Of Ripple Effects and Reverberations:

Disinformation in a Two-Level Game

by

Michal Cantrell

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2024 by the
Graduate Supervisory Committee:

Timothy Peterson, Co-Chair
Fabian Neuner, Co-Chair
Jeffrey Kubiak

ARIZONA STATE UNIVERSITY

May 2024

ABSTRACT

Why, how, and to what effect do states use disinformation in their foreign policies? Inductive accounts variously address those questions, but International Relations has yet to offer a theoretical account. I propose Putnam's two-level game (1988) as a candidate theory. A rationalist approach that jettisons the unitary actor assumption, the model accounts for previous accounts' observations and suggests their interrelation and four overarching objectives. The model also generates novel implications about disinformation in foreign policy, two of which I test via separate survey experiments.

The primary implication is that states can use disinformation to encourage polarization and in turn can reverberate into commitment problems. A survey experiment tests the first link in that chain, arguing that disinformation's effects could be underestimated due to focus on belief outcomes; potential selection bias in active-exposure studies; and probable pre-treatment effects. It hypothesizes that passive exposure to novel political dis/misinformation has ripple effects on trust, affective polarization, and participation-linked emotions even among those that disbelieve it. It thus tests both the implication that disinformation can encourage polarization and that disinformation can be used to impact multiple potential outcomes at once.

The second empirical paper tests the latter links in the disinformation-commitment problem chain. Building on a study that found U.S polarization decreases U.K. ally confidence (Myrick 2022), it argues that polarization uniquely increases chances of voluntary defection and does so not only due to government changeover risk but also weakened leader accountability. It employs a causal mediation analysis on

survey experiment data to test whether a potential partner's polarization increases their perceived unreliability and in turn decreases public cooperation preference.

The commitment problem implication receives mixed support. The first experiment evidences no impact of partisan mis/disinformation on affective polarization, though that may be due to floor effects. The second experiment finds that polarization modestly increases perceived defection risk, but this increase is not necessarily strong enough to change public cooperation preference. Beyond those findings, the first experiment also uncovers that polarization may indeed have sociopolitical impacts on even those that disbelieve it, consistent with the multiple-outcomes implication.

ACKNOWLEDGMENTS

I owe much gratitude to many: my committee members for your feedback and encouragement since well before I started the dissertation process; ASU's School of Politics and Global Studies and Graduate Professional Association for funding the surveys used in this dissertation; Thu Nguyen for her patient help in figuring out the paperwork involved in using some of said funds; the anonymous heroes who participated in the studies; the SPGS Research Lab team for going out of their way to facilitate the pilot study; Jenna Roelle and Thorin Wright for helping me identify the timeline necessary to complete the program on-time (and figure out how to accomplish the coursework part of it); Erik Nisbet, Pew Research Center, Michael Bechtel, and Kenneth Scheve for having made their datasets and/or coding available for use; my friends for helping me develop my thoughts more gooder and prepare for the dissertation defense; and my family for your steadfast support.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES ix

CHAPTER

1 DISINFORMATION IN A TWO-LEVEL GAME 1

Introduction 1

What to Make of Disinformation in International Relations? 3

Disinformation in a Two-Level Game 15

Implications & Limitations 50

Conclusion 56

2 TESTING TWO IMPLICATIONS: PAPERS 2 & 3 61

3 POTENTIAL UNDERESTIMATIONS OF MISINFORMATION'S IMPACTS: MISTRUST, EMOTION, AND A SELECTION BIAS 62

Introduction 62

Misinformation: Scope and "Ripple Effects" 64

Method 81

Analysis 94

Discussion 122

Conclusion 131

4 POLARIZATION: HOW IMPACTFUL IS IT ON COMMITMENT PROBLEMS? 133

Introduction 133

CHAPTER Page

4. POLARIZATION: HOW IMPACTFUL IS IT ON COMMITMENT PROBLEMS? 133

Theory: Polarization, Defection, and Elite/Public Cooperation Preferences.....136

Empirical Approach Overview 152

Method: Experiment 1 (Polarization and Cooperation Hesitation).....154

Results: Experiment 1.....164

Method: Experiment 2 (Affective v. Policy Polarization).....181

Results: Experiment 2.....185

Discussion.....189

Conclusion 195

5. CONCLUSION 197

REFERENCES 200

APPENDIX

A GOOGLE SCHOLAR TRENDS: DISINFO.-RELATED PUBLICATIONS ... 231

B DISINFORMATIONAL METHODS 235

C RESULTS: NOVEL MEDIATION ANALYSIS (NISBET ET AL. DATA) 237

D DISCUSSION: MTURK TREATMENT VALIDATION SURVEYS 239

E IRB APPROVAL: PAPER 2 243

F SURVEY: TREATMENT VALIDATION 245

G SURVEY: MISINFORMATION EXPERIMENT 272

H ROBUSTNESS CHECKS: BELIEF 302

I ROBUSTNESS CHECKS: TRUST 305

APPENDIX Page

J ROBUSTNESS CHECKS: AFFECTIVE POLARIZATION 313

K ROBUSTNESS CHECKS: ANGER, FEAR 319

L ROBUSTNESS CHECKS: TPE 320

M DIFFERENCE-IN-MEANS: OPT-IN/OUT BY PARTISANSHIP 324

M INFERRED PARTISANSHIP BY MISINFO. HEADLINE 329

O NON-EFFECT OF SHARER PARTY PROVISION 331

P RESULTS: CORRELATIONAL ANALYSIS OF PEW DATA 335

Q RESULTS: PILOT STUDY 344

R IRB APPROVALS: PAPER 3 357

S RESULTS: EXPERIMENT 1 (FULL FIELDING SUBSAMPLE) 360

T ROBUSTNESS CHECKS: EXPERIMENT 1 368

U DISCUSSION: EXPERIMENT 2 “COUNTRY X” ANONYMIZATION 375

V ROBUSTNESS CHECKS: EXPERIMENT 2 379

W SURVEY: EXPERIMENT 1 382

X SURVEY: EXPERIMENT 2 401

LIST OF TABLES

Table		Page
1.	Two-Level Game Overview	24
2.	Disinformational Objectives and Supporting Strategies in a Two-Level Game	28
3.	Disinformational Methods (Obj. I)	35
4.	Disinformational Methods (Objs. II-III)	42
5.	Disinformational Methods (Objs. IVa-IVb)	45
6.	Treatment Validation Content	88
7.	Difference-in-Means: Likelihood of Headline Being True	89
8.	Estimated Impact of Exposure on Belief (Odds Ratios)	97
9.	Estimated Impact of Exposure on Trust (Odds Ratios)	101
10.	Estimated Impact of Exposure on GT (Odds Ratios, Subsample)	106
11.	Estimated Impact of Exposure on OPT (Odds Ratios, Subsample)	108
12.	Estimated Impact of Exposure on Affective Polarization (OLS)	111
13.	Estimated Impact of Exposure on Affective Polarization (OLS, Subsample)	113
14.	Estimated Impact of Exposure on Emotion (OLS)	115
15.	Estimated Impact of Exposure on Anger (OLS, Subsample)	115
16.	Estimated Impact of Exposure on Fear (OLS, Subsample)	117
17.	Estimated Impact of Exposure on TPE (Odds Ratios)	118
18.	Estimated Impact of Exposure on TPE (Odds Ratios, Subsample)	119
19.	Difference-in-Means: Opt-In/Out Status	121
20.	Distribution: (Dis)belief by Subgroup	130

Table Page

21. Polarization Treatment Operationalizations (Myrick 2022) 149

22. Variable Summary.....163

23. Performance on Factual Manipulation Checks.....164

24. Assessed Aurl Defection Chances by Incorrect Unreliability Answer.....166

25. Est. Impact of Treatment Grp. on Aurl Coethnicity Perceptions (Odds Ratios) 177

26. Est. Impact of Treatment Group on Aurl Real-World Traits (Odds Ratios) 178

27. Affective Polarization Operationalizations by Treatment Group 182

28. Est. Impact of Treatment Group on Perceived Partisan Agreement (OLS) 188

LIST OF FIGURES

Figure Page

1. Novel Mediation Analysis of Nisbet et al.'s Data 79
2. Partisan Favorability Ratings: Four Misinfo. Items 90
3. Means: Misinformation Belief 96
4. Partisan Favorability Ratings: Non-Misinfo. Headlines 99
5. Means: GT and OPT 101
6. Distribution: GT, OPT (Control Group) 102
7. Means: Affective Polarization 110
8. Means: Outgroup Party Like 110
9. Distribution: Outgroup Party Like Ratings (Control Group) 112
10. Means: Anger, Fear 114
11. Means: TPE for Misinfo. Influence on Ingroup/Outgroup Party 118
12. Theory: Polarization and Defection Risk 150
13. Theory: Observer State Response to Polarization 151
14. Design: Experiment 1 (Aurl/Zerm) 156
15. Means: Offer Acceptance by Treatment Group 167
16. Est. Impact of Polarization on Offer Acceptance (Odds Ratios) 169
17. Est. Impact of Polarization on Offer Acceptance (Odds Ratios, Subsample) 171
18. Means: Gov't Changeover, Accountability by Treatment Group 173
19. Means: General Defection Probability by Treatment Group 174
20. Means: Probability Zerm Invades if Reject Aurl Offer 175
21. Means: Reliability and Cooperation Preference Outcomes 186

Figure Page

22. Means: Reliability and Cooperation Preference Outcomes (Subsample).....187

23. Means: Perception of Party Agreement Frequency.....188

CHAPTER 1

DISINFORMATION IN A TWO-LEVEL GAME

Introduction

Research regarding factually inaccurate information is on the rise,[1] and one area of focus has been state use of disinformation in their foreign policies. With particular emphasis on Russia and China, foreign affairs experts have backwards-engineered probable disinforming state motives: creating confusion and discord, and increasing the probability a target state behaves in a preferred way (Al-Rawi and Rahman 2020; Dawson and Innes 2019; Kragh and Åsberg 2017; Kurlantzick 2020; Paul and Matthews 2016; Walker and Ludwig 2017; Wu 2019).

Remarkably, international relations (IR) offers no theoretical explanation to complement those inductive findings. Four exploratory works address disinformation in the IR context, and they suggest disinformation's place in IR theory is unresolved (Anzera and Massa 2021; La Cour 2020; Lanoszka 2019; Walker and Ludwig 2017a). Lanoszka even argues it likely is of little foreign policy utility, and the other three point to mismatches between candidate theoretical explanations and international disinformation as we know it, to include its coercive nature, targeting of both foreign elites and publics, engagement with non-security topics, and employment of falsely-attributed messages. In sum, the broad questions of "why, how, and to what effect do states use disinformation in their foreign policies?" are open from a theoretical perspective.

[1] See Appendix A for supporting data.

This paper proposes a theoretical answer to those questions, placing disinformation in the rationalist, international bargaining context of Putnam's two-level game (1988). This contextualization suggests states use disinformation to alter win-sets and uncertainty toward a general objective of achieving a higher net payoff to the disinforming state. Specific objectives beneath that general umbrella are concession-extraction ("get more"), agreement acceptance ("get something"), unilateral status quo revision ("take something), and status quo maintenance ("keep something" and "keep the other guy from getting something"). The mechanisms harnessed are persuasion, side payments, and responsiveness, with anticipated effects on preferences, participation, polarization, trust, leader popularity, and leader credibility.

The paper is divided into four sections. First, the paper discusses disinformation as a concept and analyzes the four above-cited works about disinformation in IR theory. This section introduces the concept of attributional disinformation, or communications regardless of veracity that employ false attribution. Second, it proposes Putnam's two-level game as a candidate theoretical home for disinformation in IR, arguing the phenomena's compatibility with the model and describing the model's operation. Third, it applies the model to state use of disinformation in their foreign policies, generating answers to the questions of why, how, and to what effect. Fourth, it discusses the utility of a two-level understanding of disinformation in foreign policy, to include its identification of multiple pathways to status quo change or maintenance, introduction of a commitment problem as a possible disinformational effect, suggestion that disinformation can be used to play a long-game, and explanation of disinformation's utility relative to information.

What To Make of Disinformation in International Relations?

Disinformation as a Concept

As one would expect for a topic under rapidly-increasing enquiry, disinformation has several conceptualizations. Most conceptualizations frame disinformation as involving falsehood or inaccuracy in terms of message content,[2] and they make claims about propagators as well. The content propagator must have a general intent to deceive (Bennett and Livingston 2018, 124; Giusti and Piras 2020, 2; Guess and Lyons 2020, 10; Guo et al. 2021, 2; Weedon, Nuland, and Stamos 2017, 5), or specific intent to deceive in order to cause harm (de Cock Buning 2018; Wardle 2018, 954).

The intentionality element often is the distinguishing criteria between misinformation and disinformation, though the exact interrelationship can vary. Some frame misinformation as unintentional inaccuracy and disinformation as intentional inaccuracy (Guo et al. 2021, 2; Tucker et al. 2018, 3; Weedon, Nuland, and Stamos 2017, 5). Others label all inaccurate information "misinformation" and designate intentional misinformation, "disinformation" (Fallis 2015, 402; Giusti and Piras 2020, 2; Guess and Lyons 2020, 10). In either taxonomy, though, intentionality is only distinctively characteristic of disinformation.

Cold War-era propaganda literature adds an important distinction regarding the other element of inaccurate content. Content could be inaccurate with respect to its factual claims, e.g., the body of the message contains claims that are objectively

[2] One exception is Zelenkauskaite (2022, 3–4), who opts to treat disinformation as "propaganda that involves affective, deflective, and misleading, rather than false information", with the affective element predominating. However, this distinction is oddly drawn, as the author maintains the standard definition of misinformation as unintentional inaccuracy.

inaccurate. Content also can be inaccurate with respect to attribution, e.g., someone could make objectively accurate claims or subjective claims, but misrepresent their identity (Jowett and O'Donnell 2018; Martin 1982). To differentiate these two disinformation types, I employ the novel terms "message disinformation" and "attributional disinformation."

The terms are not established elsewhere in disinformation literature, likely because the recent uptick in disinformation scholarship has largely arisen independent of the propaganda literature that established the distinction.[3] Reincorporating an attributional category of disinformation is important. Its omission has narrowly focused disinformation studies on *message content* at a time when mass digital communications have enabled states to obfuscate *messenger identity* at greater scale than they could before. Scholars and practitioners have examined these efforts (Al-Rawi and Rahman 2020; DiResta et al. 2019), but largely do not examine them alongside the propagators' simultaneous and overlapping message disinformation efforts.[4] I suggest that attributional and message disinformation are best understood in-tandem, as disinformers likely do not

[3] Propaganda literature also lacks such terms, but the distinction is contained in the white/gray/black propaganda taxonomy. White propaganda is accurate in content and attribution, while gray/black propaganda is partially-wholly inaccurate on both counts (Jowett and O'Donnell 2018; Martin 1982). One work, Zelenkauskaite (2022), approaches disinformation from the perspective of propaganda, but narrowly defines disinformation as being affective and misleading rather than false. This definition encompasses attributional disinformation ("trolling") but omits message disinformation.

[4] A Facebook report described false flag attacks as a potential example of disinformation, but their implicit inclusion of attributional disinformation is not echoed in disinformational literature (Weedon, Nuland, and Stamos 2017, 5). The academic articles closest to incorporating both elements of disinformational inaccuracy tend to be about "fake news." For example, an analytical essay argued scholars ought to replace "fake news" with the concept of disinformation, defining disinformation as "intentional falsehoods spread as news stories or simulated documentary formats to advance political goals" (Bennett and Livingston 2018, 124). However, this definition requires disinformation precludes the potential of factual inaccuracy solely with respect to propagator identity.

have separate teams for "making objectively inaccurate claims" and "making claims while pretending to be someone else."

As with all concepts, these conceptualizations of disinformation do not neatly transfer from theory to empirics. First, the element of intentionality to deceive or harm is a major measurement challenge (Vraga and Bode 2020). Inaccurate influence efforts by a semi-official organization are one thing; differentiating between good-faith content inaccuracy and message disinformation another.[5] Second, the intentionality requirement can complicate empirical inquiry with little payoff. For example, an inaccurate social media post is disinformation if its author knowingly published it but is transformed into misinformation when an unwitting user republishes it. The same piece of information is disinformation in the hands of one person and misinformation in the hands of the other. One could try to distinguish the two, but with what success rate and to what end? Third, the criteria of intent-to-harm is unnecessarily restrictive. It creates a scenario in which a state could knowingly publish inaccurate claims, but provided they did not intend harm, we ought not term it disinformation. Further, actors in international relations use harm (however one defines it) to pursue particular goals. Unless we conceive of disinforming actors as sociopathic, harm is not an end in-and-of itself.

Taking those challenges into account, I define disinformation broadly: inaccurate communication intended to deceive by the original propagator. In doing so, I omit the

[5] For example: In late August 2021, the United States announced it destroyed a car bomb in Kabul. A few days later, investigative journalists asserted the strike's target was an aid worker with large water containers. An official investigation confirmed the assertion (Aikins and Rubin 2021). Was the US government's initial news release disinformation? Or merely their being misinformed? Absent leaked or declassified official communications, researchers likely will not know.

most narrow of the intent elements (intent to cause harm), as do many other definitions of disinformation (Bennett and Livingston 2018, 124; Fallis 2015, 401; Guess and Lyons 2020, 10; Guo et al. 2021, 2; Weedon, Nuland, and Stamos 2017, 5). Unlike those other definitions, though, I replace the deceptive intent requirement with a narrower requirement that only the original propagator had deceptive intent. In this, my definition is most akin to that of Giusi and Paras ("…the product of the construction of a purposeful untruth" [2020, 2]), but differs in that I incorporate language of communication, which includes not only the message content, but the messenger (McQuail and Windahl 2015).

I do not omit intent to deceive for two reasons. First, contemporary academic and popular discourse uses "disinformation" more in reference to lying than factual mistakes. Redefining it otherwise provides little value to the conversation. Second, the empirical challenges do not undermine the two follow-on empirical analyses. The analyses are focused on disinformation's effects, and mis/disinformation are interchangeable in that area. Propagator intentionality is immaterial to the question of exposure impact.[6]

Disinformation in IR Theory: A Square Peg?

The overwhelming majority of literature on disinformation in foreign policy is inductive. Governments, journalists, policy professionals, and academics have cataloged instances of foreign disinformation efforts, and some have hypothesized underlying state intent from that data. The most expansive assessments of Chinese and Russian disinformation generally assign a broad objective of target states acting in a preferred way, with supporting objectives of encouraging: division; mistrust of international allies and one's

[6] One could argue that intentionally propagated deceptions could be more persuasive than good-faith mistakes; however, that undersells the persuasive abilities of well-intended mistakes.

national government/institutions; and the perceived competency, legitimacy, and attractiveness of the disinforming state (Harold, Beauchamp-Mustafaga, and Hornung 2021; Karlsen 2019; Kragh and Åsberg 2017). Other inductive works variously identify those supporting objectives and add a further supporting objective of encouraging the election of congenial officials/parties. Often, however, they do not address the supporting objectives' ultimate end (Al-Rawi and Rahman 2020; Beskow and Carley 2020; Dawson and Innes 2019; DiResta et al. 2019; Entous, Timberg, and Dwoskin 2023; Kurlantzick 2020; Paul and Matthews 2016; 2016; Treyger, Cheravitch, and Cohen 2022).

Absent, though, is framework to organize those patchwork observations. Granted that overall goal—which likely can be said of any foreign policy tool—how do the supporting objectives interrelate? Do they adhere to an internal logic captured in IR theory? If no, disinformation is a challenging corner case that could result in theory refinement if not theory-building. If yes, the internal logic can perhaps be leveraged to identify further dynamics around disinformation in foreign policy.

IR has yet to answer that question. Four works have addressed disinformation from the perspective of IR theory and concepts and converge on a conclusion that disinformation is a concept in search of a home. These four works are Walker and Ludwig (2017), Lanoszka (2019), La Cour (2020), and Anzera and Massa (2020), and they varyingly assess disinformation from the perspectives of big "ism's" (constructivism, structural realism, rationalism), and particular concepts (soft power, sharp power, propaganda, lies).

Walker and Ludwig (2017)—Sharp Power

The first scholars to address disinformation in its theoretical context were Walker and Ludwig (Walker 2018; Walker and Ludwig 2017b; 2017a). The authors argue that IR's concepts of hard and soft power are insufficient to account for Russian and Chinese informational influence efforts abroad. Being informational, they are inconsistent with hard power; but being "malign," "aggressive," and government-directed, they are inconsistent with soft power (Walker 2018, 18; Walker and Ludwig 2017b, 13). The authors propose a new concept to encapsulate these efforts: sharp power. Authoritarian states use sharp power to "pierce, penetrate, or perforate the political and information environments in the targeted countries" (Walker and Ludwig 2017b, 6). The target is public opinion and the goal is more distraction and manipulation than attraction, as the authoritarian states seek to reduce democracies' soft power more than increase their own (Walker and Ludwig 2017b, 6, 9). Disinformation in the sharp power construct is a tool toward those goals, alongside public diplomacy, influence over expatriates, and the creation of civil society organizations that are actually state organs.

Though disinformation is consistent with the sharp power concept, it exceeds its bounds. Sharp power is wielded by authoritarian states against democratic ones. Disinformation, on the other hand, conceptually lacks a regime type qualification. Sharp power describes a pattern of international behavior since roughly 2007 (Walker and Ludwig 2017b, 8), while disinformation preexists the phenomena Walker and Ludwig seek to explain. Thus, the concept of sharp power has some utility: it identifies that disinformation can flow from public opinion through institutional channels; it does not require disinformation's content be negative, as it can be used to promote candidates and

promote the disinforming state's soft power; and it suggests opponent soft power reduction as its predominating goal. But, it does not fully encompass the research question of "why, how, and to what effect do states use disinformation in their foreign policies?" Nor did it intend to, as it sought only to explain why particular authoritarian states use it (and other tools) in a particular way.

Lanoszka (2019)— "Disinformation in International Politics."

The next work was Lanoszka (2019). His article's purpose was to provide a theoretical answer to the question of whether disinformation in IR "works" (2). His answer—that it largely doesn't—drew perhaps too-hasty conclusions.

Lanoszka opens by arguing that disinformation is an ill fit in major isms of structural realism, rationalism, and constructivism. Structural realism, with its focus on material strength and security, narrowly understands information in terms of executing or subverting intelligence collection (5). Rationalism, he says, finds disinformation "oxymoronic" (5). It fits within the mechanism of signaling, but costly signals "distinguish sincere states from insincere ones," while costly disinformational signals do no such thing (5). Similarly, Lanoszka finds disinformation in a constructivist telling illogical. If states lack common norms, disinformation will be unpersuasive, and if they possess common norms, they need not disinform in the first place (5-6).

Lanoszka then asserts disinformation's goals as armament policy or alignment change, and rejects disinformation's utility for three reasons. First, anarchy incentivizes states to disbelieve communications from other states (7-8). Second, elites and domestic audiences are difficult to persuade, as their attitudes on any given issue are more a function of preexistent preferences and biases than new information received (9-11).

Even if attitude change were easy, foreign policy can be low salience and Lanoszka finds it unclear whether public opinion influences foreign policy at all. Third, even under conditions of polarization, states can adopt countermeasures like factual interventions, censorship, and counter-campaigns of disinformation. Ultimately, he concludes that states use disinformation for three potential reasons (though he largely dismisses the second): miscalculation; long-term negative effects on a target state's institutional legitimacy or public discourse; or posturing for domestic audiences. He also identifies the possible existence of "windows of opportunity" for disinforming states to exploit new information technologies before opponent states adapt new countermeasures (21).

Overall, the article is overly strong in its rejection of rationalism and disinformation's utility. First, the article approaches rationalism with the assumption of a unitary state actor. This assumption limits disinformation's utility to signaling but is unnecessary. A rationalist approach does not preclude the influence of domestic actors on foreign policy, and when one includes domestic actors, informational vectors other than signaling are apparent. Second, the asserted incongruence of disinformation and signaling does not engage with literature on the deceptive use of signals. While Fearon's (1997) initially discussed costliness as a way to increase certainty of a signal's veracity, this does not preclude costly bluffing. Costliness is about increasing perceived credibility, which is only a proxy for sincerity (Prins 2003; Wolford 2014).[7] Poker players with a weak hand

[7] Fearon argues that states have an incentive to bluff, but that they will never bluff when making clear threats, due to the high costs and risk of escalation (1995; 1997). However, he noted that states may bluff when making fewer clear threats. Others, such as Press, explored bluffing and found that bluffing is not an infrequent strategy, and its reputational costs seemed outweighed by balance of power realities (e.g., even if a state has previously bluffed, its threats can be credible if "they are consistent with important interests and are backed by substantial power" (2004, 169)

can bluff with costly bets, and a state with a weak position can bluff with costly signals (Gartzke 1999, 584).[8]

Third, the essay's dismissal of disinformation's utility is over-strong. Elites and publics certainly are skeptical recipients of foreign messages, but attributional disinformation sidesteps such skepticism.[9] Similarly, persuasion is only a difficulty if a disinforming state seeks to achieve wide attitudinal swings. More likely, disinformation campaigns succeed the same way domestic political campaigns do: increase or maintain mobilization among probable supporters and persuade undecided actors (Haselswerdt and Sides 2019; Norris 2006; Panagopoulos 2016).

Further, the article notes but does not sufficiently address counterarguments that states may seek to increase polarization, confusion, mobilization; or to have long-term negative impacts on institutional legitimacy and political discourse. For example, on the point of confusion, it does not consider that elites will not have near-instantaneous access to clarifying intelligence on all issues; and on the point of polarization, it dismisses potential impacts by citing an article on U.S. foreign policy from 2016-2017 whose conclusion stressed its tentativeness and whose results were challenged by later scholarship (Bentley and Lerner 2022; Dombrowski and Reich 2017; Friedrichs 2022a;

[8] Further, false signals can involve costs that are counterfeit. For example, the Allies in World War II used decoy equipment and radio transmissions to create the impression their primary 1944 invasion site was somewhere other than Normandy (Holt 2010, 540). The cost of redirecting materiel and men would be exorbitant if real, and thus increased the credibility of the bluff. Finally, states also can undersell their strength by failing to send a costly signal (Slantchev 2010). This apparent weakness decreases chances of opponent concessions but can translate into a tactical upper-hand if negotiations break down.

[9] Lanoszka eventually discusses covert campaigns, but with regard to his third barrier of countermeasures (12). He does not discuss how covert disinformation ameliorates the first barrier anarchy-rooted skepticism of foreign messages.

Yarhi-Milo 2018). Finally, the article overstates countermeasure effectiveness. Countermeasures can be somewhat effective, but vary on the strength of moderating factors like polarization and mistrust in countermeasure source (e.g., government, fact-checkers) (Guess, Lerner, et al. 2020; Hameleers 2020; Hameleers and van der Meer 2020; Loomba et al. 2021). They are not a "silver bullet" solution to disinformation.

La Cour (2020)— "Theorising Digital Disinformation in International Relations."

La Cour adopts a different approach, examining disinformation in light of key works and concepts rather than isms. The key works are those of Carr, Mearsheimer, and Nye (propaganda, lying, and soft power), and related literature she surveys are: "bullshit", post-truth, emotions, information warfare, and strategic narratives. Her key takeaways from said related literature were twofold. First, emotion and post-truth environments enable disinformation. Second, interstate disinformation occurs within a context of narrative warfare: a competition for narrative dominance (708-711).

La Cour then proceeds to analyze the aforementioned IR works. The first is E.H. Carr's *The Twenty Years' Crisis* (713). She deems three of its points compatible with disinformation: propaganda leverages ideology; propagandists' goals are increased support for a given policy or changed policy preferences; and propaganda's targets are domestic and foreign publics. She assesses the work has two shortcomings when intersected with disinformation: it assumes truth is more persuasive than lies and propaganda is less useful when non-attributed (713-715). Her analysis is sound, though her assertion that Carr assumes lies are harmful seems to misrepresent his argument.[10]

[10] He argued that deceptive propaganda is more persuasive the more it incorporates truthful elements—a claim with empirical support (Mourão and Robertson 2019).

The second work she analyzes is John Mearsheimer's *Why Leaders Lie: The Truth About Lying in International Politics* (715). She finds that he discusses domestic lying more than international lying but emphasizes Mearsheimer's conclusion that leaders are more likely to lie in times of crisis. She also deems the distinction between lying, spinning, and concealment important. Spinning and concealment are biased communications, while lying is asserting an untruth. In terms of gaps, she notes three. Mearsheimer assumes the lies are rightly attributed, which is inconsistent with online disinformation. Mearsheimer's accounting of international lying also focuses on leaders and security-related content, which is too narrow. Finally, Mearsheimer assumes foreign elites are the targets of international lies (715-717). Overall, La Cour's analysis of Mearsheimer's work is sound.

