences. People clearly use reciprocity as a guide to their evaluations of foreign countries. But the effects are not uniform or unconditional: they are shaped by costs, by the visibility of reciprocal exchanges, and by the type of reciprocity involved (positive, negative, or violation of positive). The heuristic model of reciprocity captures the general pattern but is under-specified to explain important nuances in how contextual information modulates the reciprocity response.

The mechanisms of reciprocity are multiple and interact. BGR and related works on reciprocity identify four mechanisms (reputation, correction of bad behavior, moral obligation, and emotion), and all four appear to be empirically relevant across the three studies. But these mechanisms do not operate independently or uniformly. Moral obligation and emotion were particularly consistent and large mediators, suggesting that people's affect and sense of what is right or fair are central drivers of reciprocity behavior in international contexts. Reputational concerns were also consistently important, particularly when audience visibility was high. Exploring how these mechanisms combine, cancel each other out, or add to one another depending on the context is an important next step in developing a general theory of reciprocity in international relations.

Reciprocity may be more universal than group-based. The group reciprocity study found that reciprocity effects were robust regardless of group membership, while also finding that the interaction between outgroup identity and negative reciprocity produced the strongest punitive responses. This suggests that reciprocity may operate the same for most countries, but that there are important exceptions whose conditions need to be better understood.

Reciprocity reactions may depend on personal characteristics. The costly reciprocity study found in its supplementary analysis that foreign policy orientations significantly moderate the effect of cost and visibility on preferences for reciprocity behavior. Dovishness, hawkishness, and isolationist views of the world bring about different responses to negative reciprocity. Since FPO's are linked to moral foundations, reciprocity may interact with a person's value system in ways we do not yet understand.

5.4 Implications for International Relations Theory and Practice

These findings have several implications for international relations theory and practice.

For **international cooperation**, the results suggest that cooperation can be delicate. Countries should be mindful of how they treat their direct partners, not only because of how their partner feels but also because of how third parties may feel; their behaviors shape their reputation and downstream cooperative prospects. Countries that consistently return favors and abide by reciprocity norms are likely to receive more cooperative treatment from uninvolved parties. Conversely, violations of reciprocity are significant cooperative setbacks and possibly difficult to overcome given the emotional and moral mechanisms that consistently drive the effect of reciprocity.

For **group dynamics**, the findings suggest that ingroup status does not provide the cooperative insurance that political leaders might assume. Friends or partners are not forgiven more readily than strangers when they fail to reciprocate. Managing expectations around reciprocity within groups and beyond groups, like in military alliances or between historical enemies, may thus be more important than commonly recognized.

For **crisis bargaining and conflict**, the costly reciprocity findings suggest that public support for retaliation is sensitive to cost. This cost sensitivity operating through moral and reputational channels means that political leaders can potentially influence attitudes towards retaliation through specific framing. If leaders can credibly frame retaliatory acts as morally necessary or reputation-preserving, they may be able to sustain public support for costly reciprocal acts that would otherwise face opposition. The outgroup being penalized most harshly for negative reciprocity also

suggests that adversarial relationships may be more susceptible to reciprocal conflict escalation, even when you are playing by the rules of reciprocity or feel the retaliation is justified.

For **public opinion research**, the three studies collectively demonstrate that ordinary citizens hold complex and psychologically grounded views about reciprocity in international affairs. They distinguish between direct and indirect violations, between friendly and adversarial counterparts, and between costly and cheap forms of reciprocation. Understanding these distinctions may help scholars and policy-makers better anticipate public reactions to foreign policy decisions.

5.5 Limitations and Future Directions

Several limitations of this research merit acknowledgment. All three studies rely on US samples, which raises questions about generalizability to other national publics with different cultural relationships to reciprocity norms. The survey experiments are conducted in hypothetical settings, which may not fully capture the emotional and social dynamics that accompany real-world reciprocity in high-stakes international contexts. The observational study in Chapter 2 is limited in its ability to establish causation at the state level, and future work using more refined causal identification strategies would strengthen the connection between individual-level and state-level reciprocity. The lack of comparison to elite use of reciprocity also limits these studies' ability to speak about how individuals generally use reciprocity.