Joseph Nye is the final scholar whose work La Cour analyzes, and she finds his works on attractive "soft power" largely unhelpful. He anticipates deceptive propaganda will be counterproductive because it damages its author's reputation, and he does not consider the possibility of non-attribution. He also focuses on using public diplomacy for constructive ends, and therefore cannot account for disinformation's oft-negative messaging. One salient point he makes is that public diplomacy involves dialogue. La Cour notes that disinformation campaigns can have a feedback function in which disinformers research their target audiences; however, she notes that this almost certainly is not what Nye was talking about (717-718). As with Mearsheimer, La Cour's analysis of Nye's work is sound.

Overall, La Cour concludes that the literatures she surveyed offer tentative suggestions for disinforming state motives. She does not state what she believes those

proffered motives to be, but from her analysis, I would infer that they are: to shape views of the sender state or preferences regarding policy. She notes that a key shortcoming in current works is their inability to account for non-attribution. She ultimately concludes that the following major questions remain open: "how, when and why states and non-state actors use disinformation, and why disinformation appears to be ever more present in modern-day international politics" (719).

Anzera and Massa (2021)—Using International Relations Theories to Understand Disinformation

Anzera and Massa (2021) approach the question of disinformation from the perspectives of transnationalism, soft power, and media studies. They observe that technologically-aided transnationalism has enabled global information broadcasting by state and non-state actors and weakened state control of the information sphere. This "information overload" in a post-modern context allows states to use disinformation to influence foreign publics by propagating narratives (36-38). Information is thus elevated as an independent tool of power, not merely a subordinate one that supports material tools.

The authors' observation that transnationalism as facilitated by communications innovations have aided the cause of disinforming states is helpful. The bypassing of traditional media and governmental gatekeepers enables disinforming states to influence not only elites, but publics as well. Similarly helpful is their assertion of a disinforming state's broad goals: supporting the disinforming government and discrediting opponents. Finally, an important implication they identify is that soft power's effectiveness resides in the credibility of the messenger, so disinforming states will avoid attribution and seek to make the disinformation plausible with existing discourse in the target audience.

However, their work does not provide a theoretical grounding for why states disinform. Their analysis places disinformation in the realm of soft power, which has the purpose of influence via persuasion (rather than coercion). As disinformation involves deception, it is inconsistent with non-coercive soft power, as La Cour (2020) concluded. Indeed, disinformation's incompatibility with soft power is why Walker and Ludwig (2017) proposed the concept of sharp power. Further, their work does not offer answers to the questions of what mechanisms disinforming states harness, and to what effect.

Conclusions regarding Extant Literature.

The four surveyed works suggest that IR has yet to offer a comprehensive theoretical account of disinformation. The concept is not foreign to IR, but a compelling explanation of it remains outstanding. One major mismatch is how digitally-enabled disinformation is often non-attributional, sidestepping the potentially low credibility of adversary states. Another is how disinformation touches on issues not directly related to security issues like military capabilities and intentions. Two further still are how it can target both public and elite audiences, and how it is coercive, unlike soft power. A final related one is an assumption that disinforming states be non-democratic. The major questions remain: Why do states use disinformation, to what effect, and through which causal mechanisms?

Disinformation in a Two-Level Game

I turn to Robert Putnam's two-level game model to answer those questions. A rationalist account of IR outcomes that allows for imperfect rationality and affective influences, the model jettisons the unitary actor assumption that limited Lanoszka's rationalist assessment of disinformation. Further, a recent paper on public diplomacy suggests the model's suitability to the topic. The study applied the model to several states' foreign

policy communications to their domestic publics, and found it successfully accounted for state motives and outcomes (Bjola and Manor 2018). This increases confidence the model may similarly apply to states' external persuasive communications, to include deceptive ones.

Below I present the model, and then use it to build a theoretical account of disinformation as a foreign policy tool. The account handles the previous conceptual mismatches regarding attributional disinformation, breadth of disinformational topics, audience identity, coercion, and regime type. Importantly, it also offers answers to the three open questions of motive, mechanisms, and effects. Regarding motive, states have a general objective of increasing the probability of an optimum-payoff outcome[11], which can manifest in four ways depending on the context (pursuing concessions, pursuing agreement, supporting unilateral status quo revision, status quo maintenance). The how is by manipulating win-set size and uncertainty, through mechanisms of persuasion, side payments, and responsiveness. In terms of specifics, states can seek to influence preference change, preference mobilization/formation, politicization, popularity/good will, chief negotiator autonomy, and chief negotiator credibility. Those six factors are where researchers should look for immediate effects.

A Two-Level Game Approach to International Relations

The Game

Two-level games entered the IR lexicon in the late 1980s, an intersection of bargaining and second image literature (Putnam 1988). Robert Putnam proposed that international

[11] By optimum-payoff, I mean the maximum net difference between costs and benefits.

relations not be imagined as a one-level game, like chess, but as a two-level game. The first level, denoted Level I, is the international game. At that game sits each state's single international player, or "chief negotiator." There are other people at the international game—various officials and advisors can sit beside each chief negotiator—but there is only one player per state. This chief negotiator also sits at the second level (Level II) game, the domestic game. Around the domestic gameboard sit domestic elites, interest groups, and the public. The chief negotiator plays both games simultaneously. A move on one board by any player may enable or constrain a move on the other, and players' uncertain expectations about other players' behavior on both boards shapes their moves (433-436).

Imagined in real-time, this metaphorical game spirals into incomprehensible complexity (434).[12] So, Putnam does not imagine it in real time. Instead, he takes a bargaining approach, focusing on negotiations and the win-sets they involve. Each negotiating state's win-set is its range of acceptable outcomes, stretching from its ideal outcome to its minimally acceptable outcome. The wider the win-set, the greater the amount of acceptable bargaining outcomes for a given state; and the more multiple states' win-sets overlap, the more likely agreement is between the states.

Crucially, as this is a two-level game, state win-sets are determined by both their Level I and II players. Putnam incorporates this dynamic by assuming all Level I agreements are subject to ratification in each chief negotiator's Level II game. The

[12] Putnam does not use the term "infinite", however the two-level game metaphor is consistent with the concept of "infinite games" in which there is never an ultimate winner, simply relatively more advantaged players at different points in time (Gale and Stewart 1953; Karlin 1953; Sion and Wolfe 1957; Wolfe 1955).

ratification could be formal or informal, such as legislative approval or particular public opinion outcome, but it must occur. Thus, an agreement may fail not just when Level I players are unable to hammer it out or a Level I player voluntarily defects from an agreement, but also when Level II players refuse ratification and cause involuntary defection (435-441).[13]

Three factors that shape outcome probability, primarily by determining win-set size but also by altering uncertainty if one does not assume perfect information (441-442, 453). One factor is policy preferences/preference distribution (coalitions). These are the chief negotiator's preferences and the distribution of Level II player preferences.[14] Preferences are driven by the perceived costs/benefits of the status quo and potential bargain(s). The perceptions piece is key, as the model does not assume purely rational players working off perfect information. Costs/benefits are uncertain and probabilistic, and individual perception thereof will vary. Further, even in impossible conditions of perfect information, internationalist and isolationist players could list roughly the same items in a stylized cost/benefit table, but with the column headings swapped.

Distribution of those preferences (current or potential coalitions) is what matters for Level II win-set size, as ratification requires some proportion of active Level II players. The proportion size will vary by state and issue in question but will be of active players. Not all players are active. For example, most players are not incentivized to play

[13] For Putnam, "ratification" is any formal/informal influence that can approve or veto an agreement (e.g., labor union non-cooperation on an agreement as an example of ratification failure). (436)

[14] Putnam's initial description of this factor includes the assumption that Level I and Level II players' win-sets are identical (442). This limits the factor to Level II preferences/distribution only. However, he later relaxes the assumption, as the relationship between Level II players and their chief negotiator is principal-agent (456-457).

when issues have concentrated status quo costs/agreement benefits (444-445).[15] As the costs/benefits element is one of perception, Putnam identifies politicization as a factor impacting player mobilization. An issue becoming the object of political contention renders the perceived costs/benefits more diffuse, and thereby increases participation by players not otherwise incentivized to participate.

Putnam does not explain this dynamic further, but presumably the diffusion arises because politicization enmeshes the issue in prominent political narratives and identities. It increases awareness of a given issue, mobilizing some players who abstained out of ignorance; and through contestation, it can reframe the costs/benefits of an otherwise largely unimportant issue to have broader resonance with political loyalties and ideologies. Thus, mobilization is linked with preference formation. Inactive players lack either crystallized preference, a strong enough preference to merit the costs of participation, or trust that their participation will do anything. Active players must have all three. So, an inactive player may be mobilized without preference formation (if their assessment regarding participation's utility improves), but otherwise, mobilization arises through preference formation—in content and in strength.

The second factor is Level II institutions. Whether formal or informal, domestic institutions are what link Level II preferences to Level I outcomes via ratification (437, 448-450, 459). An example of a formal institutional factor is the proportion of "yes" votes required for a piece of legislation to pass; an example of an informal factor is

[15] Within those that do play, the relative rank-ordering of the preferences matter, not simply their top preference. For example, if players disagree on the preferred outcome but most agree that the potential bargaining outcomes are preferable to the status quo, then revision is more likely.

Japanese preference for wide domestic consensus before acting. Importantly, the model assumes that ratification matters not only for democracies but also for autocracies.[16] No chief negotiator is wholly free from domestic constraint.

The shadow of future ratification, combined with three negotiator incentives, produces a dynamic Putnam relates to principal-agent theory but broader literature terms "responsiveness." Chief negotiators are responsive to Level II preferences because there's "[something] in it for them" (457). Three things that can be in it for them are (1) accrual of greater political resources at Level II, or minimization of political resource loss; (2) accomplishing Level II policies that are otherwise infeasible; and (3) pursuit of national interest as they conceive of it (437, 455-458).[17] The predominating example of the first incentive is continuation in power, as Putnam's arguments often refer to electoral incentives or power entrenchment as a driving negotiator consideration (437, 449, 455-456). The second incentive appears entangled with the other two, as accomplishing domestic policies can support both one's domestic political standing and conception of the national interest.

Autonomy, whether structural or de facto, is the conditioning factor on responsiveness. Chief negotiators that are more autonomous, more insulated from domestic influences, can be less responsive without threatening their political standing and policy agenda, and have wider win-sets and weaker bargaining positions as a result

[16] Putnam critiques and intentionally deviates from "Gamma paradigm" scholarship that assumed that only democratic negotiators experienced domestic constraint (434).

[17] Theses motives are akin to those of Fenno (1973). Fenno, echoed later by Aldrich and Rohde (2001), ascribed three motives to U.S. members of congress: reelection, power in government (influence), and good policy.

(451). Less autonomy, on the other hand, means more responsiveness to Level II players, a relatively narrower win-set, and a stronger case against being able to make concessions. For example, an entrenched dictator or a wildly popular elected leader who can lead public opinion is more autonomous than a precariously-positioned dictator or less-popular elected leader, *ceteris paribus*, and will have a weaker bargaining position (449, 451).

The third and final factor is a chief negotiator's strategy with respect to the previous two factors and uncertainty. What is the optimal size of an opponent's win-set, and one's own? And how can one address or even harness uncertainty to pursue one's goals? The answer to those questions varies based on many factors, but ultimately are governed by the rationalist assumption that states engaged in negotiations want to change the status quo in a way that achieves optimum payoffs for themselves. Given that, the optimum size of an opponent win-set is clear: "[e]ach Level I negotiator has an unequivocal interest in maximizing the other side's win-set" (450). The wider an opponent's win-set, the closer the bargaining space is to one's ideal outcome.

With respect to one's own win-set, however, one could seek to widen or narrow it. Widening it increases the chance of striking a bargain. The other state is more likely to accept due to increased win-set overlap, and one's own Level II players are more likely to ratify. On the other hand, negotiators with larger win-sets could be asked to make more concessions and move within their win-set towards the opponent state's preferred outcome. Negotiators may thus prefer to have a smaller national win-set for greater bargaining leverage, but at greater risk of Level I non-agreement and Level II rejection (437-440).

Putnam's discussion is less developed regarding strategies toward uncertainty but suggests two elements a chief negotiator could manipulate. First is uncertainty about a win-set size. Modification of that sort naturally accompanies efforts to actually widen or narrow win-sets; but, being perceptual, it can also be pursued independent of actual win-set modification. Second is uncertainty about a chief negotiator's credibility. If uncertainty of chief negotiator credibility drives concerns of voluntary defection, reducing concerns about a chief negotiator increases the probability of agreement.

In terms of tactical implementation of those strategies, Putnam looks to persuasion and/or side payments, and identifies a few illustrative approaches. To extract concessions, a state could feign uncertainty that an opponent state's win-set is wide enough to support ratification (453). To narrow win-sets, states can seek to rally players in support of a particular bargaining position (450). This may actually narrow the win-set by changing the distribution of preferences, or simply decrease uncertainty that the win-set is narrow. Similarly, a state can intentionally misrepresent their win-set's narrowness to extract concessions, leveraging the inherent informational asymmetry between themselves and their opponents (453).

To widen win-sets, a state could again use persuasion, or alternatively use a side-payment at Level I or Level II (450, 454). Importantly, side payments of any stripe cause coalitional realignment not by changing preferences, but by changing the object of preference (447). This draws a distinction between side-payments and persuasion. Side-payments alter the object of preference, while persuasion alters preference.[18] Side-

[18] Putnam employs "persuasion" broadly, to describe any state effort to shift yes/no ratification odds (e.g., he describes side-payments and persuasive messages as persuasive). The distinction I draw is

payments restructure the costs/benefits themselves while persuasion alters perceptions of the costs/benefits.

Persuasive messages' impact is conditioned by a few factors. First is the audience's "generic good will" toward the communicator (451-452). To re-use a previous example, the higher the domestic popularity of a given chief negotiator, the more persuasive ability they possess. This functionally gives them greater autonomy, and for this reason, chief negotiators are always incentivized to bolster their opponent negotiator's domestic standing, as it increases the potential width of the opponent state's win set. Also for this reason, the persuasiveness of foreign state communications can be increased if the state is viewed with positive or negative affect (e.g., ally v. adversary) (456).[19]

Second, if the communicating state in question is a foreign one, it must reach its target audience in another state (455). Putnam acknowledges this is more difficult than reaching one's own domestic public, but demonstrates it nevertheless occurred with regularity in his pre-internet context (454-456). Third, it depends on a foreign state's ability to understand its opponent's Level II situation (453-454). One's ability to persuade is contingent on one's ability to overcome the inherent informational disadvantage regarding an opponent's Level II preferences and preference distribution.

---

not foreign to the model; it simply uses the language of persuasion more concisely in accordance with persuasion scholarship.

[19] Putnam specifically says, "[I]nternational pressure is more likely to reverberate negatively if its source is generally viewed by domestic audiences as an adversary rather than an ally" (456). In isolation, this could be read as referring to audience favorability, learned trust, or necessity. However, he cites an affective theory (cognitive balance theory) as justification for the observation, which narrows the interpretation to favorability (Heider 1946; 1982; Rosenberg and Abelson 1960).

When a state attempts side payment or persuasive interventions in their opponent's Level II game, they are seeking to cause domestic "reverberations" that are strong enough to impact Level I gameplay. This approach comes with risk. Reverberations may be positive for the sender, with persuasion and/or side payments widening the opposing state's win-set toward the sender's ideal point. Such an outcome is more likely when the states have close relations and the issue in question is economic rather than political-military, presumably because Putnam thinks the latter to be higher-stakes. On the other hand, reverberations also can be negative, as Level II players sometimes receive foreign communications backlash. Putnam suggests negative reverberations are empirically less likely due to messengers being strategic actors: they only message if they think it has a good chance of working (456).

Taking a broad view, then, Putnam's model identifies that bargaining outcomes depend on two overlapping factors of win-set size and uncertainty (Table I). Win-set size drives the probability of engaging in negotiations in the first place, and if negotiations are engaged in, the probability of agreement, ratification, and concessions. Uncertainty about win-set size drives

Table 1. Two-Level Game Overview

Outcome Determinants	**Factors to Influence**	**Mechanisms**	**Points of Manipulation**
- Win-set size - Uncertainty	- Preferences/distribution (*Level I, II*) - Institutions (*Level II*)	- Persuasion - Side payments - Responsiveness	- Preference change - Player mobilization/ pref. formation - Politicization - Good will/popularity - Chief negotiator autonomy - Chief negotiator credibility

concerns of involuntary defection via ratification failure, and uncertainty about the other chief negotiator's credibility drives concerns of voluntary defection. Both win-set size

and uncertainty are shaped by player preference (Level I)/preference distribution (Level II); Level II institutions; and chief negotiator strategy.

Whatever strategy a chief negotiator adopts, they pursue it via three general mechanisms of persuasion, side payments, and responsiveness. These, in turn, host six overlapping potential points of manipulation. These are preference change, player mobilization/preference formation, politicization, good will/popularity, chief negotiator autonomy, and chief negotiator credibility. Importantly, a chief negotiator can seek to influence not only the Level I game but also Level II games—both theirs and their opponent's. The goal of such influence is to encourage reverberations via responsiveness into the Level I game, and the utility of these foreign influence efforts is conditioned on the messaging state's ability to reach the target audience and understand its context (preferences, distribution thereof).

Compatibility, Utility of Framework

Disinformation is compatible with a two-level model. The model's informational vectors provide it a conceptual home. States can seek to alter payoff size and ratification chances through communication, and the model does not require the communication be accurate. Rather, Putnam even explicitly provides examples of how states can use inaccurate communications.

In terms of utility, there are two questions. Is a two-level model suited to the research question, and how does it compare to previous theoretical approaches? To the first, I answer yes. The model is comprehensive enough to allow a researcher to examine why, how, and to what effect states use disinformation in foreign policy. This answer stands against a recent critical appraisal of the two-level game and questions of foreign

policy. Noone (2019) praises the model's ability to explain negotiations' outcomes, but says it "probably will not tell us very much about why [states] chose to arrive at the bargaining table in the first place, nor will it say a lot about states that chose not to come to the bargaining table at all" (178). On the contrary, as I argue in the next section, Putnam's model does suggest motivations and methods of states not directly engaged in negotiations. To be sure, his application of the model was on negotiations, but the model itself has broader application than that singular swathe of analysis.

Regarding the second question of utility, the two-level approach promises a more thorough telling than previous ones. Its novelty lies not in its constituent parts, but in their sum. Putnam allows states to informationally influence other states at both Level I and Level II. Use of information in bargaining games at Level I is not novel, as signaling is a central concept in such literature (Fearon 1995; 1997; Morrow 1994; Powell 2002; Quek 2016; 2021; Spaniel and Malone 2019). Similarly, the ability of foreign actors to informationally "reach around" opponent negotiators to influence a foreign audience is not novel, as discussed in Carr (1939), Walker and Ludwig (2017), Lanoszka (2019), La Cour (2020), and Anzera and Massa (2020). What is novel is for both avenues of influence to coexist in a single model. Further novel is the model's non-requirement that the "reach around" information to be attractive (Anzera and Massa 2021; La Cour 2020; Nye 2008), accurately-attributed (La Cour 2020; Lanoszka 2019; Mearsheimer 2011), focused on security issues (La Cour 2020; Mearsheimer 2011), and wielded by an autocratic state against a democratic one (Walker and Ludwig 2017). Combined, this allows for a comprehensive telling of state use of disinformation in their foreign policies.

Theory: Disinformation in a Two-Level Game

The argument can now shift to the research questions. What is the purpose of disinformation as a foreign policy tool? What mechanisms does it utilize? And, to what effect?

Disinformation Objective

Disinformation fits within the three factors that shape outcome probability. The preference and institution factors are the domains in which disinformation functions via the model's three main mechanisms, which are discussed in the next section. The remaining factor of chief negotiator strategy points to the overarching goal of disinformation as a foreign policy tool: altering win-sets and uncertainty to increase chances of a preferred outcome.

The manipulations' preferred outcomes vary (Table 2). Sometimes the objective is to "get more" (ratification of a payoff-optimizing agreement) or just to "get something" (ratification of an agreement better than the status quo). Other times, it is "getting something for less:" unilateral status quo revision when the revision seems higher payoff than likely negotiation outcomes. Still other times it is maintaining the status quo, which can consist of "keeping what you have" and "keeping the other guy from getting something" by spoiling opponent cooperation.

Table 2. Disinformational Objectives and Supporting Strategies in a Two-Level Game

Status	Objective	Bargaining	
		Element Modified	**Modification**
Party to Potential or Current Negotiations	I. GET MORE *Increase chance of opponent concessions*	Win-set (*opponent*)	Widen
		Win-set (*own*)	Narrow
		Uncertainty (*own*)	Increase
	II. GET SOMETHING *Increase chance of opponent ratification*	Win-set (*opponent*)	Widen
		Win-set (*own*)	Widen
	III. TAKE SOMETHING *Increase chance of unilateral status quo revision*	Uncertainty (*opponent*)	Increase
		Win-set (*opponent*)	Widen
	IVa. KEEP WHAT YOU HAVE *Decrease chances of status quo revision attempts*	Uncertainty (*opponent*)	Increase
		Win-set (*opponent*)	Widen
Not Party	IVb. KEEP THE OTHER GUY FROM GETTING SOMETHING *Decrease chances of opponent cooperation*	Win-set (*opponent*)	Narrow
		Uncertainty (*opponent*)	Increase

The first two objectives of getting more and getting something are explicit in the model. States simultaneously pursue both seeking to widen opponent win-sets. In the broadest sense of theoretical possibility, a state could use disinformation to widen opponent win-sets in perception only (e.g., astro-turfing); however, such a goal would be counterproductive. An opponent overestimating the width of their win-set risks ratification failure. Reality is only so elastic for so long (Baum and Groeling 2010), and the true range of the win-set will assert itself in the ratification stage. Thus, disinformation is useful in negotiations to widen opponent win-sets in actuality.[20]

[20] The language of "actual" and "perceived" win-sets is useful though limited. In reality, "actual" win-sets are inaccessible, and perceptions of win-sets are what Level I negotiators must work with. I do not assume perfect information but find the actual-perceived distinction important. Speaking solely of uncertainty obscures the point that states can seek to substantively change win-set ranges with disinformation rather than simply use disinformation as an uncertainty-reducing (though false) signal of a win-set's range.

States face a tradeoff between opponent concessions and ratification chances regarding their own win-set. They can falsely narrow it to increase chance of opponent concessions (Obj. I), something they can also pursue by overstating their uncertainty about their opponent's commitment. Or, they can falsely widen their win-set to increase chance of Level I agreement (Obj. II). False win-set widening requires a state eventually defect from an agreement, and so pursuing that strategy requires one of three conditions. A chief negotiator must have a high degree of autonomy from the Level II game, the consent of the ratification majority, or an agreement structured in such a way that benefits are reaped before the ratification stage. A high degree of autonomy allows the chief negotiator to lead or manage Level II preferences, which reduces risk of involuntary defection at the ratification stage. Level II player consent does the same. Finally, some immediate payoff grants the state a window between agreement and ratification to accrue benefits before reneging.

These two objectives are explicit in the model's explication of formal negotiations. The remaining two emerge when one applies Putnam's model to the broader context of all interstate relations. Does a two-level approach have anything to say about situations in which a state does not engage in formal negotiations? Yes. It suggests states may pursue unilateral status quo revision and status quo maintenance (Objs. III-IV). A few points of the model point to the former. First, the model presupposes a core motive for all actors: payoff-optimization (under conditions of uncertainty). Second, the model indicates a chief negotiator will abandon or forgo negotiations in two situations: when they will not result in status quo revision due to lack of win-set overlap; or, when they require concessions the chief negotiator is unwilling to make. Combined, these suggest

that a state presented with a situation in which a desired status quo revision is unlikely to be achieved through negotiations may unilaterally seek to achieve the revision, given favorable payoff calculus.

In bargaining terms, unilateral status quo revision is a state choosing a particular outcome on a bargaining spectrum and implementing it without the consent and cooperation of the other parties. Helpfully for the revisionist state, a unilateral status quo change may not be near-instantaneously evident to other players. Thus, uncertainty may initially be an ally, as they can delay or even prevent another state from responding to the revision until it is too late. Another potential ally is widening opponent and potential opponent state win-sets to encompass the new status quo location. This decreases risk of contestation by the "robbed" party and bystander intervention by the potential opponents.

Also helpful for the revisionist state is the fact unilateral revision can widen the win-set of the "robbed" state, further reducing the revision's cost to the revisionist state. This is because one of the perceived costs that contribute to win-set range is that of the status quo. The more costly one perceives it, the wider one's win-set. Thus, if the "robbed" state assesses the new status quo is less costly than seeking to regain the old one, they may accept the revision as a *fait accompli*. This de facto ratification is more likely if the state's expanded win-set encompasses the new status quo location on the bargaining spectrum. Overall, these dynamics suggest that states will only pursue unilateral status quo revision if they assess contestation as improbable, or if the assessed

costs of revision defense do not lower the new status quo's payoff to the point of inutility.[21]

The second outworked objective is status quo maintenance (Obj. IVa-b). This objective can be found in the shadow cast by Putnam's discussion. He elaborates on chief negotiator efforts to widen opponent win-sets, as this increases chance of bargaining success; but win-set narrowing is also possible in the model. Is there a scenario in which narrowing opponent win-sets would be of use? Yes, evidently when one does not seek the success or even the conduct of negotiations. That is, when one seeks to maintain the status quo.

A state could have this objective in two respects. The first is when the state is party to a given status quo and wants to "keep what they have." If they prefer the current configuration of one of their international arrangements, they may seek to forestall foreign revision efforts by narrowing the opponent state's win-set. No bargaining space? No negotiations. They also may seek to manipulate their opponent's uncertainty. Increasing uncertainty about the true status quo—that is, misleading an opponent about the current situation—can forestall an opponent response to the revision.

The second respect in which states can seek to maintain the status quo is to seek to maintain it between other states, that is, to spoil their cooperation or "keep the other guy from getting something." A state has this objective when it is not party to ongoing or

[21] This explanation of unilateral status quo revision is simpler than later bargaining models of war but will suffice for the purpose at hand. A few of the differences: Putnam's bargaining spectrum consists of win-sets, which includes status quo costs, revision benefits, and defection risks. Other bargaining models introduce indifference points, current status quo location, probability of winning in case of bargaining failure, and probability of defection.

potential negotiations and prefers they fail. In the two-level metaphor, there are many players in the first circumstance. Many or even most bilateral/multilateral foreign policy engagements do not involve most Level I players. Within that majority subset, chief negotiators that find others' potential coordinated gameplay to be contrary to their own gameplay could seek to spoil that coordination.[22] For example, Russia is not party to NATO's engagement with non-member European states, but it certainly has a preference that those engagements not result in new NATO memberships. It has incentive to narrow the negotiating parties' win-sets, and to increase their defection concerns.

In sum, in a two-level game, states use disinformation to support (1) their general objective of altering win-sets and uncertainty to increase chances of a preferred outcome, and (2) four supporting objectives. States can seek to increase chance of opponent concessions or ratification ("get more" and "get something") by widening opponent win-sets. They also can pursue concessions by misrepresenting their own uncertainty and own win-set's narrowness and pursue ratification chances by overstating their own win-set's width. States also can pursue unilateral status quo revision ("take something") when that appears higher payoff than the current status quo or likely negotiated revisions. Finally, states can also pursue status quo maintenance, with two subtypes. When party to a status quo, they can seek to "keep what they have" and forestall revision attempts. When not party, they can seek to "keep the other guy from getting something" by spoiling opponent cooperation. They can pursue both by seeking to narrow opponent negotiators' win-sets,

[22] In a broad sense, one could imagine a non-party state could also seek to widen other states' win-sets. But, that inference does not travel well to lower levels of abstraction. If a state has a strong enough preference for cooperation success that it risks potential negative reverberations by interfering in other states' domestic and international politics, one would imagine it is party to the negotiations.

and when they seek to spoil other states' cooperation, they can also seek to increase defection concerns.

Mechanism and Methods

In this section, I identify and explain a variety of disinformational methods by which a state can pursue those objectives. The model is such that I cannot affirm the list is exhaustive. It produces no master matrix that systematically intersects the modified bargaining elements (both "my" and "their" win-sets and uncertainty regarding commitment), the points of view (mine and theirs), and the various points of manipulation (preference change, preference mobilization/formation, politicization, popularity/good will, chief negotiator autonomy, chief negotiator credibility).[23] Beyond the dimensional challenges, the categories themselves are often mutually reinforcing and can operate in combination, which compounds the complexity and confounds conclusive derivation.

Instead, the methods described here are drawn from asking how a state could use disinformation to accomplish those four objectives through (1) persuasion, side payments, and responsiveness; given (2) the model's six points of manipulation. After considerable thought, I have not been able to generate more methods than thirteen described below but acknowledge this could as easily be a sign of limited creativity as of completeness. At the very least, the methods illustrate the mechanisms and points of manipulation. They are organized by objective, with some methods repeating between

[23] To make a confident derivation more plausible, I could simplify the model's outputs by assuming perfect information and rationality or assuming the chief negotiator is wholly responsive to their Level II players. The scenario that results is so unrealistic that it undermines the utility of model's application.

objectives, and are listed in entire in Appendix B. Examples provided are drawn from John Mearsheimer's catalog of international lies (2011), two scholarly sources on international deception (Brown, Lupton, and Farrington 2019; Sartori 2002), and various historical sources..

Objective I ("get more") can be pursued via six methods (Table 3). One is false side payment assurances. A state can offer fallacious side payment assurances to encourage an opponent state to accept a higher-cost agreement than it would otherwise prefer. The mechanisms this method and all the other methods rely on depend on the game level of the target audience. At Level I, it relies on persuasion and side payments, as the side-payment mechanism cannot be separated from persuasion when side payment assurances are false. At Level II, this relies on all three mechanisms. Domestic players must be persuaded of a false side payment, and the resultant preference distribution shift must be strong enough that the chief negotiator will be responsive to it.

This method is high risk. Reneging on the side-payment could damage the disinforming state's reputation and increases risk of opponent defection. The risk could be somewhat mitigated by the complexity of implementing agreements ("I tried my best, but the domestic tides shifted") and the possibility the side-payment is difficult to monitor. Eventually, though, the reneging will surface, at best having purchased time for the disinforming state to extract a benefit it wanted from the agreement. This leaves disinformation in the area of side payments a pursuable but high risk-option—a risk perhaps indicated by my inability to find an example of its use.

Table 3. Disinformational Methods (Obj. I)

Objective	**Bargaining Spectrum**		**Method** **: unique methods* [1, 2, 3, 4a,45b]*: other obj. a method supports*	**Points of Manip.**				
	Element Modified	**Modifi-cation**		**Pref. Change, Mobiliz./Form.**	**Popularity**	**Politicization**	**Chief Negotiator Autonomy**	**Chief Negotiator Credibility**
I. GET MORE *Increase chance of opponent concessions*	Win-set (*opponent*)	Widen/ shift towards own	False side payment assurances [2]	X				
			Pro-agreement/anti-status quo disinfo [2,3]	X				
			Promote own soft power [2,3]	X	X			
			Promote sympathetic/undermine antagonist Lvl I/II candidates [2,3,4a-b]	X				
	Win-set (*own*)	Narrow	Misrepresent own bargaining range [2]	X				
	Uncertainty (*own*)	Increase	False concern of opponent defection *	X				

Note: I combined two of the six points of manipulation (preference change, preference formation/mobilization) into a single column for ease of display.

The second method also seeks to modify the opposing state win-set, but at lower risk. It is: targeting Level I and Level II opponent audiences with pro-agreement/anti-status quo disinformation. This has the intended effect of widening the opponent win-set via persuasion (Level I target) and/or the persuasion/responsiveness combination (Level II target). This would look like a state generating message and/or attributional disinformation highlighting the costs of the status quo and/or benefits of a given agreement. The content of the messages could be specific to the given policy, or general, such as content undermining isolationism.

The propaganda activities of the United Kingdom during World War II provide an example. The blandly-named ‘British Security Co-ordination” (BSC) engaged in many activities in the Western Hemisphere, to include disseminating news via witting and unwitting journalists (J. N. Brown, Lupton, and Farrington 2019). In the U.S., one of the main goals of these efforts was to encourage U.S. intervention by undermining the then-

powerful isolationist movement—that is, by undermining the pro-status quo movement (Stephenson 1999; Ignatius 1989).[24] For instance, some of its journalist-fronted attributional disinformation sought to emphasize the self-interested rather than selfless motives of some isolationists, highlighting their ideological and financial sympathies with disliked Nazi Germany. One BSC-authored story highlighted isolationist American businesses' friendliness with a German businessman who seemed an unofficial liaison with the Nazi government (Hemming 2019; *New York Herald Tribune* 1940). Put in the model's terms, through such communications, the BSC sought widen the United States' Level I and Level II win-sets by increasing perception of status quo's hidden costs.[25]

The BSC example also highlights how states can use attributional disinformation to sidestep credibility concerns while communicating with a foreign Level II audience. The message content can but need not be deceptive—it could be subjective or even accurate. The attribution, however, will not be the authoring state or a representative thereof. This method remains risky, but less so than false side payments. If unmasked, it

[24] The state of the histories around the BSC introduce a note of caution, as there is evidence of embellishments, and original files have never been released other than the potentially self-serving internal history. However, critiques of the mythos of the BSC suggest that the embellishments tend to regard the scale of the contributions of the BSC's head more than the particulars of its activities in the United States (Naftali 2012; Charles 2000; T. Hoffman 2002).

[25] There are more potential examples. In 1940, a rumor circulated that the United Kingdom had repelled a German invasion on its coast. This rumor may have been intentionally propagated by the United Kingdom to encourage U.S. increased commitment to the lend-lease act (Hayward [1994] 2016). In the early 1950s, the United States encouraged a European collective defense agreement by falsely denying that the agreement would permit the U.S. to withdraw troops from Europe (Mearsheimer 2011, 42). In 2002-2003, the United States encouraged coalition-formation by claiming Iraq pursued weapons of mass destruction. In 2005, the United States inaccurately identified North Korea as the origin of nuclear materials that Libya purchased (Mearsheimer 2011, 38). Present-day, China uses false attribution social media accounts in Taiwan to promote China-Taiwan unification (Dickey 2019). These examples demonstrate the challenges of attribution and of differentiating between disinformation and a mistake.

may cause negative reverberations in the opponent Level II game, perhaps counterproductively narrowing the opponent win-set. But *ceteris paribus,* it has lower probability of being unmasked than false side payment assurances.

The third method is promoting or defending one's own soft power. The concept of soft power postdates Putnam, but Putnam captured its internal logic in his discussions of good will, popularity, and standing (450-451). He noted the more popular a chief negotiator is, the wider their domestic win-set because they have an easier time persuading Level II players. He also noted state communications have greater chance of positive reception by a foreign audience to the extent the audience views the state with positive affect. These observations describe the influence of soft power on bargaining (Gallarotti 2011).[26] It widens or shifts others' win-sets to be closer to one's own ideal because the other party's preferences are endogenous: soft power shapes their preferences. As a result, states with greater soft power have greater chance of reaching not only a Level I agreement, but a more favorable Level I agreement. This observation suggests a state could reap bargaining benefits by seeking to increase their own soft power, as Walker and Ludwig argued sans the bargaining context and implications (2017a, b; 2018).

The use of disinformation to influence soft power widens the scope of disinformational message content to any source of soft power. Broadly, these are: living up to political values, an attractive culture, and a morally legitimate foreign policy (Nye

[26] The concept of "soft power" is used to describe attractive power in international relations rather than domestic politics; however, the same dynamic of shaping others' interests through non-coercive power is present in domestic politics.

2011, 84). A state could thus use message or attributional disinformation to not just to assert the costs/benefits of one of its policies, but to assert its or their moral legitimacy, or to propagate positive messages about their society. This, as does any disinformation, risks negative reverberations. But, if a state is in such a position that it seeks to manufacture its own soft power, it likely already was in a neutral or net negative position with its target audience and may have little to lose.

A potential example of its use to defend soft power include Israel's 1954 denial that it was behind false flag attacks on U.S./U.K.-related locations in Egypt. Israeli national security officials, seeking to discourage U.K. withdrawal from the Suez Canal, planned a series of bombings on western facilities in Egypt. The goal was to create the impression that Egypt was insecure enough that a Suez Canal withdrawal was unwise. The plot failed, and Egypt asserted Israel's culpability. Israel publicly denied it, calling the charges against the plotters to be "false and slanderous" (*Coventry Evening Telegraph* 1955). In part, the denial could have been to protect Egyptian Jews and the Israeli government's domestic standing, but it also served to protect Israel's image internationally (Mearsheimer 2011, 39–40; L. Weiss 2013, 63).[27]

[27] There are further potential examples. Germany falsely claimed its 1939 invasion of Poland was defensive. Israel in 1948 claimed that Arabs voluntarily abandoned their homes, and in 1953 denied its responsibility for the Qibya massacre (Mearsheimer 2011, 65, 73–74). In 1964, the United States misrepresented the Gulf of Tonkin incident of 2-4 August as unprovoked attacks by North Vietnam. In reality, Secretary of Defense Robert McNamara assessed the 2 Aug attack likely was a defensive response to U.S-supported South Vietnamese attacks in the area, and both McNamara and President Johnson assessed the 4 August attack as likely to have not happened (Hallin 1989, 15–19; Hanyok 1998; Paterson 2008). More recently, China has used attributional disinformation to create the impression that that Xinjiang's Uyghers are happy (Krolik et al. 2021). These examples illustrate how disinformation can be dual-use—communicating foreign policy legitimacy to both domestic and foreign audiences.

The fourth method is electoral interference, particularly, supporting Level I and Level II candidates who are more likely than apparent alternative candidates to support one's preferred outcomes. This widens an opponent win-set through preference change, but not by persuading one's audience regarding a given policy or general policy orientation. Instead, it changes or maintains the preferences that constitute a state's win-set by changing or maintaining the preferers in key official positions. A state could use attributional or message disinformation to support a preferred official based on the official's known positions regarding a particular issue. Or, a state could support/oppose a particular candidate or party based on their general disposition toward that state. Thus, a state could employ this method to the end of achieving opponent concessions without a particular policy revision in mind, and without the officials in question having taken a clear stance on a particular future issue under negotiation.

Examples of this includes Russian influence efforts in the 2016 U.S. elections, and ongoing Chinese influence efforts on Taiwanese elections. In the former case, Russia used attributional disinformation to support one candidate over another. In the latter case, China used both message and attributional disinformation to undermine Taiwan's pro-independence party/party figures (DiResta et al. 2019; Harold, Beauchamp-Mustafaga, and Hornung 2021; Hung and Hung 2022). In both cases, the disinforming states targeted Level II audiences in order to shape the United States' and Taiwan's Level I win-set—not by persuading chief negotiators about specific or general foreign policy stances, but by changing the chief negotiators.

The fifth method to pursue opponent concessions is misrepresenting one's own win-set range. Specifically, the disinforming state will seek to persuade its opponent that

its win-set is narrower than it is—to bluff. Its target audience will likely be the opponent Level I chief negotiator, but could also include Level II players, depending on the issue under negotiation. Accusations of and concern about this sort of bluffing are more common than clear examples,[28] but one example is that of Germany during the first Morocco Crisis (1905-1906) (Mearsheimer 2011, 36).

Germany, concerned an emerging French-British partnership would shift the balance of power against it, encouraged the partnership's dissolution by threatening war against France over its increasing domination of Morocco.[29] This threat, sent by Kaiser Wilhelm and diplomatic channels in the Level I game, communicated to France that Germany's win-set regarding European balance-of-power did not encompass France's unilateral decision to assert greater control of Morocco. That is, that Germany's win-set was narrower than that of France. If France did not pursue an outcome that was within Germany's win-set, Germany would compel them to do so by war. This threat of war was meant to strain the U.K.-France relationship, as the U.K. would be drawn into a French-German war if it remained aligned with France. The threat though, proved a bluff. Germany's win-set could actually accommodate French domination of Morocco and a U.K.-France alliance, because history shows that it did. The result of the crisis was the

[28] Possible examples are both Iranian and U.S. bargaining positions in 2010-2020s nuclear deal negotiations and Germany's possible overstatement of its military preparedness during Munich negotiations (Caquet 2019; Never-ending Nuclear Talks 2022; Ruhe 2021). Time may help identify which assessed bluffs were actually credible signals (e.g., the U.K. actually did leave the EU, and Greece defaulted on its E.U. debts (*The Economist* 2015)). However, to establish the bluff, one must also rule out that the state/leader in question was in fact bluffing, but then Level I or Level II preferences changed (e.g., David Cameron and Brexit).

[29] Another possible example is China's threats to invade Taiwan in 1950 (Sartori 2002, 139).

United Kingdom backing France's position on Morocco, and Germany did not go to war against them (Zagare 2019; Mearsheimer 2011, 36; Kaufmann 1994).[30]

The sixth method operates by the same mechanisms to the same audiences: overstating one's uncertainty regarding opponent commitment to a potential agreement. As Putnam noted, such a strategy can increase the probability of opponent concessions, as opponents seek to supplement the devaluated agreement payoff. The type of defection the disinforming state expresses concern about is important in practice but tends toward the same result. Both voluntary and involuntary defection lower the payoff of an agreement and can be used as leverage to extract concessions. A potential example of this is the U.S. defection in 2018 from the Joint Comprehensive Plan of Action (JCPOA) with Iran. The president argued a new agreement was necessary because Iran had "lie[d]" about its pacific intentions in 2015 and JCPOA provisions were too weak to identify Iranian cheating (White House 2018). The International Atomic Energy Agency (IAEA) issued a statement, though, that there were no indications of Iran's bad-faith, and the U.S. Central Intelligence Agency director stated Iran had complied with the JCPOA as far as he knew (IAEA 2018; Nomination of Hon. Mike Pompeo 2018, 41). This delta between the official U.S. justification for defection and the IAEA and CIA director statements suggests the United States may overstated its concern of Iranian defection in part to extract further Iranian concessions: to move the status quo closer to the U.S. ideal point.

[30] To be sure, the balance-of-power concerns that drove the First (and later Second) Morocco Crisis were not resolved by either crisis, as they precipitated World War I. Thus, it could be said the Germany's win-set could not ultimately accommodate the U.K.-France relationship. However, that assumes that win-sets involved were the same in 1914 as in 1905-6.

Objective II—"get something", or increasing the chances of agreement ratification—is simultaneously pursued when a state uses any of Objective I's four methods to widen an opponent win-set. A further overlap with Objective I is that states can pursue increased ratification chances by misrepresenting their own win-set, albeit in the opposite direction. That is, a disinforming state seeks to persuade its opponent that its win-set is wider than it is. If the opponent state accepts that, the disinforming state can make a Level I agreement it has no intention of keeping. As with false side payment assurances, this intention to renege comes with high risk. This risk is not only of opponent defection, but more proximately of domestic ratification failure in the disinforming state, as discussed in the previous section.

Table 4. Disinformational Methods (Objs. II-III)

Objective	**Bargaining Spectrum**		**Method** **: unique methods [1, 2, 3, 4a,45b]: other obj. a method supports*	**Points of Manip.**				
	Element Modified	**Modification**		**Pref. Change, Mobiliz./Form.**	**Popularity**	**Politicization**	**Chief Negotiator Autonomy**	**Chief Negotiator Credibility**
II. GET SOMETHING *Increase chance of opponent ratification*	Win-set (*opponent*)	Widen/ shift towards own	False side payment assurances [1]	X				
			Pro-agreement/anti-status quo disinfo [1,3]	X				
			Promote own soft power [1,3]	X	X			
			Promote sympathetic/undermine antagonist Lvl I/II candidates [1,3,4a-b]	X				
	Win-set (*own*)	Widen	Misrepresent own bargaining range [1]	X				
III. TAKE SOMETHING *Increase chance of unilateral status quo revision*	Uncertainty (*opponent*)	Increase	Mask existence/extent of status quo revision	X				
	Win-set (*third-party*)	Widen/ shift towards own	Pro-revision/anti-status quo disinfo [1,2]	X				
			Promote own soft power [1,2]	X	X			
			Promote sympathetic/undermine antagonist Lvl I/II candidates [1,2,4a-b]	X				

The third objective (III)—"take something" or unilateral status quo revision—has much in common with previous objectives: namely, three of the four methods of win-set widening. The omitted method, false side payment assurances, are moot when one is

trying for unilateral revision. New to this objective is a method of increasing uncertainty about the revised, true status quo by masking it. One can mask the revision by either making it unclear or misrepresenting it, with the goal of delaying opponent response to the revision. Ideally, this delay allows for a *fait accompli*, but would be successful if it slows it enough that the benefit of the status quo revision outweighs the cost of belated contestation.

Two examples of this method's employment be found in the Soviet Union and its successor state. During the Cuban Missile Crisis, the Soviet Ambassador to the United States falsely told the U.S. government that its ongoing missile emplacement in Cuba was purely defensive, likely to buy the U.S.S.R. time to create entrenched facts on the ground (Frankel 2002; Mearsheimer 2011, 33).[31] In the model's terms, the U.S.S.R. used message disinformation at Level I to undermine U.S. chief negotiator certainty regarding the new status quo of offensive Soviet missile emplacements near the continental United States. Five decades later, Russia used status-quo obfuscation to support its seizure of Crimea. Early in the seizure, Russian forces in Crimea wore masks and uniforms stripped of official insignia, and the Russian government publicly denied they were Russian troops. Instead, they recast them as local militias that arose organically from Crimea's residents. This deception seemed intended to create ambiguity about the status quo revision among Level I opponent audiences (e.g., Ukraine, the E.U., NATO), delaying

[31] Mearsheimer's catalog of international deception includes further examples of status quo-masking (2011, 32, 34, 67). At the turn of the 20th century, Germany inaccurately framed its naval build-up as defensive, when it in fact intended to use the Navy to compete with the United Kingdom. In the interwar period, Germany denied that it was training military personnel in the USSR in contravention of the Treaty of Versailles. Finally, in mid-1945, the USSR lied to Japan that it would not invade its territory.

opponent responses long enough for Russia to seize key terrain on the peninsula (Galeotti 2015).[32]

The risk of revision-masking disinformation is encouragement of opponent state miscalculation. Increasing uncertainty may temporarily slow responses, but it also induce miscalculations that drive states to engage in cost-inefficient interstate conflicts (Fearon 1995). So, to increase opponent uncertainty is also to increase risk of opponent miscalculation. This risk does not preclude uncertainty-encouragement, but rather suggests that states will only employ it do so only after assessing the probability nature and cost of opponent miscalculations. One likely doesn't seize Crimea if one thinks a retaliatory nuclear strike is likely.

Turning to the status quo maintenance objectives (Objectives IVa,b, Table V), states can also use status quo-masking to maintain a status quo. Closely akin to masking a status-quo revision, masking the status-quo to maintain it differs in degree. Masking a revision permits the disinforming state seeks to protect their own gains; masking an unchanged status-quo permits the disinforming state to protect what they already have. An example of this is Russia's various inaccurate explanations of its role in the 2014 Malaysian Airlines shootdown over Ukraine. Semi-official news outlets asserted variously asserted that a Ukrainian pilot had made a mistake, that Ukraine staged the shoot-down to make Russia look bad, and that MH17 had been shot down with a surface-

[32] In the same decade, Israel falsely denied to the United States that it had a nuclear weapons program, avoiding what certainly would have been U.S. opposition (Mearsheimer 2011, 33). Another potential example is that of the United States after the Cold War, falsely denying to its allies that its removal of missiles from Turkey was a result of a crisis-resolving deal with the Soviets (Mearsheimer 2011, 67).

to-air missile intended for the Russian presidential jet. These outlets reached domestic audiences in Russia, but also Level I and Level II audiences abroad through foreign-language outlets like RT (Toal and O'Loughlin 2018; Elswah and Howard 2020, 36; Demirjian 2023; Ivshina 2015). These multiple, inaccurate explanations may have served to not only defend its soft power, but to increase uncertainty about the status quo and reduce international pressure for changes to it.[33]

Table 5. Disinformational Methods (Objs. IVa-IVb)

Objective	**Bargaining Spectrum**		**Method** **: unique methods* [1, 2, 3, 4a,45b]*: other obj. a method supports*	**Points of Manip.**				
	Element Modified	**Modification**		**Pref. Change, Mobiliz./Form.**	**Popularity**	**Politicization**	**Chief Negotiator Autonomy**	**Chief Negotiator Credibility**
IVa. KEEP WHAT YOU HAVE *Decrease chances of status quo revision attempts*	Uncertainty (*opponent*)	Increase	Mask status quo	X				
	Win-set (*opponent*)	Narrow	Anti-agreement/pro-status quo disinfo [4b]	X				
			Promote sympathetic/undermine antagonist Lvl I/II candidates [1,2,3,4b]	X				
IVb. KEEP THE OTHER GUY FROM GETTING SOMETHING *Decrease chances of opponent cooperation*	Win-set (*opponent*)	Narrow	Anti-agreement/pro-status quo disinfo [4a]	X				
			Promote sympathetic/undermine antagonist Lvl I/II candidates [1,2,3,4a]	X				
			Undercut opponent soft power *	X	X			
	Uncertainty (*opponent*)	Increase	Disinfo. regarding ratification challenges (increase invol. defection concerns) *	X	X	X		
			Encourage polarization to increase opponent chief negotiator autonomy (vol./invol. defection concerns) *	X	X	X	X	X
			Disinfo. regarding chief negotiator reliability (increase vol. defection concerns) *	X				X

In addition to increasing uncertainty about the true status quo, states also can seek status quo maintenance by anti-revision/pro-status quo disinformation. Unlike increasing

[33] Another potential example of this is the USSR's denial of its continued biological weapons capability and research in contravention of treaty obligations in the 1970s-1980s (Mearsheimer 2011, 33–34). On one hand, it was status quo maintenance, as it had the program and wanted to continue having the program. On the other hand, it also included some status quo revision, as the program did not continue as a steady-state but made revisions to the status quo.

uncertainty, this method seeks to persuade opponents that status quo revision is too costly rather than unnecessary, and operates much the same as oppositely-valenced disinformation to support Objectives I-II. The state still relies on message and/or attributional disinformation to persuade Level I players and effect persuasion/responsiveness at Level II, this time highlighting the low/improbable benefits of a revision and/or the low cost of the status quo. It could do so by promoting/attacking certain general frames (e.g., promote isolationism, criticize internationalism) and by focusing on a particular issue, but the goal is to narrow opponent state win-sets. This reduces the chances of bargaining success and moreover, can forestall negotiations at all. Nazi Germany used this method, among others, to support its rearmament program in the 1930s. Initially, it concealed its rearmament program to deny France and the United Kingdom a true picture of the status quo. For example, it built up its nascent air force under the guise of a civil air service. After building up more of its capabilities, however, Germany's tactics changed. While continuing to underplay a few aspects of its rearmament, it inflated its military strength to British and French Level I interlocuters. Raising the perceived costs of status quo revision, this message disinformation encouraged acquiescence to the status quo of a rearmed Germany (Mihalka 1980; Mearsheimer 2011, 31).[34]

[34] A similar example is the USSR's inflation of its military capabilities in the 1930s to discourage a German attack after Stalin's purges hollowed the Soviet Army (Mearsheimer 2011, 31). More recently, Russia generated attributional and message disinformation in 2014-16 with the apparent intent of spoiling Swedish cooperation with NATO and Ukraine (Kragh and Åsberg 2017, 20). The disinformation included claims like NATO opposition to the United Nations and an OSCE coverup of Ukrainian corruption.

A state employing this method can but need not be party to the status quo in question (Objective IVb). The method of electoral interference can also be used regardless of party status. The remaining four methods for status quo maintenance, however, apply only when the disinforming state is not party to the status quo. First, a disinforming state can seek to undercut the soft power of at least one of the opponent parties. Such a method would be counterproductive to Objective IVa, as it would require undercutting one's own soft power, but is not so when the target is an opponent state. Using disinformation to attack a state's culture, adherence to its political values, and foreign policy legitimacy can reduce that state's attractive power and thereby weaken its influence over other states' win-sets and ideal points. This narrows their prospective partners' win-sets (or at least, shifts them farther away) and reduces the chance of agreement and ratification.

An example of this is USSR attributional/message disinformation regarding the origin of the AIDS virus. In the 1980s, with the AIDS crisis in its early years, the Soviets generated and circulated fabricated news alleging that the United States had created the virus at Fort Detrick, Maryland. The claim, initially planted in a covert Soviet newspaper in India, was intended, among other things, to encourage anti-American attitudes in audience countries (*Patriot Magazine* 1983; Panayotov and Nikolov 1985). In the model's terms, the goal was to decrease the soft power of the United States by alleging its hypocrisy in developing offensive bioweapons contrary to international norms and asserting its moral culpability for unleashing deadly virus on the world.[35] The claims,

[35] Further examples include Russia's various attributional disinformation efforts to increase political division in the United States; and a 2020 Iranian operation that sent threatening emails to U.S.

raised repeatedly by witting an unwitting Soviet agents throughout the 1980s, were circulated in many countries, and concerned the U.S. State Department enough that it risked amplifying the false claims further by publicly denying them (Selvage 2019; Geissler and Sprinkle 2013; U.S. Department of State 1987).

Second through fourth, a state can seek to spoil other states' cooperation by manipulating their uncertainty regarding defection chances. This has the effect of encouraging a commitment problem. Most simply, a state could use message and attributional disinformation to persuade one state's players that the other state's ratification chances are low. The particulars will depend upon the state's institutional features but could include amplifying concerns of public or elite opposition, gridlock due to politicization, or chief negotiator unpopularity. If effective, such disinformation increases concern of involuntary defection.

A state also could seek to raise concerns a state will defect by encouraging political polarization or amplifying its appearance in that state. Polarization, a condition of consistent politicization, may seem more useful for undercutting a state's soft power than encouraging a commitment problem. A chief negotiator in conditions of polarization has high chance of support from their supporting coalition (in part due to the entanglement of affective and ideological polarization (Hetherington 2015; Hetherington and Husser 2012; Hetherington and Rudolph 2015; Rogowski and Sutherland 2016)), and a correspondingly low a chance of support from opponents. This means they can strongly lead their coalition's preferences and effectively ignore their opponents, with a result of

leftwing voters while posing as members of an extremist-associated rightwing political organization (DiResta et al. 2019; Kramer 2020; U.S. Department of Justice 2021).

reduced domestic constraint. In the logic of the model, reduced domestic constraint means a wider win-set and therefore greater chance of bargaining success.

However, the model suggests greater autonomy also comes with increased defection risk. If a chief negotiator's actions are only weakly constrained by their Level II players, the chief negotiator can "get away with" defection from agreements.[36] This countervailing dynamic suggests two separate conditions under which a state could use disinformation to increase opponent chief negotiator autonomy via polarization. First is when the target chief negotiator has low personal credibility. In the absence of a strong responsiveness dynamic, apparent defection risk largely hangs on the credibility of the singular chief negotiator. Second is when the target state's political system is characterized by highly contested turnovers in power. In such a circumstance, even the word of a highly credible chief negotiator could be broken when their opponents eventually take the reins of government.

The first of these conditions indicates a final way a state can encourage a commitment problem to spoil another state's negotiations. It can use disinformation to undermine confidence in a chief negotiator. This could be used to support the previous method of polarization encouragement, as that method works well when the target chief negotiator has low credibility. For example, a state could use disinformation to amplify polarization (or at least polarized voices) in one target state, while using disinformation to undermine the credibility of that state's chief negotiator in another. It also could use negotiator-undermining disinformation by itself, as low chief negotiator credibility is

[36] This conceptually includes domestic audience costs, but is wider, encompassing any retrospective accountability mechanism.

sufficient to increase concerns of voluntary defection. As with nearly all the previous methods, such disinformation operates via mechanisms of persuasion and/or persuasion and responsiveness, with potential audience targets at both Level I and II.

A general example of encouraging defection concerns is contemporary Chinese disinformational efforts in Taiwan. These efforts comment on many subjects, among them a broad narrative translated as "U.S. skepticism theory." This narrative includes many elements (e.g., the U.S. is using Taiwan to develop biological weapons), to include assertions that the United States cannot be depended upon to defend Taiwan (China is Flooding Taiwan with Disinformation 2023; Gordon, Mullen, and Sacks 2023, 24; Wu 2023). The disinformation's argumentation is currently opaque in English-language publications, so I cannot assess whether it incorporates the hypothesized sub-arguments regarding polarization-promotion and attacking the reliability of U.S. chief negotiators (e.g., President). However, the general message of partner unreliability aligns with all three methods.

Implications & Limitations

What do we gain by examining disinformation in foreign policy from a two-level perspective? Beyond accounting for known datapoints, it generates new understanding and implications.

Known, noteworthy datapoints to which the telling conforms are threefold. First and second, the model expects that states not only can but should use disinformation to influence opponents' Level I and Level II players, and that they can use attributional disinformation to sidestep credibility issues. This is consistent with descriptive works that capture instances of disinformation use at Level I and in an opponent's Level II game, to

include attributional disinformation (Bittman 1981; DiResta et al. 2019; Martin 1982; Samoilenko and Karnysheva 2020). Third, states should accordingly seek to block access to their Level II game. This, too, is supported by a growing body of literature that captures state efforts to limit unwanted informational influences among domestic audiences (Brown and Peters 2018; Dukalskis and Gerschewski 2017; Frantz, Kendall-Taylor, and Wright 2020; Gunitsky 2015; Han 2015; Keremoğlu and Weidmann 2020; King, Pan, and Roberts 2014; Nocetti 2015; Rød and Weidmann 2015).[37]

In terms of new understanding and implications, a two-level bargaining focus generates specific objectives, associated strategies toward win-sets and uncertainty, and resultant disinformational methods. Previous theoretical and inductive perspectives identified these only in part. In terms of objectives, a two-level approach agrees with Lanoszka's two goals of armament policy and alignment change and adds more. Any issue area can be contested, and beyond changing the status quo, states can also seek to forestall change. Similarly, the two-level perspective shares Walker and Ludwig's perspective that states use disinformation to manipulate soft power, but do not understand that as disinformation's sole objective. Instead, the two-level perspective includes soft power manipulation as but one of three disinformational methods to manipulate opponent win-set size, and those manipulations are not ends in themselves, but support negotiation success and cooperation-spoiling. Further, the two-level perspective highlights entire

[37] Such efforts are not conditioned on regime type: democracies also work to prune or identify foreign influence efforts. It is only in their tools that they differ (e.g., social media companies removing disinformation from their platforms to reduce the risk of binding legislation) (Brown and Peters 2018)

areas of disinformational efforts vis-à-vis one's own win-set and uncertainty in which soft power plays no role.