Looking forward, several directions seem particularly promising. First, the mechanisms through which individual-level reciprocity aggregates to state-level behavior remain underspecified. Future research could also examine the pathways that transfer reciprocity-based preferences reach elite or state-level decision-making. Second, the group reciprocity findings suggest that understanding what makes a foreign country an “ingroup” member in international contexts may be important for discovering when and why groupness moderates reciprocity; future work could investigate what specific attributes or shared histories create ingroup status in international settings, and thus when individuals are likely to view that group's system of social exchange as closed or open. Third, the costly reciprocity findings open important questions about how reciprocity norms interact with

deterrence and coercive bargaining; cross-national experimental designs that vary cost salience, visibility, and the nature of the reciprocal act could further clarify the boundary conditions of the reciprocity heuristic.

5.6 Conclusion

Reciprocity is one of the oldest and most frequently invoked concepts in the study of international relations, and yet its individual-level foundations have remained surprisingly underexplored. This dissertation has taken a step toward filling that gap by testing a psychologically grounded account of reciprocity across three distinct empirical contexts. The results reveal a behavioral tendency that is powerful, widespread, and psychologically rich.

The fundamental insight is this: individuals are highly sensitive to acts of returning favors or slights, and are committed to enforcing the rules of reciprocity through cooperation or conflict. They also see reciprocity as emotional, a matter of fairness, and how they think their country should be seen. That insight, scaled up across publics and political leaders, has consequences for how states cooperate, how they manage their reputations, and how they navigate the ever-present tension between strategic interest and the human instinct to return what has been given.

Bibliography

- Aldrich, John H, Christopher Gelpi, Peter Feaver, Jason Reifler and Kristin Thompson Sharp. 2006. "Foreign policy and the electoral connection." *Annu. Rev. Polit. Sci.* 9:477–502.
- Alexander, Richard. 2017. *The biology of moral systems*. Routledge.
- Aron, Arthur, Elaine N Aron and Danny Smollan. 1992. "Inclusion of other in the self scale and the structure of interpersonal closeness." *Journal of personality and social psychology* 63(4):596.
- Axelrod, Robert. 1980. "Effective choice in the prisoner's dilemma." *Journal of conflict resolution* 24(1):3–25.
- Axelrod, Robert. 2006. *The evolution of cooperation*. Revised ed. New York: Basic Books.
- Balliet, Daniel, Junhui Wu and Carsten KW De Dreu. 2014. "Ingroup favoritism in cooperation: a meta-analysis." *Psychological bulletin* 140(6):1556.
- Baum, Matthew A and Philip BK Potter. 2008. "The relationships between mass media, public opinion, and foreign policy: Toward a theoretical synthesis." *Annu. Rev. Polit. Sci.* 11:39–65.
- Berg, Joyce, John Dickhaut and Kevin McCabe. 1995. "Trust, reciprocity, and social history." *Games and economic behavior* 10(1):122–142.
- Bergé, Laurent R., Kyle Butts and Grant McDermott. 2026. "Fast and user-friendly econometrics estimations: The R package fixest." *arXiv preprint arXiv:2601.21749*.
- Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and

- Michael Ward. 2015. "ICEWS Coded Event Data." .
URL: <https://doi.org/10.7910/DVN/28075>
- Boyd, Robert and Peter J Richerson. 1988. "The evolution of reciprocity in sizable groups." *Journal of theoretical Biology* 132(3):337–356.
- Boyd, Robert and Peter J Richerson. 1989. "The evolution of indirect reciprocity." *Social networks* 11(3):213–236.
- Brandt, Hannelore, Hisashi Ohtsuki, Yoh Iwasa and Karl Sigmund. 2007. A survey of indirect reciprocity. In *Mathematics for ecology and environmental sciences*. Springer pp. 21–49.
- Brewer, Marilynn B and Linnda R Caporael. 2013. An evolutionary perspective on social identity: Revisiting groups. In *Evolution and social psychology*. Psychology Press pp. 143–161.
- Cameron, James E. 2004. "A three-factor model of social identity." *Self and identity* 3(3):239–262.
- Caporael, Linnda R. 1997. "The evolution of truly social cognition: The core configurations model." *Personality and Social Psychology Review* 1(4):276–298.
- Carlin, Ryan E and Gregory J Love. 2013. "The politics of interpersonal trust and reciprocity: An experimental approach." *Political behavior* 35(1):43–63.
- Carson, Austin. 2016. "Facing off and saving face: covert intervention and escalation management in the Korean War." *International Organization* 70(1):103–131.
- Carson, Austin and Keren Yarhi-Milo. 2017. "Covert communication: The intelligibility and credibility of signaling in secret." *Security Studies* 26(1):124–156.
- Chiba, Daina, Jesse C Johnson and Brett Ashley Leeds. 2015. "Careful commitments: Democratic states and alliance design." *The Journal of Politics* 77(4):968–982.
- Crescenzi, Mark JC, Rebecca H Best and Bo Ram Kwon. 2010. Reciprocity in international studies. In *Oxford Research Encyclopedia of International Studies*.
- Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes." *American political science review* 88(3):577–592.