Importantly, the two-level approach not only encapsulates previously-identified objectives; it also identifies a novel objective for disinformation use: encouraging a commitment problem between opponent states. The idea of cooperation-spoiling via polarization accords with a recent analytical essay on polarization's impact on American foreign policy (Friedrichs 2022a), and this paper further identifies two additional points of manipulation (chief negotiator credibility, ratification probability) that a disinforming state can harness to the same end. This possibility of commitment-problem encouragement explains contemporary disinformational efforts that Lanoszka dismissed as resulting from miscalculation or posturing for domestic audiences.

Regarding win-set and uncertainty strategies, the two-level game generates a list of modifications a disinforming state may pursue in pursuit of its goals. The exhaustiveness of the list enables confident assessment of what modifications a state should or should not pursue via disinformation. A state seeking to increase probability of agreement can widen an opponent win-set. It should not seek to narrow the opponent win-set; nor should it seek to falsely assert its own win-set narrowness. A state seeking to spoil other states' cooperation can do so through manipulating their win-sets and uncertainty—narrowing the former and increasing the latter. It cannot do so through any other means.

With regards to methods, the interrelated and comprehensive (though not exhaustive) list of disinformational methods generated by this model does not exist in other work. As with Walker and Ludwig's soft power observation, other scholarship has

identified some of the methods individually or in small groups (Beskow and Carley 2020; DiResta et al. 2019; Harold, Beauchamp-Mustafaga, and Hornung 2021; Kragh and Åsberg 2017; Walker and Ludwig 2017a), but never laid them out entire and did not identify some at all. Wholly novel are the methods of undermining chief negotiator credibility and confidence in ratification chances. Contextually novel are some methods that have not been put under the umbrella of disinformation. Putnam points out state-propagated inaccuracies like misrepresenting one's bargaining range or uncertainty regarding defection, and these examples of disinformation are absent from contemporary scholarship on disinformation, with its digital focus.

The model's more expansive accounting of objectives, bargaining element modifications, and methods provides nuance to discussions of disinformation in foreign policy. Yes, a state may use disinformation to encourage policy change. But is the state's goal to achieve an acceptable (Obj. I) or concession-maximizing policy (Obj. II)? A state can pursue both objectives simultaneously by win-set widening methods, but not so with respect to its own win-set. A disinforming state that chooses to falsely widen rather than falsely narrow its own win-set is choosing agreement chances over concession chances. Similarly, feigning concern of opponent defection can help extract concessions, but does not simultaneously increase probability of agreement.

Further nuances can be found in the other three objectives and their supporting strategies and methods. Increasing opponent uncertainty regarding the status quo supports unilateral status quo revision and maintenance (Objectives III, IVa), but does not assist with the other objectives. Similarly, many of the methods to pursue Objective IVb are unique to it. Undercutting opponent soft power only assists in undermining opponent

cooperation, as do increasing defection concerns via polarization, undermining a chief negotiator's reliability, and questioning an agreement's ratification chances. Opposingly, though, the model suggests polarization encouragement can have multiple impacts. It is explicitly tied to opponent cooperation-spoiling—whether undercutting soft power or increasing defection concern—but in practice is likely indistinguishable from election interference in support of the other objectives. This suggests that promotion of political division is a general-purpose disinformational method. It can allow disinforming states to simultaneously encourage a commitment problem and preferred election outcomes, and even if neither is achieved, it may still undercut the target state's soft power.

In addition to these more granular points of understanding, the model also suggests why states disinform in general: inaccuracy sometimes seems more likely to pay. Put in the model's terms, a chief negotiator chooses to disinform vice inform if disinforming seems likely to have higher payoff (the payoff being defined by their objective—get more, get something, take something, or status quo maintenance). But would disinforming ever seem higher-payoff than informing? After all, disinformation in the model always risks negative reverberations. Perhaps the answer is a conditional yes: disinformation could sometimes seem higher payoff, but only due to chief negotiator miscalculation or high risk tolerance? No, in broader circumstances than that. Disinformation may seem higher-payoff if informing would be counterproductive, if disinformation is unlikely to be unmasked, and/or if the cost of negative reverberations is negligible.

The first circumstance arises from the reality that informing, like disinforming, can be costly. For example, accurate communication about one's wide bargaining latitude

increases opponents' bargaining leverage, and a foolproof way to decrease one's bargaining leverage on an opponent is to reveal that one does not think the opponent will defect. Similarly, when a state's goal is unilateral status quo revision, informing opponents about the pending revision likely would be anything but helpful. In sum, informing is not necessarily a default "safe" option for a payoff-optimizing chief negotiator.

The second circumstance—disinformation being unlikely to be unmasked—means the risk of negative reverberation is low, and thus probable cost is lower. This circumstance occurs organically for some communication topics. Win-set width is one. It can be difficult for a chief negotiator to accurately identify their own win-set bounds, much less for foreign parties to identify them (and therefore for foreign parties to identify when opponents are misrepresenting win-sets). A similarly hard-to-access topic for opponents is one's uncertainty regarding their defection chances. Beyond taking refuge in naturally lower-risk topics, states also can reduce the risk of disinformation identification through technical competence. Excellent attributional disinformational cover and disciplined counter-intelligence permit states to blunt opponent efforts to assemble an accurate picture of reality.

The third circumstance is if the costs of negative reverberations are negligible. This occurs when the relationship between the disinforming state and target audiences are so frayed that a revelation of disinformation could hardly make things worse. This also occurs when the relationship between the disinforming state and non-targeted third-party audiences is such that the disinforming state is effectively immune to negative reverberations. This occurs when unfriendly third-party audiences' low opinion could

hardly be lowered more and friendly audiences are likely to gainsay the target's allegations of disinformation. Finally, this also occurs when a disinforming state judges that failure of a desired status quo maintenance/revision is more costly than potential negative reverberations. Examples of this include high stakes security-related situations, such as when the United Kingdom fabricated intelligence to encourage U.S. participation in World War II.

As with all models, the two-level game has limitations. The first is how the list of methods it derives are not comprehensive. This is not a model that generates a neat matrix of all possible circumstances at lower levels of abstraction. The second is how the model does not account for some disinformational methods that have intuitive validity. One could easily think of a scenario in which promoting one's own soft power would increase the chance of status quo maintenance (Obj. IVa); however, soft-power promotion results in opponent win-set widening (or at least shifting), which is contrary to the model's clear implication that opponent win-set narrowing will encourage status quo maintenance (i.e., low status quo costs = narrow win-set). The third is its unsuitability to some questions raised by previous literature. For example, whether states use disinformation abroad to posture for an audience at home is beyond the scope of the model, as is the question of under what conditions a state will pursue each disinformational method. But, within its scope are the major open questions of why states use disinformation, through what mechanisms, and to what effect.

Conclusion

With an eye towards future research, three sorts of implications from the model seem most promising. The first is that disinformation can impact an opposing state's foreign

policy preferences through many pathways. Within the two general mechanisms of persuasion and responsiveness, disinformation can have overlapping influence on preference change; player mobilization/preference formation; polarization; soft power; chief negotiator credibility; involuntary defection concerns; and decision-making speed. This means that scholarship examining disinformation's impacts should explore the impacts of disinformation on belief ("I believe X piece of information"); but, for a comprehensive telling, it should also consider the other pathways to win-set content and uncertainty change. This in turn means examining not only the effects on the target audience, but potential reverberations to secondary audiences, such as increased defection concerns.

While this multiple pathway point may seem obvious, research on foreign disinformation campaigns tends to focus on belief outcomes regarding a particular policy or candidate at the expense of the other potential effects. The impacts of disinformation on belief are important and remains an open question due to measurement challenges.[38] The other impacts are nearly wholly unexplored. Some inductive work describes Russian disinformation apparently intended to sow division and interfere in elections, inherently to include impacts on preference change and mobilization/formation (Al-Rawi and Rahman 2020; Dawson and Innes 2019; DiResta et al. 2019; Etudo, Yoon, and Yaraghi 2019); but its actual causal effects remain open on those counts and with regard to

[38] The paucity is in part due to confounding variables (the people most susceptible to political disinformation are the few most predisposed to believe it); however, some of the sparsity is due to simple measurement challenges. It is challenging to conclusively identify and then measure the real-world impacts of disinformation, as Level I players aren't likely to reveal how it impacted them, while Level II players are enmeshed in a such a complex communication environment that it is challenging to make causal connections (assuming one can even identify and attribute the disinformation to begin with, and then that the disinformation you found is representative of the disinformation produced).

potential impacts on soft power, chief negotiator credibility, and involuntary defection concerns. For example, can disinformation encourage ideological polarization, and in turn, a commitment problem? Survey experiments, quantitative text analysis of news/social media content may be of use to answering that question.

A related question is whether disinformation encourages affective polarization and mistrust. The two factors are not only interrelated with ideological polarization (Hetherington 2015; Hetherington and Husser 2012; Hetherington and Rudolph 2015), but soft power as well. "Pronounced social cohesion"[39] is one of the sources of soft power (Gallarotti 2011, 23), and while Nye dismissed concern that contemporary U.S. social cleavages undermined U.S. soft power, it was because he considered the cleavages exaggerated rather than theoretically unimportant (2003, 114).

A second sort of implication has to do with the temporal elements of disinformational efforts. The model suggests that disinformation can be used to play a long-game, to generally tilt the game board in the disinforming state's favor. Promoting one's soft power can support the realization of uncrystallized and currently unknown future preferences. Likewise, polarization can decrease opponents' soft power and increase risk of defection across the board. A state could thus use those methods to their long-term benefit without having a particular apparent near-term bargain it is trying to spoil. Of empirical interest is the question of how effective this long-game is, with the attendant methodological challenge of how one could isolate such influence, given the

[39] From a different angle but tending toward the same point, Vuving (with the input of Joseph Nye) wrote in his theory of soft power that two of soft power's three bases are "positive attitudes toward other people, qualities such as kindness, benevolence, compassion, and generosity" and "commitment to values, identities, beliefs, and aspirations" (2019, 23). Affective polarization and mistrust sap both.

complexity of the information environment. Also of interest is whether there is a shift in disinformational themes from long-game, general cooperation-spoiling (promote division on all sides) to clear support for particular candidates during election cycles (promoting one side at the expense of other[s]).

Also in the temporal realm is the model's suggestion of disinformational synchronization with other, non-disinformational elements of power. The objectives and attendant win-set/uncertainty strategies are not specific to disinformation and can be pursued through other elements of power. Thus, the model anticipates those other elements of power will be used either in tandem with or at least not in opposition to the disinformational campaign. This means that examining disinformational campaigns may help point toward how a state is employing other tools of power, and vice-versa. Finally in the realm of synchronization, the model suggests that states seeking to increase polarization in one state may also use disinformation to undermine the credibility of that state's chief negotiator in the eyes of foreign audiences. Researchers that identify polarization-promoting disinformation should thus also check if there is an accompanying disinformation campaign in another state.

Overall, a two-level game understanding of disinformation is a fruitful one. It gives a more comprehensive accounting of disinformation than previous analyses and generates some novel or only partially-tested implications. Moreover, it suggests that contemporary use of disinformation in foreign policy is not paradigm-breaking. Global communications' reach is new, and they make disinforming foreign Level II audiences more feasible. But, the rationalist two-level game introduced can not only account for contemporary disinformational practices like election interference and polarization

promotion; it also relates them to specific strategies and objectives that states can and do pursue with other tools of power. Disinformation's reach in foreign policy is novel, but its place is not.

CHAPTER 2

TESTING TWO IMPLICATIONS: PAPERS 2 & 3

The dissertation's second and third papers will test two implications of "Disinformation in a Two-Level Game." The second paper focuses on the multiple pathways to disinformational impacts, to include belief, but also including underexplored effects on affective polarization, trust, and mobilization (particularly focusing on two emotional antecedents of mobilization: anger and fear). As its argument applies not only to disinformation but to the broader term of misinformation as well, it uses the term misinformation instead of disinformation.

The third paper shifts focus to impacts in third-party states, testing the implication that states could use disinformation to encourage a commitment problem. Particularly, it tests the link between polarization in one state and cooperation hesitation in another. It thus builds from the second paper's test of a disinformation-affective polarization link.. It argues that publics and elites form their cooperation preferences regarding a given state in part from that state's reputation for reliability, and that polarization can deleteriously impact that reputation.

CHAPTER 3

POTENTIAL UNDERESTIMATIONS OF MISINFORMATION'S IMPACTS: MISTRUST, EMOTION, AND A SELECTION BIAS

Introduction

What are the effects of exposure to political misinformation? The question is of rising interest not because misinformation is novel, but because the rapid, mass propagation of content via social media is. Actors can disseminate misinformation without easy attribution, and its spread is difficult to curb, given potential circulation speed and audience size (Vosoughi, Roy, and Aral 2018). wide-ranging a communications revolution c and the responsiveness of government to its public, much research has focused on political misinformation's impact on misperception belief and vote choice.

Alongside misinformation's rising salience are sets of research challenges. Some are ethical. How much misinformation of what kind can participants view before being harmed? (Flynn, Nyhan, and Reifler 2017; Guess et al. 2020). Others are empirical. Much research has focused on active exposure to or engagement with inaccurate political content, measured in "clicks", comments, shares, posts, and "likes." Fewer studies have assessed the impacts of passive exposure to inaccurate political content, which is challenging to measure due to its lack of a clear data trail (Allen et al. 2020; Guess, Nyhan, and Reifler 2020; Nelson and Taneja 2018). Beyond those research challenges, misinformation scholarship also includes a focus on misinformation's impact on belief and vote choice—which are highly theoretically-relevant but also limit the scope.

Arguing that misinformation's impact could be underestimated, this study uses a survey experiment to test the impacts of passive exposure to partisan misinformation. The

study also tests whether there is a result-deflating selection bias among people that opt-in to the data tracking required for active exposure studies. Misinformation's potential impacts include not only belief, but ripple effects to emotions and trust even among those that disbelieve the misinformation. Specifically, I hypothesize that passive exposure to partisan misinformation weakens some factors related to societal cohesion like generalized trust (GT) and outgroup party trust (OPT) and strengthens the inherently division-related factor of affective polarization. I also hypothesize it provokes political participation-driving anger and fear. I expect these impacts regardless of whether respondents are exposed to politically congenial or non-congenial misinformation (CM, NCM) due to social identity theory and intergroup contact theory.

Partisan CM exposure should decrease trust and increase anger/fear because partisans are more likely to believe the misinformation's political claims and negatively update related assessments of the outgroup party and the broader state of politics. Those exposed to partisan NCM could display the same outcomes, but for a different reason: because they disbelieve the misinformation's claims. That is, NCM encounters are an instance what I term negative informational outgroup contact (NIOC), negatively-valanced encounters with an outgroup. Such encounters increase negative outgroup perceptions, a dynamic reinforced by the third-person effect (TPE), or tendency to believe media have stronger influence on others than oneself, particularly when the message in question is undesired and the "other" is an outgroup.

Ultimately, the study finds that passive exposure's most clear effect is on misinformation belief. The study also finds that passive misinformation exposure likely can impact trust and emotions. However, these ripple effects are unexpectedly

conditioned on the combination of partisanship, misinformation congeniality, and (dis)belief. They also are tentative, as they derive from low-*n* analyses with probable floor, often only approaching standard levels of significance. Taken with that important caveat, passive CM exposure decreased generalized trust among Republicans that believed it. Passive NCM exposure, on the other hand, impacted only Democrats who disbelieved it, decreasing their generalized trust and increasing their anger and fear. Counter-expectation, CM/NCM exposure had no effect on outgroup party trust and affective polarization.

Other than floor effects, the wholly null results for those two variables and unexpected conditionality for the others could be due to asymmetrical treatment weakness or participants unexpectedly not engaging in group reasoning when processing the misinformation. I cannot rule out the former, but the latter is improbable, as study participants displayed group reasoning by updating ingroup/outgroup TPE assessments after being exposed to partisan misinformation. Finally, the study also finds no theoretically-relevant differences between those that opt-in/out to data tracking. Misinformation's impact may be underestimated in terms of ripple effects, but it is not underestimated because of selection bias in active exposure studies.

Misinformation: Scope and "Ripple Effects"

Misinformation: Is Its Influence "Fake News"?

One could assemble a hopeful case that political misinformation and its impacts are overstudied (and for ease of reading, from this point I will refer to misinformation instead of political misinformation). In terms of observational data, proportionately few social media users actively visit "fake news" sites or share misinformation (Allcott and

Gentzkow 2017; Allen et al. 2020; Altay, de Araujo, and Mercier 2021; Grinberg et al. 2019; Guess, Nyhan, and Reifler 2020; Guess, Nagler, and Tucker 2019; Nelson and Taneja 2018; Osmundsen et al. 2021). The ones that do tend to be more highly partisan and polarized. When combined with studies that find both active (information-seeking) and passive exposure to misinformation only minorly increase belief among those predisposed to believe it,[40] this could suggest that popular and academic apprehension of "fake news" is fake news.

However, this account could fall short in three respects. First, studies that measure active exposure to misinformation could suffer selection bias. Such studies examine respondents' social media or internet browsing history data, appropriately dependent on respondent consent (Osmundsen et al. 2021). Given misinformation as the subject under study, respondents who opt-in to an intrusive study could differ from those who don't in result-understating ways. Opt-in respondents are generally representative in terms of gender, race, age, education, and party, or weighted to be so (Bhadani et al. 2022; Guess et al. 2020), but some studies have found that opt-in respondents tend to be more politically involved (political knowledge, having voted), particularly in left-leaning ways (voting for Democratic candidate) (Guess, Nyhan, and Reifler 2020; Guess, Nagler, and Tucker 2019). No study has considered whether opting-in is correlated with factors such

[40] Nyhan and Reifler found conservatism increased respondent's average level of acceptance of a two politically congenial perceptions on a four-point scale by ~8-9% and ~4%, respectively (2010, 314, 317, 320). Loomba et al. found exposure to COVID vaccine misinformation caused a 6.2-6.5% decline in definite vaccination intent (2021, 340). On the other hand, Nyhan et al. found that Trump supporters in a larger experiment expressed the same mean level of belief in a misperception regardless of whether they were exposed to misinformation supporting the perception (2020, 948). Finally, Guess et al. found that exposure to a single fake news story increased average belief in two pieces of pro-Democrat/-Republican misinformation among partisans of both stripes 5.4-15% on a four-point scale (2020, 8).

as partisanship strength/polarization, or trust in the media, government, or fact checkers (Altay, de Araujo, and Mercier 2021; Garrett et al. 2014; Sanchez and Dunning 2021; Tsfati and Cappella 2003). Such a correlation could impact results concerning amount of misinformation actively shared or visited, misinformation belief, and prebunking/debunking effectiveness.

The second way the hopeful telling may fall short is its overlooking the challenge of capturing the potential impacts of passive exposure to misinformation online (Allen et al. 2020; Pasquetto, Swire-Thompson, and Amazeen 2020; Pennycook and Rand 2021a; Rogers 2020). Passive, also called incidental exposure, occurs when a person encounters a piece of information they did not seek out (Stroud, Scacco, and Kim 2022). For example, this occurs when one opens a browser homepage and encounters its rotating news stories, and when social media users scroll through their feeds. The line between active and passive exposure conceptually blurs when one considers that people can tailor their media consumption to create echo chambers, so that even passive exposure on social media can involve an element of user selection. Further, passive exposure can happen in the context of browsing, that is, general information-seeking (Tewksbury, Weaver, and Maddex 2001, 534). But, in terms of operationalization, the active-passive distinction matters because a given piece of information is likely encountered by far more people than those that intentionally sought it out.[41]

[41] A think tank report found that only 40% of Americans reported having engaged with news stories deeper than the headlines over the previous week (Media Insight Project 2014). This finding is consistent with a study that found 51-70% of news articles shared on Twitter from five major news domains redirect no traffic from Twitter to those domains (i.e., no one clicked on them), and with a big-data analysis that found the average click-through rate on recommended news articles from Microsoft's News homepage was less than 10%, regardless of subject (Nayak, Garg, and Duvvuru Muni 2023, 1399)

As passive exposure lacks an easily-accessible data trail, this presents inherent challenges to misinformation research. Passive exposure literature tends to rely on self-assessments of passive exposure—measures whose unreliability due to inaccurate respondent memories and social desirability bias is likely even more pronounced for questions of misinformation exposure (Xiao, Su, and Lee 2021). One potential source for more reliable-data would be Facebook, which captures passive exposure data using an algorithm that analyzes scrolling behavior, but that data is publicly unavailable (Facebook, n.d.; Yu and Tas 2015). As a result, misinformation researchers largely rely on experiments, which come with their own challenges.

Experimental settings typically present one-two exposures of already-circulating misinformation in a survey format or in format that partially approximates a Facebook recommended stories feed. Neither context closely resembles the complex environment in which people encounter new information (Guess et al. 2020; Luo, Hancock, and Markowitz 2020). Participants may have encountered the treatment misinformation prior to the experiment, and the treatment delivery may make it easy for some participants to guess the treatment's rough purpose. This potential result-deflation due to the social desirability of not being fooled by misinformation, in combination with pre-treatment effects, makes it somewhat remarkable that single-exposure treatments of already-circulating misinformation find an average effect at all.

The final potential shortcoming of the hopeful account is its focus on belief and voting outcomes (e.g. Who believes/disbelieves? Does belief change vote choice? Are fact checking interventions effective in reducing impacts on belief or vote choice?). Those outcomes are highly theoretically appropriate, as responsiveness literature is

premised on public beliefs impacting government policies via participation. As I argue in the next section, however, they are not the only theoretical outcomes of misinformation exposure, especially when one considers that misinformation can impact those that disbelieve it.

Belief, Ripple Effects, and How Misinformation Impacts Even Those that Disbelieve It

Four Effects of Political Misinformation

A foremost theoretical impact is on belief. Passive exposure to information via traditional and new media is associated with learning (Cairns, Hunter, and Herring 1980; Janiszewski 1993). This has been established with respect to political knowledge, but more generally in advertising and marketing (Baum 2002; Nanz and Matthes 2022; Takano et al. 2021; Weeks, Lane, and Hahn 2022; Yoo 2009). The operation of directionally-motivated and accuracy-motivated reasoning, while conceptually distinct, together point in the same direction: that belief is most likely among those to whom it is most concordant, or congenial. Whether for identity-protection—as proposed by social identity theory—or due to differing assessments of messenger trustworthiness, people are more likely to accept new information that most accords with their priors (Druckman and McGrath 2019; Huddy and Bankert 2017; Jefferson, Neuner, and Pasek 2021; Kunda 1990; Lee et al. 2022, 2022; Taber and Lodge 2006).

Regardless of preexistent beliefs, people of all partisan stripes have increased susceptibility to misinformation if exposed to it multiple times. This is due to the illusory truth effect, or people's general tendency to believe something the more they encounter it. Attested to in cognitive psychology (Begg, Anas, and Farinacci 1992; Hasher, Goldstein, and Toppino 1977; Wang et al. 2016; Whittlesea 1993); marketing (Hawkins

and Hoch 1992; Roggeveen and Johar 2002); and two recent studies about misinformation (Fazio, Rand, and Pennycook 2019; Pennycook, Cannon, and Rand 2018), the illusory truth effect particularly matters for those in social/digital networks that are politically engaged. The increased probability of repeated exposure to political misinformation increases the probability of illusory magnification of the misinformation's accuracy.

Beyond belief, passive exposure to misinformation can have interrelated sociopolitical "ripple" effects, of which I will focus on three: fear/anger, affective polarization, and trust. Importantly, these theoretical ripple effects are not limited to those that believe misinformation, as disbelief can lead to updated appraisals of associated attitude objects, e.g., interpersonal deception results in negative appraisals of the deceiver (Tyler, Feldman, and Reichert 2006). Further, social identity theory and intergroup contact theory suggest that people learn not only about the singular communicator. They also learn about groups associated with the communicator, and in circumstances of intergroup competition, the learning is biased positive toward ingroup(s) and negative toward outgroup(s) (Abrams and Hogg 1990; Brown and Ross 1982; Tajfel et al. 1979; Turner, Brown, and Tajfel 1979). These are precisely the circumstances of political misinformation in the United States, with high competition between distinct political parties (Greene 2004; Huddy and Bankert 2017).

First, regarding emotion: cognitive appraisal theory suggests that certain emotions will follow certain cognitive appraisals of new information, and the theory does not condition the emotion-appraisal patterns on information accuracy or the passivity of exposure (Folkman et al. 1986; Lerner and Keltner 2000; Roseman 1996). Now, there is

nothing to suggest anything unique about the range of emotions believed misinformation can provoke. Sadness, enthusiasm, anger, and fear can result from true communications as well as false ones. However, disbelieved misinformation likely majors in negative emotions. Particularly in the context of contemporary U.S. polarization, disbelief in political misinformation likely leads to inferences about its sharer and their associated groups. These appraisals will be biased toward one's ingroups and against one's outgroups, meaning: disbelieved outgroup misinformation is more likely to result in negative emotions, like anger and fear. The result is that a piece of misinformation that angers or inspires fear in those that believe it may inspire the same emotions in those that disbelieve it.

That two-fold emotional impact has political implications.[42] Strongly-emotive information is more likely to be encoded in detail in memory (Kensinger 2007) and therefore be accessible for decision-making. Moreover, emotions impact political information-seeking and participation. Anger tends to increase people's reliance on heuristics and preexistent viewpoints, and accordingly decreases their desire to seek out and countenance opposition viewpoints (MacKuen et al. 2010), increases reported intent to participate in politics through voting and non-voting means (Valentino et al. 2011; Weber 2013), and lowers risk perception (Lerner et al. 2003; Lerner and Keltner 2000). Fear may similarly increase intent for non-voting participation (Valentino et al. 2011). But, unlike anger, fear increases risk perception and tends to make people less reliant on preexistent opinions while also increasing the probability they seek information that

[42] Sadness also may impact participation; however, its impact is less strong than anger and fear (Brader 2005; Valentino et al. 2011; Weber 2013).

confirms rather than ameliorates their fear (Gadarian and Albertson 2014; Lerner et al. 2003; Lerner and Keltner 2000; MacKuen et al. 2010).

The few studies to have tested the link between misinformation exposure and emotions find emotional impacts, to include increased fear and anger. However, their research questions differ from mine, and therefore so does their findings' scope.[43] Social media users who actively engage with misinformation are their focus (e.g., comments, likes, shares), which excludes those of most interest to me: the majority of exposed users who do not interact with it (McLoughlin, Brady, and Crockett 2021; Vosoughi, Roy, and Aral 2018). Two other studies examine passive exposure, but do not address its causal impacts. One experimental design permitted analysis of emotional outcome by misinformation item or respondent characteristics, but did not include control group and likely was subject to demand effects (Horner et al. 2021).[44] The other was a cross-sectional survey study that established a positive correlation between perceived misinformation exposure and self-reported negative emotions regarding misinformation's influence (Lo, Xiao Zhang, and Lu 2023). This is consistent with the inference that disbelief in encountered misinformation produces negative affect but does not test it.

[43] Most instead focus on emotional antecedents of misinformation susceptibility, such as anger and negative affect (Featherstone and Zhang 2020; Greenstein and Franklin 2020; Martel, Pennycook, and Rand 2020; Rathje, Van Bavel, and van der Linden 2021; Sanchez and Dunning 2021; Weeks 2015).

[44] Respondents may have been able to guess the researchers' objective for two reasons. First, the treatment was not masked in the experiment (e.g. the inaccurate news headlines were not embedded among other questions or in an environment like a social media feed that might have made the researchers' interest less clear). Second, the questions elicited feedback directly and only about the treatment (e.g., "Knowing that this headline is FALSE, how would you feel if you saw it being shared on social media?" (2021, 1048, supplemental materials).

The second potential ripple effect the hopeful account overlooks is affective polarization—strong dislike for one's outgroup and like for one's ingroup. One study has examined the effect of misinformation exposure on affective polarization, and its results were indeterminate with respect to passive exposure (Guess et al. 2020).[45] [46] As Guess et al. argue, though, an effect is anticipated for at least CM because pro-attitudinal information exposure increases affective polarization (2020, 3-4; Garrett et al. 2014; Levendusky 2013a,b; Suhay, Bello-Pardo, and Maurer 2018). I extend their argument by arguing that NCM exposure can also increase affective polarization.