- Fearon, James D. 1997a. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1):68–90.
- Fearon, James D. 1997b. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1):68–90.
- Fearon, James D. 1998. "Bargaining, enforcement, and international cooperation." *International organization* 52(2):269–305.
- Fehr, Ernst and Simon Gächter. 2002. "Altruistic punishment in humans." *Nature* 415(6868):137–140.
- Fehr, Ernst and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and human behavior* 25(2):63–87.
- Fehr, Ernst, Urs Fischbacher and Simon Gächter. 2002. "Strong reciprocity, human cooperation, and the enforcement of social norms." *Human nature* 13:1–25.
- Frank, Robert H. 2011. "The strategic role of the emotions." *Emotion Review* 3(3):252–254.
- Gartzke, Erik. 1998. "Kant we all just get along? Opportunity, willingness, and the origins of the democratic peace." *American Journal of Political Science* pp. 1–27.
- Gartzke, Erik and Jon R Lindsay. 2015. "Weaving tangled webs: offense, defense, and deception in cyberspace." *Security Studies* 24(2):316–348.
- Gartzke, Erik and Kristian Skrede Gleditsch. 2004. "Why democracies may actually be less reliable allies." *American Journal of Political Science* 48(4):775–795.
- Gartzke, Erik and Kristian Skrede Gleditsch. 2006. "Identity and conflict: Ties that bind and differences that divide." *European Journal of International Relations* 12(1):53–87.
- Gibbs, David N. 1995. "Secrecy and international relations." *Journal of Peace Research* 32(2):213–228.
- Goldstein, Joshua S. 1991. "Reciprocity in superpower relations: An empirical analysis." *International Studies Quarterly* 35(2):195–209.

- Goldstein, Joshua S and John R Freeman. 1990. *Three-way street: Strategic reciprocity in world politics*. Chicago: University of Chicago Press.
- Goldstein, Joshua S and John R Freeman. 1991. "US-Soviet-Chinese relations: Routine, reciprocity, or rational expectations?" *American Political Science Review* 85(1):17–35.
- Goldstein, Joshua S and Jon C Pevehouse. 1997. "Reciprocity, bullying, and international cooperation: Time-series analysis of the Bosnia conflict." *American Political Science Review* 91(3):515–529.
- Gray, Kurt, Adrian F Ward and Michael I Norton. 2014. "Paying it forward: generalized reciprocity and the limits of generosity." *Journal of experimental psychology: General* 143(1):247.
- Grieco, Joseph M. 1988. "Anarchy and the limits of cooperation: a realist critique of the newest liberal institutionalism." *International organization* 42(3):485–507.
- Gummerum, Michaela, Masanori Takezawa and Monika Keller. 2009. "The influence of social category and reciprocity on adults' and children's altruistic behavior." *Evolutionary Psychology* 7(2):147470490900700212.
- Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath et al. 2001. "Cooperation, reciprocity and punishment in fifteen small-scale societies." *American Economic Review* 91(2):73–78.
- Holsti, Ole R. 2004. *Public opinion and American foreign policy*. University of Michigan Press.
- Huth, Paul K. 1988. "Extended deterrence and the outbreak of war." *American Political Science Review* 82(2):423–443.
- Huth, Paul K. 1999. "Deterrence and international conflict: Empirical findings and theoretical debates." *Annual Review of Political Science* 2(1):25–48.
- Hyde, Susan D. 2015. "Experiments in international relations: Lab, survey, and field." *Annual Review of Political Science* 18:403–424.
- Imai, Kosuke and In Song Kim. 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis* 29(3):405–415.

- Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1):51 – 71.
URL: <https://doi.org/10.1214/10-STS321>
- Jervis, Robert. 1976. *Perception and misperception in international politics: New edition*. Princeton University Press.
- Jervis, Robert, Keren Yarhi-Milo and Don Casler. 2021. "Redefining the debate over reputation and credibility in international security: Promises and Limits of New Scholarship." *World Politics* 73(1):167–203.
- Keohane, Robert O. 1986. "Reciprocity in international relations." *International organization* 40(1):1–27.
- Kertzer, Joshua D. 2022. "Re-assessing elite-public gaps in political behavior." *American Journal of Political Science* 66(3):539–553.
- Keysar, Boaz, Benjamin A Converse, Jiunwen Wang and Nicholas Epley. 2008. "Reciprocity is not give and take: Asymmetric reciprocity to positive and negative acts." *Psychological Science* 19(12):1280–1286.
- Kydd, Andrew. 1997. "Game theory and the spiral model." *World Politics* 49(3):371–400.
- Kydd, Andrew. 2000. "Trust, reassurance, and cooperation." *International Organization* 54(2):325–357.
- Kydd, Andrew H. 2007. *Trust and mistrust in international relations*. Princeton, NJ: Princeton University Press.
- Larson, Deborah Welch. 1988. "The psychology of reciprocity in international relations." *Negot. J.* 4:281.
- Leach, Colin Wayne, Martijn Van Zomeren, Sven Zebel, Michael LW Vliek, Sjoerd F Pennekamp, Bertjan Doosje, Jaap W Ouwerkerk and Russell Spears. 2008. "Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification." *Journal of personality and social psychology* 95(1):144.

- Leng, Russell J. 1993a. *Interstate Crisis Behavior, 1816-1980*. Cambridge University Press.
- Leng, Russell J. 1993b. "Reciprocating influence strategies in interstate crisis bargaining." *Journal of Conflict Resolution* 37(1):3-41.
- Lindskold, Svenn. 1978. "Trust development, the GRIT proposal, and the effects of conciliatory acts on conflict and cooperation." *Psychological Bulletin* 85(4):772.
- Marcus, George E, Michael MacKuen, Jennifer Wolak and Luke Keele. 2006. The measure and mismeasure of emotion. In *Feeling politics: Emotion in political information processing*. Springer pp. 31-45.
- McDermott, Rose, Anthony C Lopez and Peter K Hatemi. 2017. "'Blunt Not the Heart, Enrage It': The Psychology of Revenge and Deterrence (November 2017)." *Texas National Security Review*.
- Mellers, Barbara, Alan Schwartz and Ilana Ritov. 1999. "Emotion-based choice." *Journal of experimental psychology: General* 128(3):332.
- Miller, Steven V. 2022. "peacesciencer: An R Package for Quantitative Peace Science Research." *Conflict Management and Peace Science* 39(6):755-779.
- Nowak, Martin A and Karl Sigmund. 2005. "Evolution of indirect reciprocity." *Nature* 437(7063):1291-1298.
- Nowak, Martin A and Sébastien Roch. 2007. "Upstream reciprocity and the evolution of gratitude." *Proceedings of the royal society B: Biological Sciences* 274(1610):605-610.
- Oneal, John R. and Bruce Russett. 2001. "Clear and Clean: The Fixed Effects of the Liberal Peace." *International Organization* 55(2):469-485.
- Osgood, Charles E. 1962. "An alternative to war or surrender."
- Oskamp, Stuart. 1971. "Effects of programmed strategies on cooperation in the Prisoner's Dilemma and other mixed-motive games." *Journal of Conflict Resolution* 15(2):225-259.
- Perugini, Marco, Marcello Gallucci, Fabio Presaghi and Anna Paola Ercolani. 2003. "The personal norm of reciprocity." *European Journal of Personality* 17(4):251-283.

- Poznansky, Michael and Evan Perkoski. 2018. "Rethinking secrecy in cyberspace: the politics of voluntary attribution." *Journal of Global Security Studies* 3(4):402–416.
- Rabin, Matthew et al. 1992. *Incorporating fairness into game theory and economics*. University of California at Berkeley, Department of Economics.
- Rajmaira, Sheen and Michael D Ward. 1990. "Evolving foreign policy norms: Reciprocity in the superpower triad." *International Studies Quarterly* 34(4):457–475.
- Rathbun, Brian C. 2011. *Trust in international cooperation: International security institutions, domestic politics and American multilateralism*. Vol. 121 Cambridge University Press.
- Rathbun, Brian C, Christopher Sebastian Parker and Caleb Pomeroy. 2025. "Separate but unequal: Ethnocentrism and racialization explain the "democratic" peace in public opinion." *American Political Science Review* 119(2):621–636.
- Rathbun, Brian C, Joshua D Kertzer, Jason Reifler, Paul Goren and Thomas J Scotto. 2016. "Taking foreign policy personally: Personal values and foreign policy attitudes." *International Studies Quarterly* 60(1):124–137.
- Reiter, Dan and Allan C Stam. 2002. *Democracies at war*. Princeton University Press.
- Renshon, Jonathan and Ryan Powers. 2024. "Identity and the Social Construction of Reputation in World Politics."
- Rhodes, Carolyn. 1989. "Reciprocity in trade: The utility of a bargaining strategy." *International Organization* 43(2):273–299.
- Romano, Angelo, Daniel Balliet and Junhui Wu. 2017. "Unbounded indirect reciprocity: Is reputation-based cooperation bounded by group membership?" *Journal of Experimental Social Psychology* 71:59–67.
- Roskos-Ewoldsen, Andrew, Morgan Ellithorpe and Brandon J Kinne. 2026. "The Effect of Reciprocity on Trust: International Cooperation and COVID Aid." *Journal of Conflict Resolution* 70(1):115–139.