On its face, my argument runs counter to a body of literature that suggests counter-attitudinal information exposure has de-polarizing effect (Cheng, Marcos-Marne, and Gil de Zúñiga 2023; Guess and Coppock 2020; Lee and Cho 2023; Levy 2021; Lin, Haridakis, and Zhang 2020; Zhu, Weeks, and Kwak 2021). However, other results suggest the relationship between counter-attitudinal exposure and polarization is not settled (Kubin and von Sikorski 2021, 198).[47] A month of counter-attitudinal political news exposure on Twitter increased the Republicans' political opinion strength (Bail et

[45] An unpublished thesis also examined the subject. Particularly, it tested whether misinformation ("fake news belief") mediated social media use's impact on affective polarization, and it found a null result. Importantly, though, the study's model assumed fake news belief rather than fake news exposure would result in affective polarization (Daoyenikye 2023).

[46] Particularly, they found that real-world active exposure to untrustworthy news sources via web browsers on personal computers increased affective polarization, but controlled exposure to a single partisan fake news article did not. The tension in the results may be because the respondents were accidentally exposed to a media literacy intervention due to a programming error, and thus the magnitude of the "false and inflammatory news exposure treatments may underestimate the true effects of these articles" (2020. 10). Further, the partisan slant of the articles was qualitatively assessed by the authors rather than tested via a survey.

[47] "We found the literature unanimously agrees that exposure to like-minded media increases polarization. However, there is less agreement on the role of counter-attitudinal media in political polarization." (Kubin and von Sikorski 2021, 198)

al. 2018), and people who reported more encounters with counter-attitudinal cable news programs in the prior month displayed greater partisan affective polarization in another study (Gill 2022).

The tension about counter-attitudinal information's impacts may be due to selection bias and variation-masking aggregation. The selection bias may arise because those that report greater exposure to counter-attitudinal information or opt-in to follow outgroup partisan media are likely those more open to attitude modification to begin with (Bertrand and Duflo 2017, 380–81; Heatherly, Lu, and Lee 2017; John and Dvir-Gvirsman 2015).[48] The potential aggregation issue emerges from the no-backlash studies' operationalization of counter-attitudinal exposure. They typically operationalize it based on source characteristics, e.g., a news outlet more associated with one party (e.g., CNN) is counter-attitudinal for opposition party members (Lin, Haridakis, and Zhang 2020). However, not all information from an outgroup source may be equally counter-attitudinal. An encounter with an outgroup member's social media profile can reduce negative assessments by increasing perception of shared similarities , while encounters with partisan cable news that lambasts one's in-group can increase negative out-group perceptions (Gill 2022).

The potential variation beneath the aggregation is consistent with intergroup contact theory. Just as outgroup contact's effect on prejudice can depend on the valence of the contact, so the impact of what I call informational outgroup contact (IOC) may

[48] For example, Chen et al. (2022) found that the de-polarizing effect of counter-attitudinal exposure is concentrated among people most open to incorporating opposition viewpoints into their own. Those that reported lower perceived utility of opposition viewpoints actually were more polarized after exposure to them.

hinge on the valence of the contact (Barlow et al. 2012; Hangartner et al. 2019; Paolini, Harwood, and Rubin 2010). Depending on the congeniality of the information in question, one would expect negative informational out-group contact (NIOC) to sometimes be washed out or overpowered when aggregated with positive informational out-group contact (PIOC) (Trilling, Van Klingeren, and Tsfati 2017, 20). For example, Lin et al. (2020) found no relationship between counter-attitudinal news exposure regarding the U.S. 2012 election and intergroup bias/competition. However, the study design assumed that MSNBC and CNN were counter-attitudinal for Republicans and Fox News was counter-attitudinal for Democrats. As the authors note, the net valence of the content could actually have been positive or neutral rather than negative—something that seems probable given the above-discussed potential of selection bias (2020, 2457). Outgroup media content does not all deride one's ingroup, and one is probably more likely to self-select into precisely that sort of outgroup media content.

Largely unexplored is a third potential ripple effect: trust.[49] As with emotions and affective polarization, passive exposure to misinformation may impact trust through two pathways. Those who believe CM attacking or promoting certain groups may experience correlative mistrust/trust in said group, and those who disbelieve NCM may mistrust those who author it and are apparently influenced by it. For example, Nisbet, Cooper, and Garrett (2015) found that exposure to counter-attitudinal science information decreased trust in scientists, and Ognyanova et al. (2020) found that exposure to pro-Trump

[49] Trust, both an affective and cognitive concept, is a "social bond... characterized by feelings of security and confidence in others' good intentions and good will" (Tropp 2008, 91).

administration misinformation decreased trust in government among strong liberals and had the opposite effect on conservatives.

No study has tested that backfire effect on trust in general and in political outgroups. Indeed, no study has tested the impact of misinformation exposure on those two categories of trust at all, as the literature's focus is on trust in various communicators e.g., news media, scientists, fact checkers (Agley and Xiao 2021; De Coninck et al. 2021; Krishna 2021; Nisbet, Cooper, and Garrett 2015; Sullivan 2019; Vinck et al. 2019). Generalized trust is one's trust in others regardless of their identity, and is at times conceptualized as a dispositional trait rather than a knowledge-based trust of a particular person based on their previous history (Freitag and Traunmüller 2009; Schilke, Reimann, and Cook 2021, 243). Important to the well-functioning of democracies, this sort of trust is a driving factor in positive engagement with society beyond one's immediate social circles (Putnam 2000). Trust in political outgroups similarly matters, with impacts on political participation, policy content, and cooperation.[50]

Third Person Effect: Amplifying Misinformation's Effect on Those that Disbelieve It

Beyond the evidence discussed above, there is another reason to expect that misinformation will impact even those that disbelieve it: the "third person effect" (TPE). TPE is the tendency for people to think that others are more influenced by mass media

[50] Some trust scholars consider trust in political in-groups to be an example of particularized trust, which varyingly is identified as trust in people like the truster or knowledge-based trust of known persons (Freitag and Traunmüller 2009; Smith 2010; Uslaner 2002; Yamagishi and Yamagishi 1994). By extension, one might consider out-group trust to be an example of generalized trust. However, several studies confirm a third sort of trust based on group identity exists in the radius between narrow particularized trust and broader trust in people in general/strangers (Carlin and Love 2018; Freitag and Bauer 2013; McPherson, Smith-Lovin, and Cook 2001).

than they themselves are (Davison 1983; Sun, Pan, and Shen 2008).[51] Generally, the more undesirable a person finds a message's potential influence, the more likely they assess it will influence others, especially outgroups. This pattern is not straightforward in all contexts, but is evidenced with respect to political misinformation and counter-attitudinal political information.[52] With regard to the former, people in cross-national contexts assess distant others as less likely than themselves or close others to identify "fake news" or inaccurate information and more likely to have it influence their beliefs and behavior (Altay and Acerbi 2023; Chen, Yu, and Liu 2023; Cheng and Luo 2020; Chung and Wihbey 2022; Yang and Tian 2021; Yoo, Kim, and Kim 2022). With regard to the latter, the more undesirable (more negative) a respondent finds the potential influence of a piece of political information, the greater the effect they assess it will have on political outgroups (Baum, Meissner, and Krasnova 2021; Hyun and Seo 2021; Lee and Kim 2022; Meirick 2004; Wei, Chia, and Lo 2011).

If people were unskilled at identifying NCM as misinformation, little NCM exposure would result in TPE-heightened concern about its influence. However, that is

[51] Lyons (2022) found TPE is not necessarily a bias, as 60-70% of respondents in three studies who had high TPE with regard to false news identification were also better at false news identification.

[52] TPE's relationship with in-/out-group dynamics is not always straightforward because perceived distance (self-outgroup, message-outgroup) interacts with message desirability and social learning in complex ways (Cho and Boster 2008; Huge, Glynn, and Jeong 2006; Tsfati and Cohen 2004). Social distance and message desirability generally reduce TPE, but there likely are other mechanisms, as a study that sought to disaggregate by mechanism (self-candidate distance, group-candidate distance) had indeterminate results (Meirick 2004). Social learning's impact varies, depending on the outgroup and message in question. For example, all three main U.S. partisan groups (Republicans, Democrats, Independents assessed that Independents were most likely to be influenced by a presidential debate, and young U.S. voters assessed that political comedy would more influence Democrats than Republicans (Hoffner and Rehkoff 2011; Wei, Lo, and Zhu 2019). This likely is because (1) respondents viewed independents as most likely to be swing voters, and (2) nearly unexceptionally, U.S. political comedy shows are left-leaning.

not the case. People are generally effective at correctly identifying NCM as inaccurate (Chen, Yu, and Liu 2023; Corbu et al. 2020, 170–73; Pennycook and Rand 2021b; Ștefăniță, Corbu, and Buturoiu 2018; Tang, Willnat, and Zhang 2021).[53] Thus, NCM encounters are likely to result in identifications of the misinformation as misinformation, and in turn to be followed by appraisals that others are influenced by its message—particularly relevant outgroups. Given the influence of NCM on its probable audience is undesirable, one would not only perceive that its audience is being duped, but that their being duped will have negative impacts to their opinions and behaviors. These perceptions feed into the three aforementioned ripple effects, a relationship likely to be particularly pronounced in the United States. Its partisan polarization and social sorting increase distinguishability between parties in terms of platform and narrative, thereby increasing peoples' ability to discern the congeniality and non-congeniality of a given piece of political information (Iyengar, Konitzer, and Tedin 2018; Mason 2018).

The potential that NIOC in the form of outgroup misinformation negatively impacts assessments of the outgroup via TPE is consistent with novel analysis I conducted of a survey dataset from Nisbet et al. (2020; 2021) (Appendix C). Nisbet et al. originally used the dataset to test the relationship between TPE about misinformation and satisfaction with U.S. democracy.[54] Importantly, their study asked respondents about

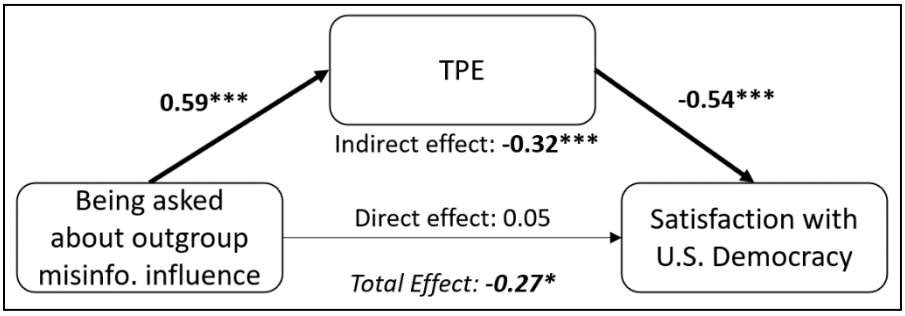
[53] People are better at correctly identifying non-congenial than congenial misinformation as misinformation. They do have better truth discernment with respect to congenial misinformation, but that is because they tend to have lower belief levels in accurate non-congenial political information (Pennycook and Rand 2019; 2021b; 2021a).

[54] The study found that greater TPE with regard to misinformation's influence on other voters correlated with greater dissatisfaction with U.S. democracy. Influence on other voters was operationalized as an index measure, that included perceived influence on other voters' election issue opinions and vote choice This finding is echoed in Ross et al. (2022). They found that U.S. partisans' perceptions that their opponents' 2016 vote choice was influenced by Russian disinformation correlated with decreased trust in

various misinformation sources, to include conservative and liberal groups/organizations. This design element permitted me to identify the in/outgroup association of the queried misinformation for each respondent, based on respondent partisanship.[55] When I used this in/outgroup status as the independent variable in a mediation analysis with TPE as the mediator, the results show the anticipated outgroup asymmetry (Mize n.d.; Preacher and Hayes 2008). There is a statistically significant relationship between being asked about outgroup misinformation influence and democratic unsatisfaction, wholly mediated by TPE (Figure 1). The effect size is small, as the total effect of -0.27 represents a 2.7% change in the 10- point range of democracy satisfaction. However, the effect could be stronger if respondents were asked about misinformation influence on partisan outgroups rather than voters

Figure 1. Novel Mediation Analysis of Nisbet et al.'s Data

| ρ values: 0.1, 0.5**, 0.01*** |*



Controls were same as those used by Nisbet et al., with the addition of a dichotomous partisanship measure: attention to politics, voted in the 2016 U.S. election, ideology, income, whether they were self-identified evangelical Christians, sex, race, age.

---

the election results and decreased satisfaction with U.S. democracy (the exact relationships varied by partisanship).

[55] I operationalized (1) being *asked about outgroup misinformation* as whether or not a Republican or Democrat respondent was asked to evaluate the influence of "false or misleading news stories from liberal[conservative] political groups and organizations", and (2) *TPE* as the difference between the respondents' reported perceived influence of misinformation on other voters and on themselves (Nisbet et al. 2021, 12).

in general, and if respondents were exposed to outgroup misinformation rather than asked to assess its influence in the abstract.

This result does not establish a relationship between NIOC in the form of outgroup misinformation and negative outgroup assessments, but it does establish that TPE is higher for outgroup misinformation than ingroup misinformation in general. Further, the results do not support the general null hypothesis that there is no relationship between assessments of outgroup misinformation influence and negative assessments of related attitude objects.

Theoretical Expectations

Based on the above literature and the exploratory empirical analysis, I have two expectations. First, I expect passive exposure to ingroup political misinformation (CM) to cause belief in that misinformation. Whether through accuracy- or directionally motivated reasoning, partisans are likely to accept information that most conforms to their preexisting beliefs. Second, I expect passive exposure to partisan political misinformation to decrease trust in one's outgroup party and people in general, increase affective polarization, and to be able to increase the anger and fear of both those who find it congenial and non-congenial. The reasons for this expectation both converge and diverge between the groups.

For those that find the misinformation congenial, the expected effects ripple from misinformation belief. Whether the result of ego protection—as suggested by social identity theory—or simple Bayesian updating regarding a socially-distant, competitor outgroup, acceptance of negative information about an outgroup likely is accompanied by

both attitudinal and affective results. For those that find the misinformation non-congenial, the expected effects ripple from misinformation disbelief. This, too, may be the result of ego protection and/or Bayesian updating, and the perceived inaccuracy of the information further introduces the amplifying dynamic of TPE. As suggested by intergroup contact theory and TPE literature, NIOC such as identifying undesirable misinformation intended for a political outgroup increases the perceived influence of misinformation on that outgroup. This perceived influence likely results in updated negative affect toward and assessments of that outgroup, to include affective polarization and mistrust in outgroups and others in general.

The ripple effects' magnitude may differ between the two groups (CM, NCM), but I expect effects to be significant. Finally, I also ask if respondents who opt-in to share social media and browser activity meaningfully differ from those who opt-out. Below are the associated hypotheses:

Hypothesis 1 (Belief): passive exposure to CM will cause belief in that misinformation.

Hypothesis 2a-b (Trust): passive exposure to partisan misinformation will decrease (a) GT and (b) OPT, regardless of whether the misinformation is CM or NCM.

Hypothesis 3 (Polarization): passive exposure to partisan misinformation will increase affective polarization, regardless of whether the

misinformation is CM or NCM.

Hypothesis 4 (Emotion): passive exposure to partisan misinformation will evoke (a) anger and (b) fear, regardless of whether the misinformation is CM or NCM.

Hypothesis 5 (TPE): passive exposure to partisan misinformation will increase TPE regarding misinformation's influence on its likely audience, provided one disbelieves the misinformation. In other words, (a) disbelief in CM will increase TPE regarding one's ingroup party, and (b) disbelief in NCM will increase TPE regarding one's outgroup party.

Hypothesis 6 (Opt-In/Out): respondents who opt-in to share their digital history differ from those who don't in terms of partisanship strength, trust in outgroups/government/media/experts/fact checkers, affective polarization, misinformation belief, and/or misinformation active exposure.

Method

I conducted a between-subjects survey experiment, with three treatment conditions: (1) control, (2) CM, (3) NCM. In each condition, respondents were exposed to four headlines. The treatment groups saw two misinformation and two real headlines, and the control group saw the same two real headlines along with two other real ones (eight

headlines total used). The experiment's participants were a convenience sample from CloudResearch Connect (*n* = 868). The sample had quotas for partisanship to permit comparison between Republicans and Democrats.[56] Prior to the experiment, I validated the misinformation treatment items were roughly equal in terms of plausibility/partisan valence via a separate convenience sample from CloudResearch Connect (*n* = 200). I originally attempted the treatment validation with Mturk, but found major sample quality issues akin to those described by Webb and Tangney (2022) (Appendix D).[57] In all the samples, respondent compensation was $1.42. All surveys were IRB-approved; their approval letter and question wordings are in Appendix E-G).

Data & Operationalization

The study's independent variables are passive exposure to CM/NCM and opting-in to grant researchers access to their anonymized social media activity. The conceptual dependent variables are belief, trust (GT, OPT), emotion (anger, fear), and affective polarization. They are operationalized as follows.

Passive exposure to CM/NCM (Independent Variable): Misinformation exposure is captured in three dichotomous variables that indicate whether respondents saw *CM*, *NCM*, or no partisan misinformation (*control*). Misinformation is categorized as CM if the presented misinformation favors the respondent's party, and NCM if it favors the respondent's outgroup party. Respondents' partisan identities were operationalized as

[56] The official quotas were 35% Democrat, 40% Republican, and 25% Independent. This was not because I wanted an unbalanced sample between Democrats and Republicans, but because I knew from a treatment validation study that Independents on CloudResearch had more Democrat-leaners than Republican leaners.

[57] Large majorities of the responses appeared to be inauthentic or inattentive.

follows: Republicans(Democrats) were those respondents who identified as Republican(Democrat) or Republican(Democrat)-leaning. Thus, true moderates/independents are dropped from the sample for all analyses save the opt-in/out analysis.

Opt-in (Independent Variable): Opt-in is a dichotomous measure capturing whether respondents agree to a question asking whether they are willing to grant researchers access to anonymized web browser history and/or social media activity. The study has no way to capture or access such data; the goal of the question is to identify opt-in/out willingness.

Belief (Dependent Variable): Belief is a family of five-point measures capturing respondents' assessment of how likely each of the eight headlines is to be true ("likely" to "unlikely").

GT (Dependent Variable): The primary measure of *GT* attitudes are responses to the statement: "You can't count on strangers." Answers are dichotomous ("more or less agree", "more or less disagree") (Glaeser et al. 2000).[58] Though single measures of latent variables tend to be less reliable than indexed measures, it is the primary measure because it consistently correlates with self-reports of trusting behavior and trusting behavior in "trust games" (Capra, Lanier, and Meer 2008, 44; Gächter, Herrmann, and Thöni 2004, 521; Glaeser et al. 2000, 826–29).

[58] The question is a modification of Glaeser et al.'s question, which was "You can't count on strangers anymore" (2000). I omit the time element because it changes the question from trust in strangers to trust in strangers relative to the past.

Due to disagreement regarding the quality of various trust measures, two alternate GT measures are also captured. First is the embattled but longstanding dichotomous question: “Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?” It only sometimes correlates with trusting behavior in game settings (Bauer and Freitag 2018; Capra, Lanier, and Meer 2008), possibly captures caution as well as trust (Miller and Mitamura 2003), and can be interpreted differently by different respondents (Sturgis and Smith 2010). However, the question remains standard due to its simplicity, lack of a better alternative, cross-comparability with previous studies, questions of whether one should expect GT to result in reciprocal behavior in a trust game, and evidence that some countries correctly interpret “most people” to mean unknown rather than known persons (Delhey, Newton, and Welzel 2011; Doyle 2021; Uslaner 2015). It is most useful for the purposes of this survey for the latter three reasons, especially as the United States is one of the countries that interprets the question to mean “most people.” The second alternate measure is the three-item measure from the General Social Survey (GSS), comprising the standard question and two questions on most people’s fairness and honesty.[59] Each of the three questions are dichotomous, and responses are averaged together (Cronbach’s alpha: 0.84).

[59] The questions are: “Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves?” and “Do you think that most people would try to take advantage of you if they got the chance or would they try to be fair?”

Outgroup Party Trust (OPT) (Dependent Variable): Self-reported trust attitudes toward one's outgroup is measured via a question regarding trust in a given group of partisans (Democrats, Republicans), on a five-point scale ("none" to "a great deal"). The outgroup party's identity is the opposite of the respondent's partisanship. This measure of trust is a less-generalized version of the World Value Survey's current measure of outgroup trust, which is an index of "people you meet for the first time", "of another religion", and "of another nationality" (Delhey, Newton, and Welzel 2011).

Emotions (Anger, Fear; Dependent Variables): *Anger* and *fear* five-point index variables drawn from a battery of questions that asks "Generally speaking, how do you feel about the way things are going in the country these days? Please tell us how much you feel each of the following emotions" (Valentino et al. 2011, 163). The emotions measured are angry, disgusted, outraged, afraid, nervous, hopeful, proud, and happy, with the latter three emotions being chaff. As anticipated from Valentino et al., factor analysis confirms that the three anger-related variables load onto a different factor than the two fear-related ones. From these two groups of variables, I created two averaged index variables for anger and fear (respective Cronbach's alpha of 0.91 and 0.86).

Affective Polarization (Dependent Variable): Affective polarization is measured as the difference between two feeling thermometers regarding Republicans and Democrats, with the outgroup being subtracted from the ingroup.

Political Knowledge, Political Interest, Partisanship Strength, Frequency of Religious Service Attendance, Age, Ethnicity, Education, Trust in Media, Trust in Government, Trust in Experts, Trust in Social Media Fact Checkers, Active Exposure (Other Independent Variables): These variables are included on the basis of their known

relationship with GT and political misinformation susceptibility in the United States (Schilke, Reimann, and Cook 2021; Tucker et al. 2018). *Political knowledge* is a two-point index of three items adapted from Miller et al. (2016). The three items had low inter-reliability (Cronbach's alpha = 0.2805), so I only employ the measure in the opt-in/out analysis.[60] *Political Interest* was a five-point index variable averaging respondents' interest in this year's campaigns and politics in general (Cronbach's alpha = 0.8059). The *partisanship strength* measure is ordinal. Respondents who self-identify as a strong Republican/Democrat are coded as 3, not very strong Republican/Democrats are a 2, and independent leaners are a 1.

Religious service attendance frequency (*religious attendance*) is a five-point measure ranging from "never" to "every week."[61] *Age* is continuous from 18-99. *Ethnicity* is a dichotomous variable indicating white (0) and non-white (1). *Education* is an eight-point measure from "less than a high school diploma" or "no formal education" to "professional" or "doctorate degree." The latter variables of *age*, *ethnicity*, and *education* are provided by the polling firm, and therefore do not appear in the survey questions. *Trust in media, government,* and *social media fact checkers* are measured like

[60] The items were: which party is more conservative, which party has a majority in the House of Representatives, and the size of Congressional majority required to overturn a presidential veto. Respondents' political knowledge was high, with 96.43%, 75.35%, and 72.4% answering the respective questions correctly. This could suggest that CloudResearch respondents are more likely to look up political knowledge questions. The polling firm touts their respondents' attentiveness, perhaps creating a perverse incentive to answer questions "correctly." However, only 4.26% of respondents correctly answered an initial question intended to identify respondents who look up answers and nudge them not to ("In what year did the Supreme Court of the United States decide Geer v. Connecticut?") (ANES 2021, 311–12). This suggests that rather than looking up answers, the CloudResearch sample simply is more politically knowledgeable than the general population.

[61] I do not include religion as an independent variable because the studies regarding the relationship between religion and GT are done at a higher level of analysis (country) than this study (Delhey and Newton 2005)

OPT on a five-point scale ("none" to "a great deal"). *Trust in experts* is also measured on a five-point scale taken from ANES 2020, ranging from "[trust ordinary people] much more" to "[trust experts] much more." Finally, *active exposure* contains a series of eight dichotomous variables (one per headline) capturing whether the respondent subsample who initially saw the headline self-reported being likely to read the headline.

Experimental Procedures

The study had two parts: treatment validation, and survey experiment. The treatment validation permitted identification of four misinformation items (two that favor each party) that were of roughly equal plausibility and had clear partisan slant. The validation consisted of asking Connect respondents to evaluate sixteen researcher-generated misinformation headlines along those two dimensions. Below is the wording of the questions and the sixteen misinformation items (Table 6).

Table 6. Treatment Validation Content

Dimension Questions	Below are sixteen headlines. What is each headline's likelihood of being true? *(note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true).*
	Assuming the headlines are entirely accurate, how favorable do you find them to Republicans versus Democrats? *(note: a headline is more favorable to one party if you think it makes that party look good and/or their opponent party look bad).*
Pro-Dem. Misinfo.	White governor names own child after racial slur to protest NFL team name change
	NRA claims liberal activists conduct school shootings to undermine gun rights
	Poll: alongside abortion bans, most Evangelicals want to force women to marry their rapists
	Repub. candidate DeSantis signals support for law to execute women that receive abortions
	Senior Alabama senator: arm the homeless to protect against active shooters, provide dignity of work

	Trump directed Mar-e-Largo aides to burn classified documents to prevent FBI discovery
	Fox News journalist says swastika-carrying Nazis are 'misunderstood'
	Texas' proposed history textbooks suggest room and board compensated slaves for their labor
Pro-Rep. Misinfo.	MSNBC's secret internal policy to fire commentators critical of welfare fraud
	Large minority of migrant "families" crossing border are child sex traffickers and their victims
	New York Times covered-up illegal FBI surveillance of conservative candidates
	Russia decided to invade Ukraine after learning "weak" Biden would succeed Trump as president
	Secretary of Educ. plans to fire kindergarten teachers that refuse to encourage all students to question their gender
	Dem. governors manufacture prison "overpopulation" to justify early release of Dem. voters
	Black Lives Matter leaders privately confirmed their support for violent redistribution of wealth
	Saving the planet or lining their pockets? California politicians promote self-serving climate change regulations

Based on the results (Table 7), I selected "Trump directed Mar-e-Largo aides to burn classified documents to prevent FBI discovery" and "Texas' proposed history textbooks suggest room and board compensated slaves for their labor" as the pro-Democrat misinformation.[62] The pro-Republican misinformation was "Large minority of migrant 'families' crossing border are child sex traffickers and their victims" and "New York Times covered-up illegal FBI surveillance of conservative candidates." I selected the items based on four criteria.

First, they were in the top four of their respective misinformation groups in terms of difference-in-means between the Republican and Democratic respondents. That is, they were among the headlines whose truthfulness the two parties most disagreed about. Second, their respective means indicate that one party assessed the headline as being more likely than not (mean > 3) while the other assessed the opposite (mean < 3). Third, the headlines' believability to the partisan ingroup was roughly equal. The most plausible misinformation item for Democrats and Republicans respectively were the "burn

[62] After fielding, I learned that I had misspelled Mar-a-Lago as Mar-e-Lago. However, the error does not seem to have impacted the headline's anticipated believability.

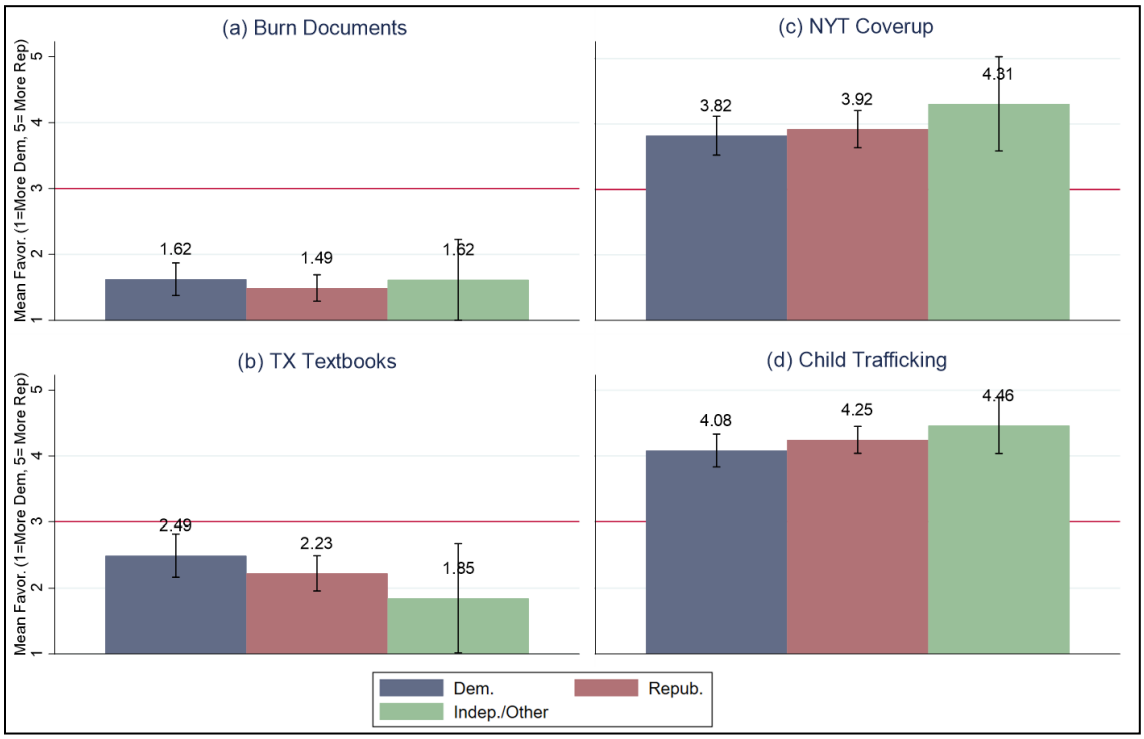
documents" and "NYT coverup" items (means of 3.81 and 3.67) and the next most plausible were those regarding "TX textbooks" and "child trafficking" (means of 3.24, 3.21).

Table 7. Difference-in-Means: Likelihood of Headline Being True

	Repub.	Dem.	Diff	P-Stat
True: Racial Slur Name	1.627	1.729	0.102	0.50
True: NRA False Flag	2.245	2.471	0.225	0.29
True: Evangelicals	1.412	2.059	0.647	0.00***
True: DeSantis	1.304	2.047	0.743	0.00***
True: Arm Homeless	1.755	1.941	0.186	0.28
True: Burn Documents	2.461	3.812	1.351	0.00***
True: Fox Reporter	1.804	2.824	1.020	0.00***
True: TX Textbooks	2.255	3.235	0.980	0.00***
True: MSNBC Policy	2.902	2.306	-0.596	0.00***
True: Child Trafficking	3.206	1.871	-1.335	0.00***
True: NYT Coverup	3.667	2.271	-1.396	0.00***
True: Russia Invasion Timing	3.049	2.271	-0.778	0.00***
True: Question Gender	2.490	1.788	-0.702	0.00***
True: Prison Overpopulation	2.716	1.659	-1.057	0.00***
True: Climate Grift	4.186	3.424	-0.763	0.00***
True: BLM Marxist	2.382	1.612	-0.771	0.00***
Observations	187			

Fourth, the difference-in-means for party favor confirmed the headlines had a clear valence in the expected direction (Figure 2). The bipartisan means for the two pro-Democrat headlines were firmly on the favors-Democrats side of the scale (< 3, Figures 2a-2b), while the bipartisan means for the two pro-Republican headlines were firmly on the favors-Republicans side of the scale (> 3, Figures 2c-2d). Further, the magnitude of the party favor was roughly balanced between the Democrat and Republican headlines. Both sets of headline sets had one headline with strong partisan slant ("Burn Documents", "Child Trafficking") and another with moderate partisan slant ("TX Textbooks", "Child Trafficking").

Figure 2. Partisan Favorability Ratings of the Four Misinfo. Items



I then conducted the survey experiment. CloudResearch directed respondents to a Qualtrics survey advertised as a social media dynamics study. To reduce risk of non-U.S. respondents taking the survey, I employed a free service called IPHub to screen out participants who had IP addresses outside the United States or IP addresses associated with VPNs or similar location-masking services (Winter et al. 2019).[63] Respondents were notified of this requirement in the consent form.

The consent form page had a "trap" question that was visually hidden and therefore only answerable by bots. It was the first of three sample quality questions. After the consent form and automatic IPHub screening came the next two sample quality

[63] I learned of an improvement to be made to the IPHub screening protocol: namely, how to not accidentally exclude respondents from Puerto Rico. Puerto Ricans are U.S. respondents, but their IP address block is listed separate from the United States, and so a simple filer for all respondents without a U.S. IP address is insufficient.

questions, which the survey termed "common knowledge questions."[64] The first asked respondents to identify AI-generated photorealistic images of a flashlight, a pair of sandals, and a tire. This served to identify the proportion of respondents that learned English outside United States (the former would tend to refer to those objects as a torch/torchlight, chappals, and a tyre).[65] The second question asked respondents "What wrd is undrelned in tihs qestuion" to assist in bot identification. The descriptive statistics from the three sample quality questions suggest the sample was actual humans from the United States. No one answered the hidden trap question, only 1.27% answered the image identification question with answers involving "torch" or "tyre" (no answers included "chappals"), and only 0.58% did not answer "in" in some form to the third question.

Next was the pre-survey. It asked respondents' religious service attendance, partisanship/partisan strength, political interest, political knowledge, and social media habits. There are three items of note for this question section. First, one of the political interest variables was censored for 550 respondents. These respondents only had four options for how often they pay attention to politics and government ("always", "most of the time", "about half of the time", and "some of the time"), as the fifth option of "never" was accidentally overwritten with gibberish. No respondents selected the gibberish. Second, the political knowledge questions were prefaced with language to discourage

[64] Because such simple questions could cause some respondents to wonder if the questions were trick questions, the survey prefaced the questions with "Important! The questions are very easy and the answers will seem obvious, because they are."

[65] Of course, valid respondents could have learned English outside the United States, and so this measure was not singly used to disqualify any respondents. It instead served to highlight whether an unrealistic proportion of respondents seemed to use non-US English.

looking up answers, drawn from the American National Election Survey.[66] Third, the social media section was mostly chaff questions to obfuscate the true focus of the study, but within that section was the key question about whether respondents were willing to grant researchers access to their social media/browser history.

Then came treatment exposure. Respondents were presented four article headlines purportedly "shared by randomly-selected social media users" who had "endorsed the articles as being interesting, important, or useful to know."[67] All of the respondents saw the same true headlines ("Facial recognition firm Clearview AI used nearly 1m times by US police", "Afghanistan girls' education activist arrested by Taliban"). For the other two headlines, some respondents saw the pro-Democrat misinformation, others saw the pro-Republican misinformation, and others saw two further real headlines ("Flat-packed pasta could help revolutionize food production", "Brain cancer patient prepares to run London Marathon").[68] To encourage passive exposure beyond the simple reading task and provide a plausible reason for the headlines being in the survey, respondents were asked to imagine scrolling through their social media feed and coming across any of the

[66] The language is: "we are interested in the guesses people make when they do not know the answer to a question. We will ask you four questions. Some may be easy, but others are meant to be so difficult that you will have to guess." That statement is followed by two answer options: "I promise to try my best without looking up any answers" and "I do not want to make that promise." Finally, the instructions end with a question to catch cheaters and nudge them to no longer cheat. Respondents who correctly identify the year an obscure U.S. Supreme Court case was decided receive a follow-up question asking them if they'd already known that answer or just looked it up.

[67] This language is drawn from Bachleda et al. (2020).

[68] The real headlines were from *BBC* (AI, Afghanistan), *ABC* (pasta), and *SWLondoner* (marathon) (Choi 2021; Clayton and Derico 2023; Haase 2024; Yong 2023). The pasta headline was slightly modified, dropping "groovy" from the original headline of "Groovy flat-packed pasta could help revolutionize food production" due to its semantic ambiguity in the absence of an accompanying picture. Since selection, the AI headline appears to have been expanded by the BBC, as it has an addendum I do not recall having been there: "[…]it tells the BBC."

articles. They were then asked which of the articles they would read, if any, with the ability to select as many as applied. This provided the active exposure data I use in the opt-in v. opt-out analysis.

After treatment exposure, respondents were first asked about the key DVs. Namely, they were asked the emotion battery, how they felt toward the two main political parties, their GT (all three measures), and a particularized trust battery that captured trust in Republicans, Democrat, the media, government, and social media fact checkers. They also were asked their trust in experts v. ordinary people. Finally, respondents were also asked their belief in the headlines they had seen as well as the other four headlines in the study. Belief's placement as the last question in the section accomplished two things. First, it avoided prompting respondents to evaluate the misinformation's accuracy when they might not have done so of their own accord (Pennycook and Rand 2019). Such priming could have influenced the other DV results. Second, it permitted me to capture the control group's belief in the misinformation headlines, as their responses for the other DVs were already registered before encountering the misinformation for the first and only time in the belief question.

The survey concluded with three manipulation checks, a placebo question, and the final DV question of TPE. The first manipulation check, an instructional manipulation check, queried respondents' awareness of the information I provided about the purported social media sharers' partisanship. The second, a subjective manipulation check for the treatment groups, asked what party they thought the article sharer belonged to. This served to check whether respondents inferred the sharer's party from the misinformation as expected. Next, a placebo question asked all respondents if they remembered "having

seen any of the article headlines" before the study participation. This served to identify if there was systematic illusory truth underlying the belief results. After that was the final manipulation check (subjective) for control group respondents only, which asked the party favorability of the four headlines they originally saw. This was to check whether the four real headlines were neutral in terms of partisan slant, as I had not included them in the treatment validation.

The final question regarded TPE, asking respondents how much influence they thought "inaccurate information (misinformation)" had on the "political opinions" of themselves, people in general, Democrats, and Republicans. It was asked last to avoid tipping off respondents that the study was about misinformation. Its order, however, means that there is no true control group for the TPE analysis, as the control group was exposed to the four misinformation headlines in the belief question. The study concludes with a debriefing message that identified: me as the source of the article headlines, which headlines were inaccurate, that I had no access to their social media activity/web browser history, and the purpose of the survey. Respondents were then automatically redirected to the CloudResearch Connect site for compensation.

Analysis

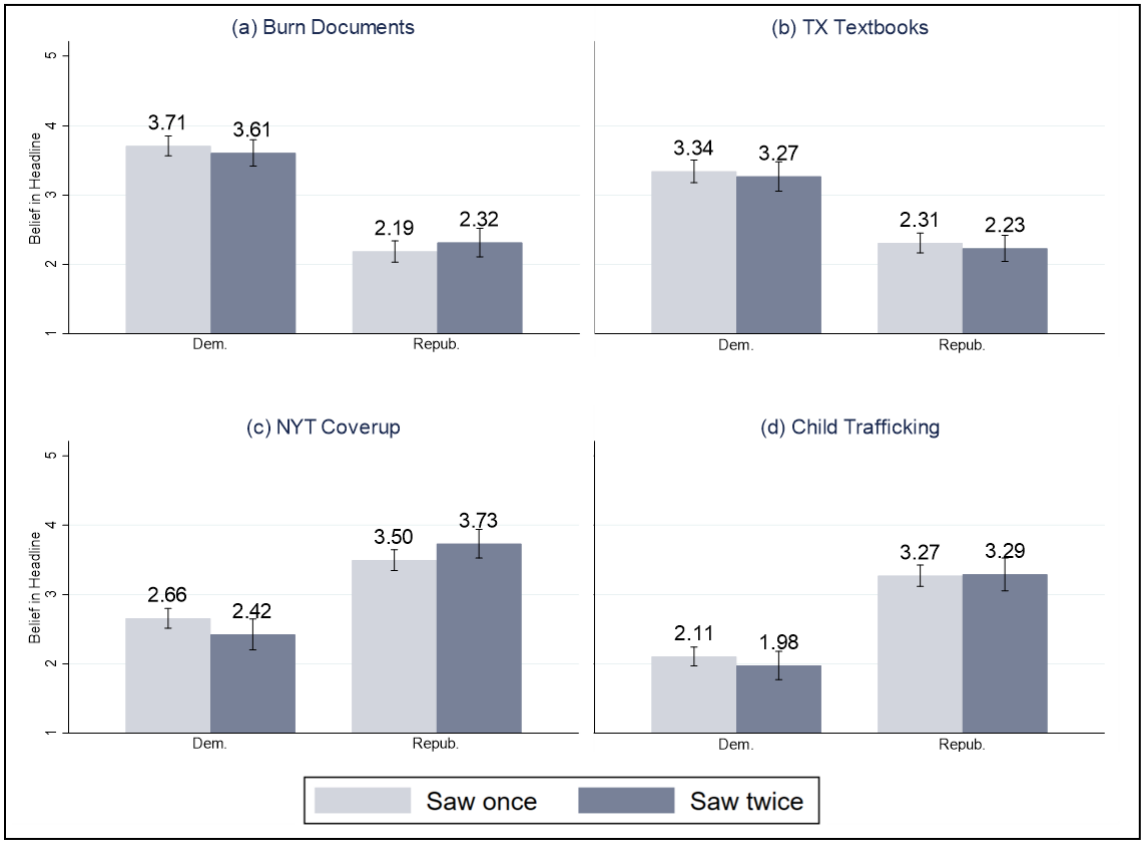
The results supports Hypothesis 1 (belief), provides limited support for Hypotheses 2, 4, and 5 (GT, anger/fear, TPE), and does not support Hypothesis 3 (affective polarization) or Hypothesis 6 (opt-in/out).

Hypothesis 1 (Belief): Supported

Difference-in-mean and regression results support Hypothesis 1. Democrats on average rated pro-Democrat misinformation as more likely to be true than not, with mean ratings

being between “equally likely and unlikely” (3) and “somewhat likely” to be true (4) (Figure 3a-3b). Republicans, as expected, found the pro-Democrat misinformation less plausible, with average belief scores between “equally likely and unlikely” (3) and “somewhat unlikely” (2). The pro-Republican misinformation demonstrates the same pattern (Figure 3c-d). Republican belief scores were over 3 for both items, while Democrat belief scores were under 3. These results are unsurprising, as the treatment validation had demonstrated that the partisan-slanted headlines had a corresponding partisan asymmetry in belief.

Figure 3. Misinformation Belief by Partisanship and Exposure Count



Beyond replicating the treatment validation’s result, the survey experiment further permits analysis of illusory truth. All respondents saw their two unique headlines twice—

once in treatment exposure and once in the belief question—and only saw other groups' unique headlines in the belief question. This repeated exposure could have increased belief. Simply put, though, the means do not show an illusory truth effect. There are no statistically significant differences between those that saw a given misinformation headline once versus twice, regardless of respondent partisanship. This does not contradict the hypothesis but could suggest that simple repeated exposure to a claim over the course of a few minutes does not necessarily increase belief in that claim.

Also supporting Hypothesis 1 is a regression analysis using an ordinal logistic model (Table 8). Being Republican versus a Democrat corresponds with 73-86.9% lower odds of belief in the two pro-Democrat misinformation items (burn documents, TX textbooks). Similarly, being a Republican corresponds with a 111.9-240% increase in odds of belief in the two pro-Republican misinformation items (NYT coverup, child trafficking). These results are found while controlling for repeated exposures to the misinformation during the experiment, the interaction of repeated exposure and partisanship, perceived illusory truth, trust in media, and low attention to the scenario (misidentifying what info one was given about the article sharers' partisanship). Brant tests suggest that these congeniality results come with precision limitations due to violations of the model's parallel regression assumption. However, the violations do not undermine the validity of the results. Further, these results are robust to specifications excluding the potential confounders (perceived illusory truth, trust in media, low attention). See Appendix H for the robustness checks.

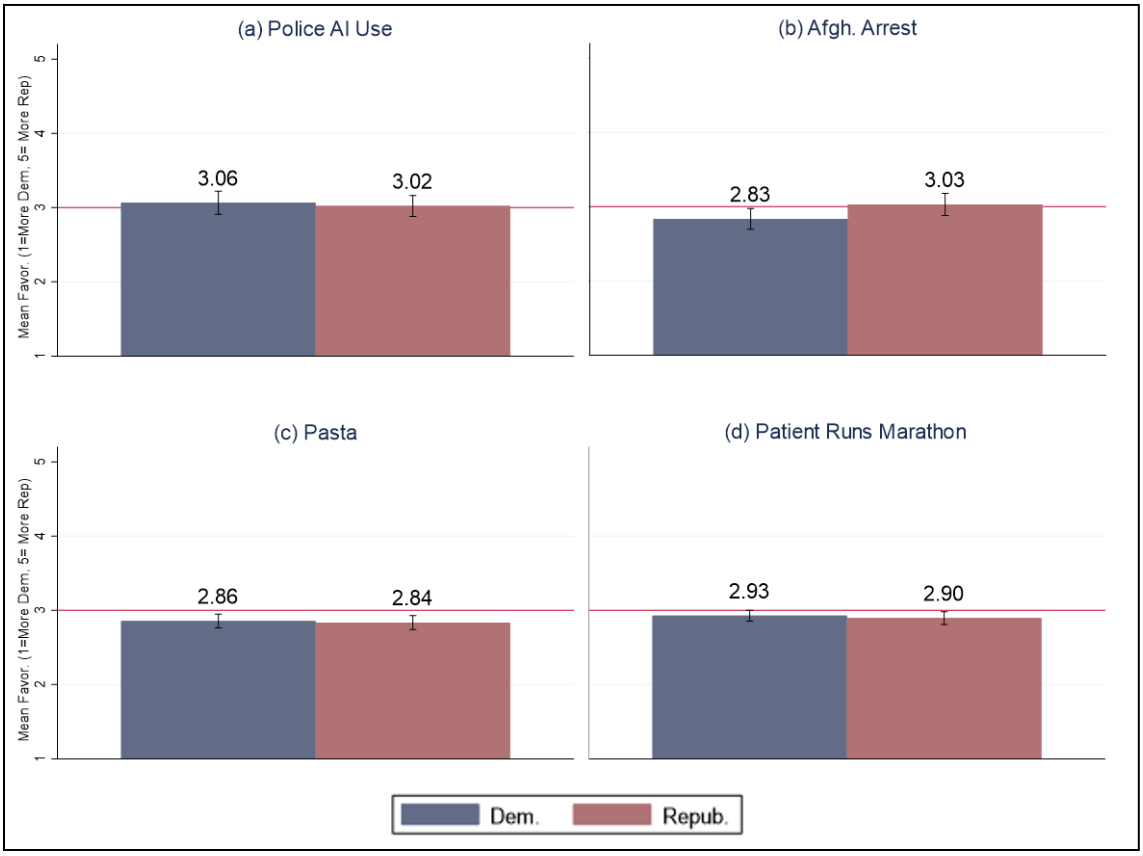
Table 8. Estimated Impact of Exposure on Belief (Odds Ratios)

VARIABLES	(1) Burn Docs	(2) TX Textbooks	(3) NYT Coverup	(4) Child Trafficking	(5) Police AI Use	(6) Afgh. Arrest	(7) Pasta	(8) Patient Runs Marathon
Republican	0.141***	0.270***	2.119***	3.400***	0.814	0.746**	0.664**	0.708**
	(0.026)	(0.047)	(0.350)	(0.577)	(0.117)	(0.109)	(0.109)	(0.119)
Saw Twice	0.820	0.903	0.618**	0.676*	--	--	1.281	1.481**
	(0.154)	(0.173)	(0.121)	(0.136)			(0.242)	(0.292)
Repub.*Twice	1.373	0.931	1.991**	1.424	--	--	1.218	0.596*
	(0.371)	(0.248)	(0.553)	(0.399)			(0.327)	(0.163)
Perceived Illus. Truth	2.457***	2.141***	1.983***	2.038***	1.487***	1.456***	1.178	1.182
	(0.237)	(0.240)	(0.232)	(0.210)	(0.161)	(0.151)	(0.158)	(0.142)
Trust in Media	1.321***	1.076	0.566***	0.709***	0.893	1.065	1.136*	0.934
	(0.103)	(0.083)	(0.045)	(0.054)	(0.069)	(0.084)	(0.086)	(0.073)
Low attent.	0.932	1.136	1.279	1.178	0.578***	0.619**	0.831	0.634**
	(0.182)	(0.215)	(0.243)	(0.223)	(0.110)	(0.121)	(0.156)	(0.122)
Observations	811	811	811	811	811	811	811	811

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Model: ordinal logistic regression

Beyond the data supporting the hypothesis, two other trends emerge of note. First, there was a partisan asymmetry in real headline belief. Being a Republican versus a Democrat decreased one's odds of higher belief in three of the four real headlines by 25.4-33.6% (Table 8, columns 6-8). This low-grade skepticism to the real headlines save "Police AI Use" is perhaps due in part to perceived partisan slant, as Republicans rated two of the three headlines as minutely favoring the Democrats on average ("Pasta", Patient Runs Marathon", Figure 4c-d).

Figure 4. Partisan Favorability Ratings (Non-Misinfo. Headlines)



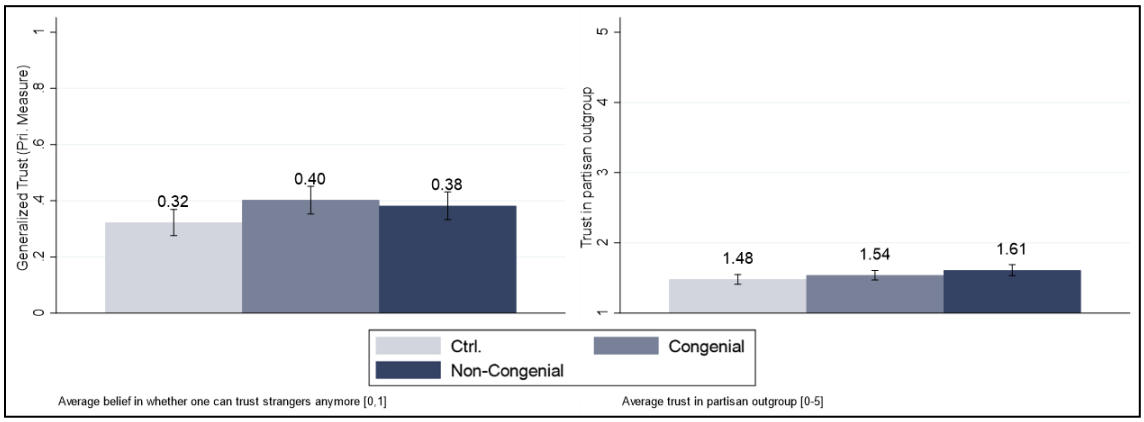
Supporting this inference is the fact that the one real headline that had no partisan difference in belief in Table 8 ("Police AI Use") was rated as neutral by both parties (Figure 4a). However, partisanship's impact on real headline belief seems outsized relative to the slight perceived partisan slant of the two headlines. Further, Republicans were skeptical of a headline that they rated as roughly partisan neutral ("Afghanistan Arrest", mean of 3.03). Given the model already incorporates some other major explanations (trust in media, lack of repeated exposure, perceived illusory truth, low attention), the decreased Republican belief could perhaps point to a partisan asymmetry in the perceived quality of news shared on social media.

The second trend of note is one that diverges from the difference-in-means analysis: repeated exposure at times impacted belief. Repeated exposure to the two pro-Democrat misinformation headlines had no impact on belief in the headlines (Table 8, columns 1-2), but repeated exposure to the two pro-Republican misinformation headlines decreased belief in the headlines (Table 8, columns 3-4, 32.4-38.2%). These results were at times conditioned on partisanship. Partisanship did not condition the effect of repeated exposure on belief in the "Child Trafficking" headline (1.424, $p < 0.1$): meaning, repeated exposure decreased odds of belief among both Republicans and Democrats. But, being a Republican exposed to the headlines twice increased odds of higher belief for the pro-Republican NYT coverup headline (99.1% increased odds, $p < 0.05$). Overall, this suggests that repeated exposure to misinformation, even over the course of a few minutes, can increase belief in the misinformation item among some, but the opposite can be just as true. The relationship is not straightforward.

Hypothesis 2a-b (GT, OPT): Limited Support for GT, No Support for OPT

Differences-in-means do not support Hypothesis 2a or 2b regarding lower GT and OPT. For GT, the CM group has a significantly *higher* mean GT than those in the control group rather than the hypothesized lower one (0.40 v. 0.32, $p = 0.05$). Further, the NCM group's mean GT of 0.38 was not lower than that of the control group. It indeed was slightly higher, though its difference of 0.06 only approached standard levels of significance ($p = 0.15$). Similarly counter-expectation are the OPT results. I expected the CM and NCM groups to have lower mean OPT than the control group, but the CM group does not significantly differ from the control group (1.54 v. 1.48, $p = 0.32$), and the NCM group has a statistically significant higher mean (1.610 v. 1.480, $p = 0.04$) (Figure 5).

Figure 5. Means: GT and OPT



Regression results using a logistic model amplify these counter-expectation differences-in-means (Table 9). For GT, CM is the only treatment with a significant impact, increasing rather than decreasing the odds of higher GT (43.2%, $p < 0.1$). For OPT, the inverse is true. Only NCM has treatment effects, unexpectedly increasing odds of higher OPT (41.4%, $p < 0.05$). These results are robust to a simpler model specification including only treatment group assignment (Appendix I). I included the three theoretically relevant confounders in this main analysis to rule out the possibility that result significance was due to demographic flukes in the randomization process. I have no reason to expect such flukes, but it seemed easier to check up-front and include the simpler specification in an appendix.

Table 9. Estimated Impact of Exposure on Trust (Odds Ratios)

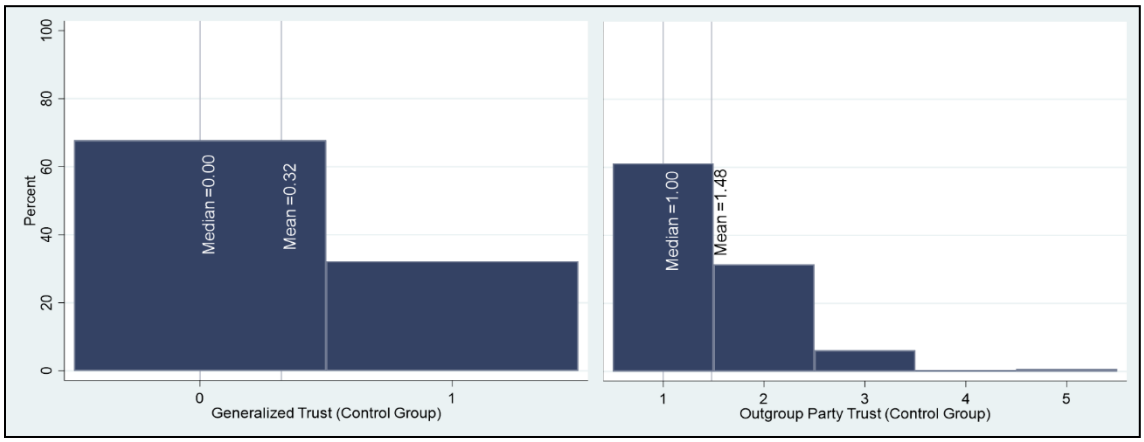
VARIABLES	(1) GT	(2) OPT
CM	1.432*	1.242
	(0.268)	(0.222)
NCM	1.327	1.414**
	(0.244)	(0.246)

Republican	0.728**	1.268
	(0.111)	(0.184)
Religious Attendance	0.981	1.154***
	(0.053)	(0.057)
Low attent.	0.918	1.139
	(0.206)	(0.246)
Observations	811	811

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1
GT: logistic model. OPT: ordinal logistic model

A clear accounting of these results is elusive. For GT, perhaps the CM group had their GT increased not by the CM per se, but by the study's assertion that randomly-selected social media users had shared the congenial content. Trust in strangers could go up if an assertedly random draw of social media posts contains largely congenial information. However, one would expect a corresponding GT decrease in the NCM group (when half the randomly-selected strangers offered non-congenial views). That decrease is not evident. A floor effect could be responsible for the lack of a decrease, as 67.77% of the control group selected the "0" GT value, leaving little room for mean downward GT movement (Figure 6). A floor effect is similarly plausible for OPT, as over 60% of control group respondents reported having no trust in their outgroup party whatsoever (value of 1). The mean increase in GT and OPT would seem to belie the possibility of a floor effect, as floor effects inherently censor decreases. But, if most of the respondents have no ability to show a trust decrease due to censorship, a spurious net increase could result, as increase is the only change that can be captured.

Figure 6. Left-Skewed Distribution of GT, OPT (Control Group)



Another possible explanation of these results is that there were unanticipated countervailing dynamics in the treatment groups. A theoretically-supported countervailing dynamic is differential impact by (dis)belief in the misinformation. Perhaps among those who believed CM and disbelieved NCM, GT and OPT went down, but this was washed out by an overall strong average positive effect of treatment exposure. I cannot test this explanation with high confidence due to the sample size, but an exploratory regression analysis offers limited support.

For the exploratory analysis, I took into account respondents' (dis)belief status. This meant I had to disaggregate the analysis into pro-Republican and pro-Democrat misinformation groups and further disaggregate the control group into a CM control group and NCM control group (i.e. by partisanship). This was for three reasons.

First, I needed to compare the (dis)belief-disaggregated treatment groups against the (dis)belief -disaggregated control group. The disaggregated control groups provide a baseline against which the disaggregated treatment groups can be causally compared, as the control group answered the trust questions before being exposed to the misinformation headlines in the belief question. Second, disaggregating the control group

by (dis)belief status meant I had to further sub-divide the sample by their (dis)belief in headline groups (pro-Republican/Democrat). This was because the control group had no assigned misinformation headlines. I could easily create a measure for whether the treatment groups (dis)believed their assigned misinformation; but I inherently could not do so for the control group, which had no assigned misinformation. To do so would be to compare apples (treated groups' [dis]belief in their assigned two headlines) to oranges (the control group's [dis]belief in some combination of all four misinformation headlines).

Third, as I had to disaggregate by headline group, I also had to disaggregate the control group by partisanship. The CM and NCM groups exposed to the pro-Democrat misinformation respectively all-Democrat and all-Republican, with the inverse being true for those exposed to pro-Republican misinformation. Thus, the control group also needed to be disaggregated by partisanship as well. To compare the results of a single-party CM and NCM group against a control group containing members of both parties would again be a case of apples and oranges.