- Russett, Bruce M and John R Oneal. 2001. "Triangulating peace: Democracy, interdependence, and international organizations." (*No Title*) .
- Sanger, David E. 2012. *Confront and conceal: Obama's secret wars and surprising use of American power*. Crown.
- Schelling, Thomas C. 2008. Arms and influence. In *Strategic Studies*. Routledge pp. 96–114.
- Schroeder, David A, Julie E Steel, Andria J Woodell and Alicia F Bembenek. 2003. "Justice within social dilemmas." *Personality and Social Psychology Review* 7(4):374–387.
- Shinada, Mizuho, Toshio Yamagishi and Yu Ohmura. 2004. "False friends are worse than bitter enemies: "Altruistic" punishment of in-group members." *Evolution and Human Behavior* 25(6):379–393.
- Spence, Michael. 1978. Job market signaling. In *Uncertainty in economics*. Elsevier pp. 281–306.
- Tajfel, Henri. 1970. "Experiments in intergroup discrimination." *Scientific american* 223(5):96–103.
- Tajfel, Henri. 1974. "Social identity and intergroup behaviour." *Social science information* 13(2):65–93.
- Tomz, Michael and Jessica LP Weeks. 2021. "Military alliances and public support for war." *International Studies Quarterly* 65(3):811–824.
- Tomz, Michael, Jessica LP Weeks and Keren Yarhi-Milo. 2020. "Public opinion and decisions about military force in democracies." *International Organization* 74(1):119–143.
- Tomz, Michael R and Jessica LP Weeks. 2013. "Public opinion and the democratic peace." *American political science review* 107(4):849–865.
- Uslaner, Eric M. 2002. *The Moral Foundations of Trust*. Cambridge University Press.
- Ward, Michael D and Sheen Rajmaira. 1992. "Reciprocity and norms in US-Soviet foreign policy." *Journal of Conflict Resolution* 36(2):342–368.
- Ward, Michael Don. 1982. "Cooperation and conflict in foreign policy behavior: Reaction and memory." *International Studies Quarterly* 26(1):87–126.

- Weisiger, Alex and Keren Yarhi-Milo. 2015. "Revisiting reputation: How past actions matter in international politics." *International Organization* 69(2):473–495.
- Wilson, Warner. 1971. "Reciprocation and other techniques for inducing cooperation in the Prisoner's Dilemma game." *Journal of Conflict Resolution* 15(2):167–195.
- Wu, Junhui, Daniel Balliet and Paul AM Van Lange. 2016. "Reputation, gossip, and human cooperation." *Social and Personality Psychology Compass* 10(6):350–364.
- Yamagishi, Toshio, Nobuhito Jin and Toko Kiyonari. 1999. "Bounded generalized reciprocity: ingroup boasting and ingroup favoritism. Adv." *Group Process* 16:161–197.
- Yamagishi, Toshio and Toko Kiyonari. 2000. "The group as the container of generalized reciprocity." *Social psychology quarterly* pp. 116–132.
- Yarhi-Milo, Keren, Joshua D Kertzer and Jonathan Renshon. 2018. "Tying hands, sinking costs, and leader attributes." *Journal of Conflict Resolution* 62(10):2150–2179.
- Zheng, Wenqian, Shengli Cao, Yangyang Wang, Ke Yang, Yunyun Chen and Guoping Song. 2021. "The impact of social value orientation, game context and trust on cooperative behavior after cooperative video game play." *Psychological Reports* 124(3):1353–1369.