The end result was I regressed the GT measure on CM/NCM treatment status, (dis)belief, and the interaction thereof among four samples using an ordinal logistic model (odds ratios). The first two samples permitted analysis of CM exposure, being composed of Democrats[Republicans] exposed to CM along with their untreated co-partisans in the control group. The latter two samples permitted analysis of NCM exposure, being composed of Democrats[Republicans] exposed to NCM along with their untreated control group co-partisans. *Belief*, used in the two CM samples, was dichotomous, with 1 indicating a rating of "somewhat likely" or "likely" for both CM

items and 0 indicating all other rating combinations. *Disbelief*, used in the two NCM samples, was similarly operationalized, with 1 indicating a rating of "somewhat unlikely" or "unlikely" for at least one NCM item and 0 indicating all other rating combinations. Due to the small sizes of these subsamples (the largest was $n = 283$), I included a fourth level of significance (80%) alongside the standard three levels of 90%, 95%, and 99%. The 80% level I will cautiously treat as potentially indicative of significance in a larger sample. As before, I include partisanship, religious service attendance, and low respondent attention in the specification, and the results are robust to a simpler specification excluding those "controls" (Appendix I).

The logistic models offer limited support for the inference that the expected negative impacts on trust can be found when the sample is disaggregated by party/headline group (Table 10). Starting with GT, belief in CM headlines did not show the expected negative impact among Democrats (column 1, 28% increase, $p < 0.1$), but did at a nearly-significant level among Republicans (column 2). Republicans who believed their CM treatment had lower odds of higher GT than Republicans in the control group (-53.9%, $p = 0.169$). NCM also shows a partisan asymmetry, with Democrats displaying the anticipated negative relationship at nearly-significant levels (column 4). Democrats who disbelieved the encountered NCM had lower odds of higher GT (-61.9%, $p = 0.192$) than did Democrats in the control group.

Table 10. Estimated Impact of Exposure on GT (Odds Ratios, Subsample)

	CM		NCM	
VARIABLES	(1) Dem.	(2) Repub.	(3) Dem.	(4) Repub.
CM treatment	1.654#	1.651#		
	(0.534)	(0.565)		
Believed both headlines	0.647	0.787		
	(0.248)	(0.308)		
CM treatment*Believed both headlines	1.028	0.461#		
	(0.536)	(0.259)		
Low attent.	0.696	1.306	1.132	0.997
	(0.283)	(0.522)	(0.531)	(0.557)
Religious Attendance	0.852#	1.052	0.939	1.029
	(0.093)	(0.087)	(0.106)	(0.088)
NCM treatment			4.308**	0.892
			(2.985)	(0.580)
Disbelieved 1+ headline			2.914*	1.309
			(1.690)	(0.680)
NCM treatment*Disbelieved 1+ headline			0.381#	1.019
			(0.282)	(0.723)
Constant	0.733	0.446**	0.213***	0.349**
	(0.219)	(0.147)	(0.124)	(0.184)
Observations	283	261	264	276

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1 # p<0.2
Model: logistic model

The results also highlight the countervailing, positive dynamic captured in the initial difference-in-means and regression analyses. Exposure to CM and NCM increased one's odds of higher GT for three out of the four subgroups at significant or nearly-significant levels (columns 1-3). CM treatment increased one's odds by 65.1-65.4% regardless of respondent partisanship at levels approaching statistical significance (p =

0.119-0.143), and NCM treatment increased odds among Democrats (column 3, 330.8%, $p < 0.05$).

Overall, the difference in direction between the CM/NCM treatment odds (positive) and the interaction odds of CM/NCM and (dis)belief (negative) suggests the positive impact of misinformation treatment on GT is an aggregate of countervailing underlying variables. CM and NCM exposure could decrease GT among those that (dis)believe it; however, even if a larger sample bore out that conclusion, the average decrease in GT odds from (dis)belief is less than the average increase in GT odds associated with CM/NCM treatment overall. There are more dynamics at work than simple (dis)belief, to the extent that the negative impact of (dis)belief is "washed out" by the positive impact of exposure overall. Further, the effect of (dis)belief was not consistent across parties, with Republicans having lower odds of high GT when they *believed* CM, and Democrats having lower odds of higher GT when they *disbelieved* NCM. These results were largely robust to the two alternate GT measures (Appendix I)

Unlike GT, the regression results for OPT still and entirely diverge from my expectations (Table 11).[69] The subsamples in columns 2-3 display the anticipated decrease in OPT among treated (dis)believers, echoing the GT analysis, but their parameters do not approach standard levels of significance ($p = 0.239$-0.391). Belief and disbelief matter by themselves with significant or near-significant decreased odds of outgroup party trust in all four columns (43.1-67.1%). But again, this effect is not an

[69] For these regressions, I had to collapse the OPT measure's two highest categories (4-5) into its next lowest category (3) because the upper categories had so few observations ($n = 10$) that they prevented me from running a Brant test. The operationalization change did not meaningfully change any of the parameters' magnitude, direction, or significance.

isolated causal effect, as the parameter includes control group members who expressed (dis)belief in the headlines after trust was measured. Finally, the CM and NCM treatment largely do not impact OPT in the aggregate. There is no significant relationship between treatment and OPT in columns 1, 3, or 4, with the only effect being found among Republicans exposed to CM (113.1% increased odds, $p < 0.5$). Beyond being theoretically unexpected (why would exposure information that favors one's ingroup and disfavors one's outgroup increase trust in said outgroup?), this result also unexpectedly diverges from the initial regression results, which showed that NCM and *not* CM had an aggregate positive and significant impact on OPT.

Table 11: Est. Impact of Exposure on Outgroup Party Trust (Odds Ratios, Subsample)

	CM		NCM	
	(1)	(2)	(3)	(4)
VARIABLES	Dem.	Repub.	Dem.	Repub.
CM treatment	0.893	2.131**		
	(0.282)	(0.689)		
Believed both headlines	0.507*	0.329***		
	(0.194)	(0.129)		
CM treatment*Believed both headlines	1.373	0.625		
	(0.717)	(0.343)		
Low attent.	1.482	0.721	2.707**	1.284
	(0.575)	(0.287)	(1.228)	(0.653)
Religious Attendance	1.339***	1.142*	1.245**	1.097
	(0.134)	(0.090)	(0.127)	(0.085)
NCM treatment			1.782	1.379
			(0.987)	(0.753)
Disbelieved 1+ headline			0.569#	0.474#
			(0.241)	(0.219)
NCM treatment*Disbelieved 1+ headline			0.483	1.195
			(0.298)	(0.722)
Observations	283	261	264	276

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1 # p<0.2
Model: ordinal logistic model

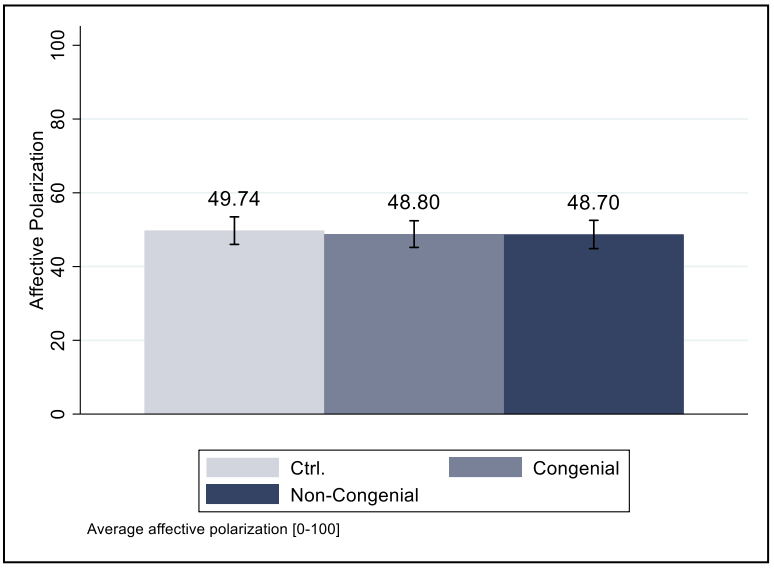
These unexpected null results could be due to floor effects. A slightly lower proportion of the sample had left-justified answers for OPT than GT, but it still was a majority (~60%). Average downward movement would have to be quite large for 40% of the sample to register an effect against the other 60%. One explanation for the unexpected results that I can rule out is the models' violation of the parallel regressions assumption. As with the previous belief analysis, the ordinal logistic analysis for OPT violates the assumption, but not in a way that is conclusion-altering (Appendix I).

In sum, these analyses presented limited support for Hypothesis 2a (GT) and no support for Hypothesis 2b (outgroup party trust). GT decreased among Republicans that believed their treatment CM and Democrats that disbelieved their treatment NCM, but this decrease averages out to no effect when taken in tandem with CM and NCM's overall positive effect on GT. These results approach but do not reach standard levels of significance. Unlike GT, OPT did not decrease among those in the treatment groups that believed CM or disbelieved NCM or in the treatment groups overall when disaggregated by partisanship.

Hypothesis 3 (Affective Polarization): Unsupported

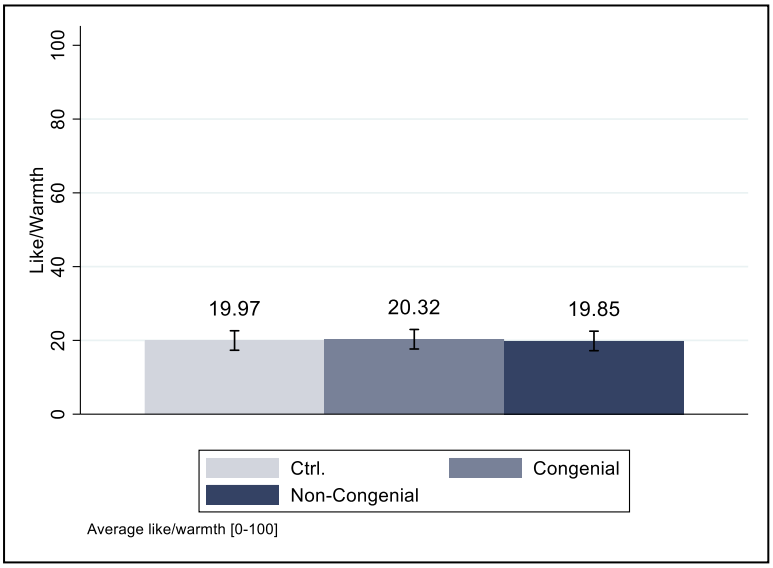
The results do not support Hypothesis 3. Starting with differences-in-means, the groups I expected to exhibit increased affective polarization show no difference from the control group, with their means covering a range of only 48.70-49.74 (Figure 7).

Figure 7. Means: Affective Polarization



These closely-grouped means do not mask underlying variation in the measure's two components of outgroup and ingroup like. As Figure 8 shows, outgroup party like does not significantly vary by treatment group, and necessarily ingroup party like does not vary either. Misinformation's non-impact on affective polarization and outgroup party

Figure 8. Means: Outgroup Like



like is further illustrated in Ordinary Least Squares regression results (Table 12).[70] Neither CM/NCM treatment has a significant effect on affective polarization nor outgroup party like. Indeed, neither treatment even approaches standard levels of significance, with the lowest p-value being only 0.708 (CM, column 2). These results are robust to a simpler model specification that excludes the confounders (Appendix J).

Table 12: Estimated Impact of Exposure on Affective Polarization (OLS)

VARIABLES	(1) Affect. Polariz.	(2) Party Outgroup Like
CM treatment	-0.547	0.739
	(2.786)	(1.973)
NCM treatment	-0.558	-0.154
	(2.713)	(1.921)
Republican	-5.541**	4.487***
	(2.192)	(1.552)
Low attent.	-2.215	-1.258
	(3.385)	(2.396)
Constant	58.037***	13.300***
	(3.774)	(2.672)
Observations	811	811
R-squared	0.009	0.011

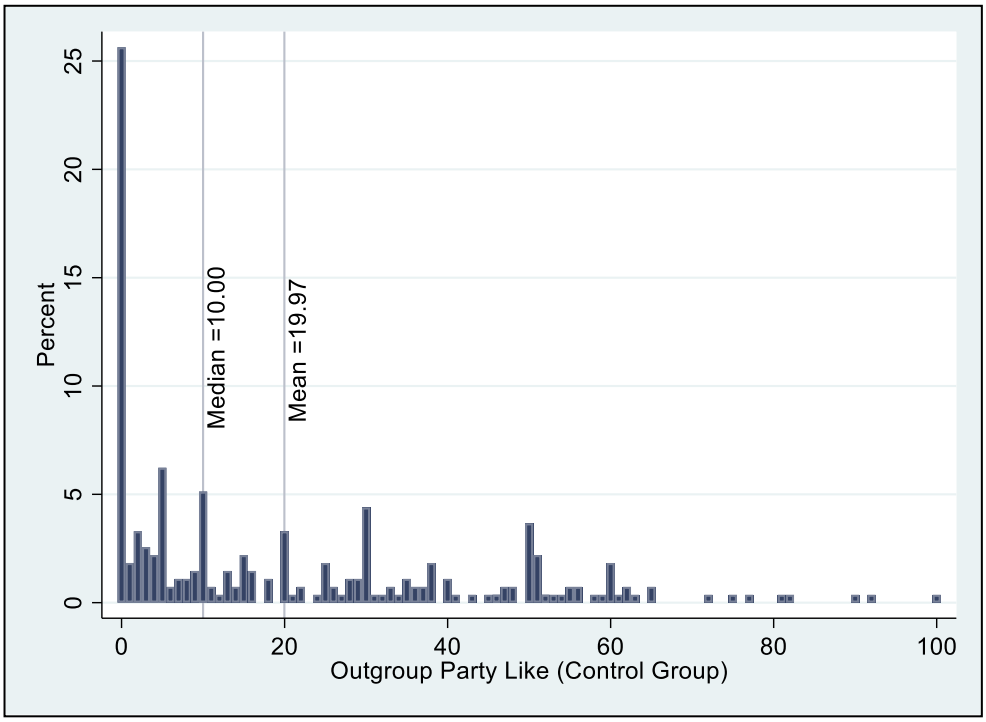
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Model: OLS

Respondent unresponsiveness to the treatment could in part be due to a floor effect with outgroup party like, which is low. In the control group, 25.64% of respondents assigned a 0 rating for their outgroup party, and half of them did not score the outgroup

[70] OLS was a suitable model for *affective polarization*, as it was an index variable of two component variables. An ordinal logistic regression would be more suited to ordinal *outgroup party like*; however, I used OLS to enable easier comparison of coefficient sizes.

party higher than 10 (Figure 9). The result is that many respondents had little room on the scale for meaningful more-negative appraisals of their outgroup party.

Figure 9. Distribution: Outgroup Party Like Ratings



Whether there is an intervening floor effect or no, the pattern of the treatment non-significance continues when the effect of (dis)belief is taken into account (Table 13). Neither belief nor disbelief mattered to treated respondents' affective polarization levels in any of the four subgroups. Following the pattern established in the trust analysis, the anticipated direction (positive) is found among treated Republicans who believed CM and treated Democrats who disbelieved NCM (1.197, 4.946), but with respective p-values of only 0.883 and 0.586. As with the previous regression analyses, these results are robust to a simpler specification (Appendix J). Further, the results' insignificance remains if outgroup like is regressed instead of affective polarization (Appendix J).

Table 13: Estimated Impact of Exposure on Affective Polarization (OLS, Subsample)

	Congenial		Non-Congenial	
	(1)	(2)	(3)	(4)
VARIABLES	Dem.	Repub.	Dem.	Repub.
CM treatment	2.199	-4.387		
	(4.375)	(5.240)		
Believed both headlines	12.018**	19.142***		
	(4.860)	(5.742)		
CM *Believed both headlines	-1.304	1.197		
	(6.829)	(8.157)		
Low attent.	-6.451	2.639	-12.216*	-1.713
	(5.438)	(6.084)	(6.670)	(8.363)
NCM treatment			-1.769	7.378
			(8.315)	(9.407)
Disbelieved 1+ headline			12.902**	27.467***
			(6.295)	(7.674)
NCM*Disbelieved 1+ headline			4.946	-10.069
			(9.059)	(10.345)
Constant	47.111***	40.620***	41.240***	24.760***
	(3.030)	(3.475)	(5.701)	(7.023)
Observations	283	261	264	276
R-squared	0.043	0.087	0.059	0.067

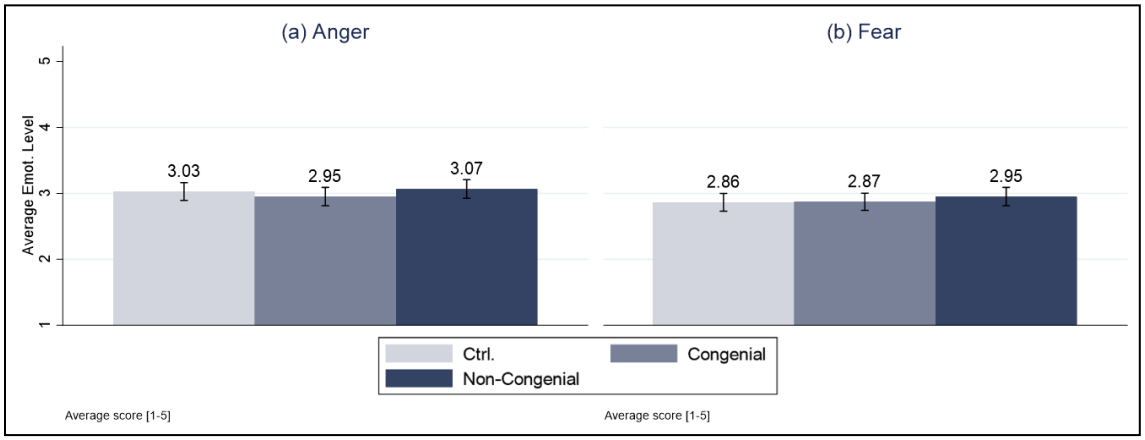
*** p<0.01, ** p<0.05, * p<0.1

Overall, this study offers no support to Hypothesis 3's expectation that exposure to partisan misinformation will increase affective polarization regardless of the misinformation's congeniality. Even taking into account differential effects on affective polarization by (dis)belief, the result remains null. This could be due to a noted floor effect; but as the data stands, the results only support the null hypothesis that partisan misinformation exposure has no effect on affective polarization.

Hypothesis 4 (Anger, Fear): Limited Support

Again, difference-in-means and regression results show no net impact of misinformation exposure on anger or fear, regardless of misinformation congeniality (Figure 10, Table 14). The means for anger and fear respectively range 2.95-3.07 and 2.86-2.95, demonstrating no statistically significant differences-in-means. Further, none of the parameters on CM or NCM approach standard levels of significance in Table 14's regressions (see Appendix K for robustness check).

Figure 10. Means: Anger, Fear



Factoring in (dis)belief changes the null finding, but in a very limited way (Table 15). I expected CM belief and NCM disbelief to increase anger. This result occurs only once, and only approaches standard levels of significance (column 3). Democrats in the NCM group who disbelieved at least one of the headlines had higher anger than Democrats in the control group who disbelieved (0.507 increased odds, $p = 0.158$). The pattern is not evidenced for disbelieving Republicans in the NCM group (column 4), nor

Table 14: Estimated Impact of Exposure on Emotion (OLS)

VARIABLES	(1) Angry	(2) Afraid
CM	-0.086	-0.025
	(0.103)	(0.100)
NCM	0.026	0.087
	(0.101)	(0.097)
Republican	0.139*	-0.327***
	(0.081)	(0.079)
Low attent.	0.051	0.112
	(0.126)	(0.121)
Constant	2.820***	3.351***
	(0.140)	(0.135)
Observations	811	811
R-squared	0.006	0.023

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Model: OLS

for believing Republicans in the CM group (column 2). Puzzlingly, the expected pattern is not only nonexistent for believing Democrats in the CM group, but inverted (column 1). Their anger decreases relative to believing Democrats in the control group at levels approaching statistical significance (-0.424%, p = 0.110). CM belief made them less angry. These conclusions are robust to a simpler model specification omitting low attention (Appendix K).

Table 15. Estimated Impact of Exposure on Anger (OLS, Subsample)

	CM		NCM	
VARIABLES	(1) Dem.	(2) Repub.	(3) Dem.	(4) Repub.
CM	0.013	-0.009		
	(0.170)	(0.188)		
Believed both headlines	0.365*	0.961***		

	(0.189)	(0.206)		
CM*Believed both headlines	-0.424#	-0.311		
	(0.265)	(0.292)		
Low attent.	0.004	0.421*	0.030	-0.426#
	(0.211)	(0.218)	(0.264)	(0.295)
NCM			-0.404	0.126
			(0.329)	(0.332)
Disbelieved 1+ headline			-0.188	0.537**
			(0.249)	(0.271)
NCM* Disbelieved 1+ headline			0.507#	-0.026
			(0.358)	(0.365)
Constant	2.851***	2.704***	3.147***	2.625***
	(0.118)	(0.124)	(0.226)	(0.248)
Observations	283	261	264	276
R-squared	0.018	0.124	0.008	0.039

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1 # p<0.2

Moving to a regression analysis of fear, the hypothesis is again only weakly supported when (dis)belief is taken into account (Table 16). Of the four groups, only Democrats who disbelieved NCM have a higher level of fear relative to their equivalent co-partisans in the control group at levels that approach statistical significance (column 3, 0.467, p = 0.203). Neither Republicans who disbelieved NCM (column 4) nor either party who believed CM (columns 1-2) show the expected pattern. Interestingly, the parameter on the *CM treatment*Belief* interaction is in the expected direction in column 2, as with the anger analysis, affective polarization analysis, and GT analysis. But, it does not meaningfully approach standard levels of statistical significance (0.290, p = 0.315). Nor did it do so in a robustness check employing a simpler specification in Appendix K.

Table 16. Estimated Impact of Exposure on Fear (OLS, Subsample)

	CM		NCM	
	(1)	(2)	(3)	(4)
VARIABLES	Dem.	Repub.	Dem.	Repub.
CM	-0.164	-0.049		
	(0.167)	(0.185)		
Believed both headlines	0.193	0.178		
	(0.186)	(0.203)		
CM*Believed both headlines	-0.029	0.290		
	(0.261)	(0.289)		
Low attent.	0.173	0.311#	-0.023	-0.203
	(0.208)	(0.215)	(0.269)	(0.282)
NCM			-0.256	0.180
			(0.335)	(0.318)
Disbelieved 1+ headline			-0.185	0.177
			(0.254)	(0.259)
NCM* Disbelieved 1+ headline			0.467#	-0.086
			(0.365)	(0.349)
Constant	2.992***	2.581***	3.220***	2.509***
	(0.116)	(0.123)	(0.230)	(0.237)
Observations	283	261	264	276
R-squared	0.013	0.037	0.009	0.006

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

These results could evidence the impacts of NIOC among Democrats, but otherwise does not support Hypotheses 4a-b. Democrat anger and fear increases when they disbelieve encountered NCM. However, these increases only approach standard levels of significance. The non-relationship between (dis)belief and anger/fear for the other three subgroups supports the null hypothesis.

Hypothesis 5a-b (TPE): Limited Support

As with the previous outcomes, neither in- nor outgroup TPE varies by treatment group. The TPE means regarding the ingroup party have a tight range of 0.93-0.97, and the mean range regarding the outgroup party is similarly tight (1.71-1.79) (Figure 11). Simple CM exposure did not increase ingroup party TPE, nor did simple NCM exposure. Further, ordinal logistic regressions show no significant impact of CM or NCM treatment on any of the sorts of TPE relative to the control group (Table 17). This result is robust to an alternate specification that omits the low attention "control" (Appendix L). The result is also not wholly surprising, as the control group was exposed to the misinformation in the headline-belief question prior to being asked about their perceptions of misinformation's influence.

Figure 11. Means: TPE for Misinfo. Influence on Ingroup/Outgroup Party

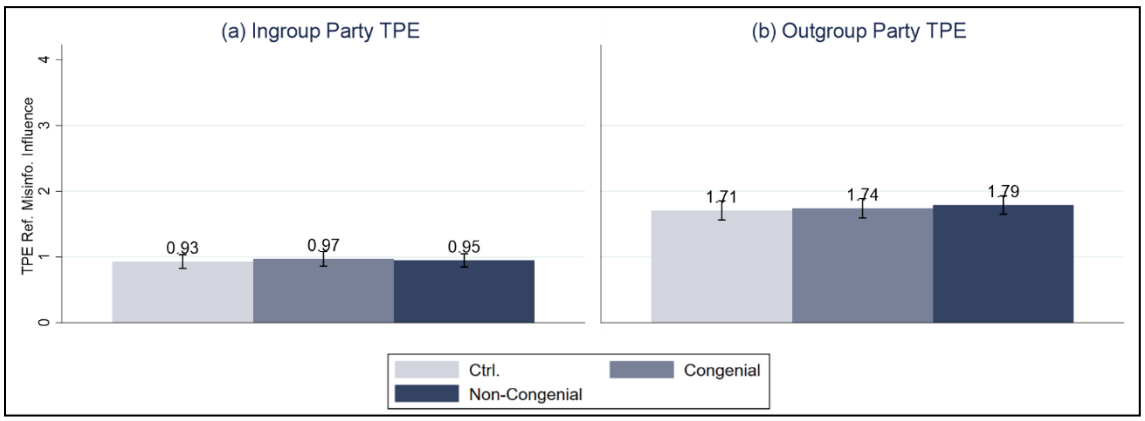


Table 17. Estimated Impact of Exposure on TPE (Odds Ratios)

VARIABLES	(1) Ingroup Party	(2) Outgroup Party
CM	1.025	1.083
	(0.168)	(0.170)
NCM	1.016	1.088
	(0.161)	(0.168)
Low attent.	1.194	0.843

	(0.236)	(0.168)
Observations	811	811

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

However, the control group's passive misinformation treatment meaningfully differed from that of the treatment groups in that it lacked social context. Control group respondents just encountered headlines; treatment group respondents were told randomly-selected social media users had shared the headlines and identified them as worthwhile. The importance of this difference is perhaps evidenced by the party/headline group-disaggregated regression results, as they provide limited support for the hypotheses (Table 18). The limited support is found in the Democrat subsamples, for whom CM disbelief increases TPE regarding the associated partisan group (ingroup) and NCM disbelief does the same regarding its associated party (outgroup) (240.4-244.2% greater odds, $p < 0.05$). The Republican subsamples, however, do not demonstrate the expected pattern. Neither disbelief in CM nor NCM impacted ingroup nor outgroup TPE, respectively (-0.09 to 85.1%, $p = 0.270$- 0.985). These results are robust to the simpler specification (Appendix L).

Table 18. Estimated Impact of Exposure on TPE (Odds Ratios, Subsample)

	Ingroup TPE (CM)		Outgroup TPE (NCM)	
	(1)	(2)	(3)	(4)
VARIABLES	Dem.	Repub.	Dem.	Repub.
Treatment	0.654	1.047	0.359**	0.766
	(0.187)	(0.322)	(0.182)	(0.387)
Disbelieved 1+ headline	0.341***	1.155	1.111	1.307
	(0.113)	(0.381)	(0.423)	(0.560)
Treatment*Disbelieved	3.422***	0.991	3.404**	1.851
	(1.552)	(0.468)	(1.897)	(1.033)
Low attent.	1.100	1.107	1.328	0.807
	(0.401)	(0.389)	(0.565)	(0.356)

Observations	283	261	264	276

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In sum, these results suggest that misinformation exposure can impact party-specific TPE, but that the relationship is conditioned on partisanship. Republicans that disbelieved misinformation showed no TPE change regarding the misinformation's associated parties, while that of Democrats increased. This provides evidence that partisan misinformation encounters can result in updates to related group assessments. Taking into account TPE's theoretical role in the ripple effects of NIOC, this also suggests that NCM's ripple effects would be stronger among Democrats than Republicans, as their perception of NCM's influence on inferred Republican audiences is higher. The other hypotheses' results support this finding, as all the significant and near-significant impacts of NCM disbelief on the other outcomes occurred among Democrats.

Hypothesis 6 (Opt-In/Out)

A difference-in-means analysis of those that opted-out of data tracking suggests that those who opt-out do not meaningfully differ from those that opt-in. These findings hold even when disaggregated by partisanship.

First, across the entire sample, there were many differences, but few of note (Table 19). Those who opted-out of the tracking were slightly older than those who opted-in (45 v. 42 years). They also were somewhat more conservative (3.038 v. 2.734 on a scale of 1-5) and somewhat more Republican (52.6% v. 43.1). They also were less likely to use social media. They reported slightly lower frequency of social media usage, and fewer of them had accounts on Facebook and TikTok than those who opted in (79.2% v. 94.2% and 38.4% v. 53.3%, respectively). Accompanying these lower rates of

use were lower rates of news-related social media posts for all sorts of news save the catch-all "other news."

Table 19. Difference-in-Means: Opt-In/Out Status

	Opt-out	Opt-in	Diff	P-Stat
Female	0.521	0.467	-0.054	0.14
Education	4.233	4.129	-0.105	0.30
Age	**45.034**	**41.971**	**-3.063**	**0.01*** **
Rural	0.178	0.191	0.013	0.64
Suburban	0.549	0.533	-0.016	0.67
Urban	0.273	0.276	0.002	0.95
Religious Attendance	2.025	2.015	-0.010	0.92
Pol. Ideology (L-C)	**3.038**	**2.734**	**-0.303**	**0.00*** **
Republican	**1.526**	**1.431**	**-0.095**	**0.01*** **
Partisan Strength	3.136	3.118	-0.018	0.78
SM: Have Unfollowed/Muted	0.817	0.820	0.003	0.92
SM: Freq. of Use	**5.842**	**6.077**	**0.235**	**0.01*** **
SM: Reddit	0.703	0.754	0.051	0.12
SM: Facebook	**0.792**	**0.842**	**0.050**	**0.08***
SM: TikTok	**0.384**	**0.533**	**0.149**	**0.00*** **
SM: YouTube	0.924	0.919	-0.005	0.78
SM: Instagram	0.681	0.728	0.047	0.17
SM: No SM Usage	0.007	0.000	-0.007	0.18
SM News: Sci/Tech	**0.559**	**0.699**	**0.140**	**0.00*** **
SM News: Polit.	**0.498**	**0.603**	**0.105**	**0.00*** **
SM News: Sports	**0.409**	**0.526**	**0.116**	**0.00*** **
SM News: Busin.	**0.369**	**0.456**	**0.087**	**0.02** **
SM News: Celeb.	**0.344**	**0.430**	**0.086**	**0.01*** **
SM News: Other	0.320	0.298	-0.023	0.50
SM News: None	**0.195**	**0.132**	**-0.062**	**0.03** **
Trust: News Media	1.992	2.029	0.038	0.58
Trust: SM Fact-Checkers	**2.289**	**2.456**	**0.167**	**0.04** **
Trust: Experts	3.295	3.360	0.065	0.46
Trust: Gov't	2.428	2.386	-0.042	0.45
GT	**0.351**	**0.412**	**0.061**	**0.08***
Alt: GT (Standard)	**0.351**	**0.426**	**0.076**	**0.03** **
Alt: GT (GSS Index)	0.426	0.471	0.045	0.15
Trust: Ingroup Party	2.943	2.939	-0.004	0.95
Trust: Outgroup Party	1.542	1.545	0.003	0.95
Pol. Knowledge (Index)	0.814	0.815	0.001	0.95
Pol. Interest: Campaign	2.203	2.254	0.051	0.31
Pol. Interest: General	3.398	3.441	0.044	0.58
Affect. Polariz.	48.646	50.098	1.452	0.54
TPE: General	**1.450**	**1.313**	**-0.137**	**0.09***

TPE: Ingroup	0.963	0.919	-0.044	0.58
TPE: Outgroup	1.791	1.642	-0.149	0.18
True: Burn Docs	2.955	3.044	0.089	0.38
True: TX Textbooks	2.799	2.813	0.014	0.89
True: NYT Coverup	3.094	3.059	-0.035	0.72
True: Child Trafficking	2.723	2.581	-0.142	0.16
True: Police AI Use	**3.903**	**3.750**	**-0.153**	**0.04****
True: Afg. Activist Arrest	**4.242**	**4.103**	**-0.139**	**0.05****
True: Flat Packed Pasta	2.955	3.081	0.126	0.14
True: Patient Runs Marathon	3.847	3.875	0.028	0.72
Click: Burn Docs	0.284	0.312	0.027	0.63
Click: TX Textbooks	0.322	0.398	0.076	0.20
Click: NYT Coverup	0.371	0.355	-0.016	0.80
Click: Child Trafficking	0.326	0.344	0.018	0.76
Click: Police AI Use	**0.440**	**0.544**	**0.105**	**0.00*****
Click: Afg. Activist Arrest	0.281	0.390	0.109	0.08
Click: Flat Packed Pasta	0.265	0.299	0.033	0.58
Click: Patient Runs Marathon	0.222	0.267	0.045	0.41
Observations	868			

Statistically significant differences-in-means are bolded

Those who opted-out also reported some trust levels that were slightly lower than those in the opt-in group. Their trust in fact-checkers on social media was lower (2.289 v. 2.456 on a scale of 1-5), as was their GT by two measures (0.351 v. 0.412/0.426 on a scale of 1-5). However, the two groups did not differ significantly with regard to trust in media, government, experts, government, ingroup party, or outgroup party. They also had no significant differences in political knowledge or political interest (both campaigns and in general).

Noteworthily, those who opted-out did not differ from those who opted-in with regard to either misinformation belief or self-reported interest in reading any of the misinformation articles. This result was not conditioned on respondent partisanship (Appendix M). Democrats who opted-out were no less likely to believe the pro-Democrat misinformation than Democrats who opted-in, and the same was true for Republicans

regarding pro-Republican misinformation. This is not to say there were no differences by partisanship. Democrats who opted-out had less campaign interest than those that opted in (2.209 v. 2.329, $p = 0.08$) and were more often female (58.1% v. 48.6, $p = 0.07$). Republicans who opted-out tended to be older than those that opted in (48.946 v. 45.377, $p = 0.05$). They also had higher TPE regarding misinformation's influence on the political opinions of people in general, their ingroup party, and their outgroup party (0.983-1.401 v. 0.774-1.142).

However, none of these partisan differences suggest a result-altering selection bias in active exposure studies. A handful of differences exist between those that opt-in and -out of the requisite data tracking, but within and across parties, respondents who opted out were no more likely to believe or express interest in reading CM than those that opted-in. As perhaps suggested by the lack of difference in partisanship strength; trust in media, government, and experts; and affective polarization, the lack of difference in misinformation belief and engagement may arise from a lack of difference in predisposing social and behavioral factors.

Discussion

Ripple Effects: Trust and Emotions

The results suggest that misinformation has socio-emotional ripple effects, but that these effects largely will not be evident unless conditioned on (dis)belief, partisanship, and misinformation congeniality. However, due to the resultant major disaggregation and the potential of floor effects on the trust outcomes and affective polarization, my results are only suggestive. Many of the effects of note (e.g., decreased GT among Republicans that believed CM and Democrats that disbelieved NCM; increased anger/fear among

Democrats that disbelieved NCM) only approach standard levels of statistical significance.

The results point to four broader implications. First, and as theorized, misinformation can impact those that disbelieve it, or, at least, Democrats that disbelieve it. The effects were not large enough to be evident in the aggregate, as they often concentrated among the subgroup of Democrats in the NCM group, and they also were not evident for OPT or affective polarization. But they were there for GT, anger, fear, and in/outgroup TPE. Disbelief in NCM decreased Democrats' GT and increased their fear, anger, and outgroup TPE at levels that approached standard statistical significance.

Second and relatedly, misinformation does not always impact those that disbelieve it. NCM disbelief had no impact on OPT and affective polarization for either party, and had null effects on Republicans' GT, anger, fear, and TPE (outgroup). The null results of NCM disbelief (and CM belief) on OPT and affective polarization could suggest that people do not update partisan group assessments based on the inferred partisanship of misinformation sharers or likely audiences. That would be theoretically unexpected but statistically plausible, as the subgroup of treated respondents who were not told the article sharers' partisanship reported only weak partisan inferences about them (Appendix N). The two pro-Democrat headlines were rated as being only marginally more likely to have been shared by a Democrat (2.62-2.97 on a scale of 1-5), regardless of respondent partisanship. Similarly, the two pro-Republican headlines were only rated marginally more likely to have been shared by a Republican (3.08-3.28), with Republicans actually making no partisan inferences about the NYT coverup article sharer (2.98). Assuming these self-reported partisan assessments are accurate, respondents do

not seem to have strongly inferred sharer partisanship, which in turn would short-circuit the theorized updates to party assessments.

However, party-based reasoning is evident in the TPE results for Democrats at standard levels of statistical significance. Partisan misinformation exposure ought not have impacted partisan TPE at all if inferred partisanship and resultant party-based assessments were nonexistent. Further, sharer partisanship seems likely to have been inferred because the addition or omission of sharer partisanship did not significantly alter results within the treatment groups (Appendix O). Of course, this result equivalence could suggest that sharer partisanship did not matter at all. But, given the impacts to Democrats' TPE and the contemporary context in which partisan group identity hardly lacks salience, it is more likely that sharer partisanship does matter and some other countervailing dynamic accounts for the null results. A plausible underlying dynamic is the previously-discussed floor effects. In the control group, the distributions of both OPT and outgroup like are left-skewed. This would censor any impacts on those participants perhaps most likely to make negative, group-based inferences regarding their outgroup—those who already think poorly of their outgroup party.

Thirdly and unexpectedly, misinformation's ripple-effects on those that believed it were remarkably limited. Belief in CM decreased GT among Republicans at a statistically significant level, but had no effect on OPT, affective polarization, or fear for either party. Further, the effect was at one point counterintuitive, as CM belief *decreased* anger among Democrats. Various explanations could contribute to the null effects. One potential explanation is that CM belief has weaker impacts than NCM, as disbelief in NCM constitutes NIOC. It would thus be unsurprising that the CM results are relatively

weaker than the NCM results, though their weakness to the point of insignificance remains unexpected.

Another potential explanation of CM belief's null effects is that it simply does not have a strong impact on the measured outcomes. In the context of political polarization, sophisticates most able to link CM content with related political assessments likely already possess strong general opinions about their outgroup party and feelings about how things are going in the country, and a limited exposure to CM may be too weak to alter those opinions and feelings. Another potential, related explanation is floor effects. To paraphrase a cognitive interviewee who piloted the survey, "You can't go any lower [in your appraisals of how it's going in the country] if you already have such a low opinion. [The misinformation's claims were] just more of the same old thing."

The potential of floor effects impacting the CM and NCM results suggests that future research into misinformation's ripple effects consider non-survey based measures. Some of this study's survey measures could have benefited from including more points in their scales (e.g., the dichotomous trust variables), but even measures that had a 100-point scale were strongly left-justified. Two potentially useful alternate approaches are interviews or use of micro-expression-capturing software like Emotient. Both methods could provide researchers access to emotion or trust impacts the participants may register but cannot communicate via survey instruments. Another potentially useful measure could include behaviors in trust games or other cooperative scenarios. Researchers could vary participants' awareness about a potential partner's sharing of or endorsement of misinformation and examine their trust/emotional disposition towards that partner. The potential of floor effects also suggests that future research examine misinformation's

potential to maintain high levels of negative emotion or mistrust. If misinformation exposure causes emotion or mistrust censored by floor effects, the censored outcomes could still register a "shadow" effect in terms of their temporal longevity relative to those not exposed to misinformation.

The single unanticipated significant interaction effect—CM belief decreasing Democrats' anger—could be due to broader contextual reasoning beyond the misinformation's content. Democrats' anger may have decreased not because they believed the CM, but because they viewed the believed stories' publication as a positive development. Believing that a major outgroup figure did something wrong could be angering, but the wrong being unmasked and shared by strangers on social media could have an opposite and even larger effect.

Fourth, there are partisan asymmetries in how Democrats and Republicans encounter partisan misinformation. Generally, NCM disbelief was most impactful among Democrats, and CM belief was most impactful among Republicans, though to markedly lesser extent. This result could be because Democrats are more responsive to NIOC than Republicans, as the TPE analysis would suggest, or that Democrats are more likely to make stronger group-based assessments. This result also could be because Republicans feel marginalized in the media ecosystem, and so encounters with congenial, perceivedly accurate news was more impactful to them. Opposingly, the partisan asymmetry could result from unidentified differences in the treatment strength between the pro-Democrat and pro-Republican misinformation (see limitations section).

Opt-In/Out Imbalances

The study has one further set of implications related to the opt-in/out analysis. While there was no selection bias in the opt-in/out samples that suggests extant research underestimates active misinformation exposure and its impact, there were several imbalances between the two samples that could be relevant to future research. Those who opted-out of tracking had marginally lower self-reported frequency of social media use. This imbalance was echoed for Facebook, TikTok, and YouTube, with all opt-out respondents reporting lower rates of Facebook and TikTok use and Democrat opt-out respondents also reporting lower YouTube use. An implication is that frequency of active exposure to misinformation on social media—especially on those three platforms—may be *higher* among those that opt-in than those that opt-out, due to simple volume of use. Relatedly, those who opted-out reported moderately lower rates of posting news on social media, to include political news. This suggests researchers may similarly overestimate the prevalence of political misinformation sharing if they only examine data from participants who opt-in to tracking.

Another imbalance, this one counteracting the two previous ones, is in the area of trust in fact-checkers. Those who opted-out reported moderately lower trust in fact-checkers on social media than those who opted-in. This suggests that research into fact check effectiveness estimates may be overestimated relative to the general social media population if study samples are composed of people who opted-in to tracking. As the selection bias for fact check trust was most pronounced for pure independents, though, this will be of less concern for analyses focused solely on partisans/partisan leaners. Finally, turning to partisan analyses: Democrats who opt-out reported moderately lower interest in this year's political campaigns, and opt-out Republicans reported moderately

higher TPE regarding misinformation (general, ingroup, and outgroup). Thus, active exposure studies could overestimate active exposure rates among Democrats, assuming campaign interest corresponds with active exposure to campaign-related misinformation, and among Republicans, assuming TPE regarding misinformation actually derives from misinformation encounters.

Limitations

The study has six main limitations. First and foremost, the (dis)belief analysis may be subject to post-treatment bias. Best practice dictates moderators be measured pre-treatment, but (dis)belief in a given misinformation item necessarily could not be measured until after respondents had seen the misinformation item. It therefore was measured post-treatment, which introduces the possibility that disbelief could have a confounding relationship with the treatment via an omitted variable (Montgomery, Nyhan, and Torres 2018; Sheagley and Clifford 2023).

I can rule out general political anger or fear out as confounders, as treatment had no impact on them. However, there are innumerable factors I cannot rule out. For instance: outgroup like. Partisan misinformation exposure could perhaps have decreased outgroup like in ways censored by this study's floor effects. Outgroup like, in turn, could have increased treated respondents' propensity to express stronger (dis)belief than those in the control group. Here, the dichotomous operationalization of (dis)belief reduces risk of bias, but not entirely. Moderate impacts have much lower chance of registering in a dichotomous variable, which is inherently less sensitive to variation. However, the risk for bias introduction is higher at the (dis)belief variables' 0/1 seam. The move from 0 to 1

is a move from “equally likely and unlikely” to be true and “somewhat (un)likely” to be true: a modest impact that an omitted cofounder could effect.

Second, setting aside the posttreatment bias concern, the results regarding (dis)belief were conducted with relatively small sample sizes. The four partisan/headline subgroups ranged in size from 261-283 people, but the skewed distribution of (dis)belief within the subgroups meant they differed in size by 22.6% (CM, Republicans) to 63.76% (NCM, Democrats) (Table 20). This means that the key interaction effect of (dis)belief on misinformation exposure may have been underestimated due to the small size of the (dis)belief comparison group, particularly for the two NCM subgroups.

Table 20. Distribution of (Dis)belief by Subgroup

	(Dis)believed	**Didn’t (Dis)believe**
CM: Dem.	109 (38.52%)	174 (61.84%)
CM: Rep.	101 (38.70%)	160 (61.30%)
NCM: Dem.	215 (81.44%)	49 (18.56%)
NCM: Rep.	226 (81.88%)	50 (18.12%)

Third, I cannot rule out that the pro-Republican and pro-Democrat headline treatments varied in strength. The two headline sets were roughly balanced in terms of relative believability and partisan favorability, but they possibly varied on some omitted variable. There also could have been an asymmetrical interaction effect between the treatment’s two misinformation items and the two real items. In either case, the partisan asymmetries by (dis)belief across the various outcomes would be due not to some social-psychological difference between Republicans and Democrats, but treatment differences. For example, NCM disbelief only increased anger and fear among Democrats. That could perhaps be because one or both of the pro-Republican headlines was more outrageous or fear-provoking to outgroup partisans than the pro-Democrat headlines: a possibility I did

not test for. Overall, though, the potential of treatment differences would not call into question my main conclusion that misinformation can have ripple effects conditioned on (dis)belief status.

Fourth, the results' external validity is limited to circumstances in which content sharers are not well-known to the passively-exposed individual. The study provided little information about the purported article sharers. Respondents were told the sharers were randomly-selected and endorsed the shared articles as being "interesting, important, or useful to know," and half of the treated respondents were also informed of the sharers' partisanship. This low-context exposure method is more akin to encountering strangers' posts on social media than the posts of known figures, acquaintances, or friends. Absent is the exposed individual's running estimate of trust in a given person as a communicator. Moreover, absent are the heuristics respondents could use in real life to rapidly appraise a stranger on social media (e.g., name, race, age, etc.). This does not discount my results but does suggest that more complicated information exposure environments be included in future research.

Fifth, the research design did not fully approximate passive exposure in a key way: study participants were likely more attentive to the headlines than the average social media user. CloudResearch Connect participants can have their pay denied if they do not perform well on attention checks, and so they tend to be attentive. Further, the exposure environment was not complex, as it lacked a constant stream of information competing for respondents' attention. Combined, these factors likely made study participants much more attentive to the passively-exposed misinformation than the average internet user.

Sixth, the opt-in/out measure had limited external validity. In the study, participants were simply asked their willingness to grant access to their anonymized browser history and/or social media activity. In reality, the process to opt-in to researcher tracking is multi-step, and typically involves downloading a browser extension and likely involves reading a detailed agreement with the polling firm or researchers. The study's measure of opt-in/out preference likely captures participant's initial openness to participation in such research but cannot capture the cognitive and time commitments required to actually grant researcher access to one's data. It is possible selection bias emerges at the later steps in the real-world process not emulated in this study, as respondents drop out of the sample who are uncomfortable with downloading a browser extension or become uncomfortable after reading agreement documents associated with the tracking.

Conclusion

In January 2020, Guess et al. published a paper entitled "'Fake News' May Have Limited Effects Beyond Increasing Beliefs in False Claims." My results broadly echo their conclusion. I argued a theoretical case that misinformation's effects are potentially underestimated, and the strongest, clearest effects I found was on belief, conditioned on partisan congeniality. The analysis also showed some limited ripple effects. Passive misinformation exposure had no effect on affective polarization and OPT, and only conditional and occasionally countervailing effects on GT, anger, and fear that only approached standard levels of statistical significance. Effects were largely concentrated among Democrats who disbelieved NCM, though Republicans who believed CM registered one significant effect in the form of decreased GT. Overall, these results

suggest potential impacts to political participation via anger/fear and social cohesion via GT, but not strong, neatly consistent ones. Misinformation's effects are wider than belief, but these ripple effects are highly conditioned.

Further, I found that the standard method to access active misinformation exposure data does not evince theoretically-relevant selection bias. The few differences that existed between those that opted-in to tracking and those that opted-out did not include belief in any of the misinformation items or self-reported interest in reading the misinformation items (regardless of partisanship). They also did not include affective polarization, trust in media, trust in government, or trust in outgroup party. Studies that examine active misinformation exposure and its outcomes do not have the deck stacked by inadvertent selection bias among those that opt-in to internet tracking.

Despite limiting factors such as floor effects and small subsample sizes, this study's results suggest that misinformation's effects are not "fake news." Partisan misinformation can be highly plausible to its respective ingroup audiences, with passive exposure causing belief. Further, misinformation's effects may be underestimated with regard to ripple effects, most notably in this study: backfire effects among Democrats. However, the effects will not be evident in the aggregate due to its conditionalities on (dis)belief, partisanship, and misinformation congeniality. On average, misinformation may not matter much beyond belief. But beneath the surface, it may increase participation-linked emotions and deleteriously decrease cooperation-linked trust outcomes.

CHAPTER 4

POLARIZATION: HOW IMPACTFUL IS IT ON COMMITMENT PROBLEMS?

Introduction

What are the international implications of polarization? Scholars, largely focused on the United States, have explored polarization's genesis and domestic implications (Druckman et al. 2021b; Iyengar et al. 2019; Rogowski and Sutherland 2016). They even have debated its existence (Abramowitz 2010; Fiorina, Abrams, and Pope 2006). Less explored are its implications for international politics, particularly if the polarization in question is not that of the United States. What does polarization mean for politics beyond "the water's edge?"

Polarization literature offers some answers with mixed applicability beyond the United States (e.g., polarization's impact on support for liberal internationalism) (Chaudoin, Milner, and Tingley 2010; Kupchan and Trubowitz 2007; Myrick 2022, 6–7). Another answer, though, touches on a fundamental calculus of international bargaining: defection chances. Schultz (2017), Friedrichs (2022a), and Myrick (2022) all variously argue that U.S. domestic polarization increases concern of its defection. Myrick even finds evidence that U.S. policy polarization decreases U.K. public confidence in U.S. reliability and preference for future cooperation.

Several open questions remain. Is perceived unreliability the mediator from polarization to cooperation preference, or is an alternate explanation like perceived co-ethnicity in play? If polarization does increase perceived unreliability, why does it do so? Is it because of gridlock and the risk that government changeover means policy

changeover? Further, are Myrick's findings generalizable to polarization beyond the U.S. case and to cooperation outcomes more concrete than general cooperation willingness?

Building on Myrick (2022), this study argues that polarization decreases cooperation preference via perceived unreliability. Bargaining literature suggests concern of voluntary rather than involuntary defection drives this relationship, as involuntary defection is due more to negotiator misjudgment than gridlock. Two pathways flow from polarization to voluntary defection chances. First is risk of policy changeover due to government changeover. If, as Myrick argues, competing coalitions have widely diverging platforms, then change in government will likely result in change in policy. Second is a reason I argue in addition to Myrick's: weakened leader accountability. Increasing group distinctiveness increases coalition influence on uncrystallized and moderate strength preferences, which increases coincidence between supporter and leader policy positions. These real or even just potential impacts of polarization theoretically matter because elites and publics learn from the domestic situation in other states when making foreign policy assessments.

Focusing particularly on public perceptions, a survey experiment (n = 750) and its pilot study of undergraduate students (n = 300) use an international politics scenario to test my argument that (1) perception of polarization decreases cooperation preference and (2) perceived unreliability accounts for a portion of that decrease. It also tests a causal logic hypothesis that polarization perception (3a) increases perceived probability of defection in case of government changeover and (3b) decreases perceived probability that a leader's supporters will hold them accountable for defection. Finally, it also tests whether (4) Myrick's finding that affective polarization had no impact on unreliability

perception/general cooperation preference was an artifact of operationalization. The operationalization emphasized affective polarization's social vice policy implications, which could have understated its strength relative to policy polarization.

The results support none of the hypotheses. A potential partner state's polarization had neither net, direct, nor indirect/interaction effects on respondents' willingness to accept a cooperation offer from that state. This was true even when examining the subsample of respondents that were above the 75th percentile in political sophistication. Polarization does have a significant direct effect and near-significant indirect/interaction effect among the subsample of sophisticates who passed both factual manipulation checks about the treatment; however, I cannot treat those results with high confidence due to potential post-treatment bias.

Other than null results regarding cooperation preference, the study also finds both anticipated causal logics to be inoperative. Polarization modestly increased perceptions the partner state would defect overall but did not increase perceptions the partner state would defect due to a government changeover. Nor did it impact perceptions about leader accountability for defection. Alternate explanations of perceived coethnicity or treatment information inequivalence do not seem to account for polarization's impact on general defection chances in the absence of the two anticipated dynamics. Finally, the study also finds that Myrick's (2022) finding that affective polarization does not encourage to commitment problems is robust to my alternate operationalization.

Combined, these results indicate that polarization does increases concern of defection beyond the U.S. case—but only moderately so. These modest impacts are not sufficient to induce a commitment problem by themselves and simply increase its

probability. That increased probability is moot or near-moot when the prospects of non-cooperation seem worse than those of partner defection. The results also suggest that public and foreign policy elites could reason about polarization differently, as polarization was more impactful in the pilot study's subsample of attentive undergraduates majoring in politics-related subjects than the full fielding's quota sample. Finally, given the undergraduate sample did not meaningfully differ from Myrick's quota sample of U.K. adults, the study suggests that foreign publics that hear a partner state is affectively polarized are more open to cooperation with them than if they hear the state is policy polarized. Applied, that means that polarization's impacts on perceived general defection chances in this study arise from concerns related to policy polarization rather than affective polarization.

Theory: Polarization, Defection, and Elite/Public Cooperation Preferences

This section argues that polarization increases state voluntary defection chances. It also argues that, given salience, elites and publics in other states accordingly perceive a polarized state as more unreliable and decrease their preference to cooperate with that state.

Polarization: Entangled Social-Ideological Division

Polarization is the condition of a polity that has high levels of social-ideological division (Schultz 2017, 7–8).[71] The term can refer to distinct concepts of affective polarization (hostility towards the political "other") and policy/ideological polarization (divergence of policy preferences) (Iyengar et al. 2019). The differences between these two concepts are

[71] My approach is similar to that of Schultz (2017, 7–8), which I encountered late in the writing process. He similarly takes polarization to refer to interrelated phenomena, listing policy (ideological) polarization, social sorting, and affective polarization. My discussion assumes the social sorting piece.

important, as they appear to have partially different origins and effects, and therefore suggest different approaches to polarization reduction (Abramowitz 2010; Huddy and Yair 2021; Iyengar et al. 2019; Levendusky 2018). In practice, though, the two are strongly interrelated to the point of being mutually constitutive. Policy polarization increases affective polarization (Rogowski and Sutherland 2016; Webster and Abramowitz 2017), and affective polarization reinforces policy polarization (Druckman et al. 2021b; 2021a; Lelkes 2021).[72] Their chicken-and-the-egg ordering is debated; the relative magnitude of one may be greater than the other; and some apparent polarization may be due to expressive responding (Iyengar, Sood, and Lelkes 2012; Mason 2015; Peterson and Iyengar 2021; Rogowski and Sutherland 2016; Schaffner and Luks 2018; Webster and Abramowitz 2017). Ultimately, though, social groups impact ideas, and vice-versa.

As illustrated by the citations in the previous paragraph, polarization's implications for domestic politics and especially U.S. domestic politics are subject to much discussion (Abramowitz 2010; Campbell 2018; Carmines, Ensley, and Wagner 2012; Fiorina, Abrams, and Pope 2006; Hetherington and Rudolph 2015; Mason 2018). Others have explored polarization's impact on international politics, with particular emphasis on its impacts to trade agreement outcomes, international negotiations, and cooperation prospects (De Vries, Hobolt, and Walter 2021; Friedrichs 2022a; 2022b;

[72] Interestingly, Levandusky and Malhotra (2016) found that people exposed to media coverage of polarization reported higher levels of affective polarization and lower extremity of their own issue positions. However, these self-reported results could be due to social desirability effects, as respondents could have blamed the outgroup for the polarization while underestimating the extremity of their own issue positions via a self-serving bias.

Friedrichs and Tama 2022; Myrick 2021; 2022; Schultz 2017). I focus on the latter category: expanding on and extending Myrick's argument that U.S. domestic polarization should encourage a commitment problem on the part of its longstanding U.K. ally. Particularly, I argue that polarization is most likely to impact voluntary defection chances and propose weakened leader accountability as a causal logic alongside government changeover risk. Then, discussing responsiveness dynamics in the context of international reputation, I make the case that these two causal logics are likely to be found in contexts broader than the United States, and therefore, so will polarization's encouragement of commitment problems.

Polarization's Institutional Impacts on Voluntary Defection Chances

Defection comes in two types: voluntary and involuntary. Voluntary defection occurs when a state chooses to renege on an agreement, and involuntary when a state fails to ratify an international agreement (Putnam 1988, 438). Polarization's impacts are highly likely most strong on voluntary defection. Polarization could impact chances of involuntary defection, as opposing sides necessarily have diverging policy platforms and therefore have difficulty reaching consensus (Friedrichs 2022a, 18; Myrick 2022, 7; Schultz 2017, 10–15). However, involuntary defection is less about a failure to reach domestic consensus and more about a misjudgment by agreement negotiators from all involved states.

In creating the agreement, negotiators consider not only their range of acceptable outcomes, but what their respective states are likely to ratify. They do so because risk of involuntary defection incurs side-payments to other involved states, and actual involuntary defection carries reputational consequences (Putnam 1988, 439, 453). As

probability of involuntary defection is included in agreement of negotiation, the driver of involuntary defection is not polarization but negotiator misjudgment about what the domestic consensus could support. A polarized legislative body certainly could fail to ratify an international agreement, but so could an unpolarized body. Legislative gridlock does not presuppose involuntary defection, but negotiators misjudging an agreement's ratification chances does. If anything, involuntary defection risk should decrease when a state is polarized, as the relative proportions of the distinct political coalitions is easier for negotiators to identify when hammering out agreements.

Polarization in a polity has a stronger theoretical impact on voluntary defection. One reason polarization impacts voluntary defection—also argued by Myrick and Schultz in the U.S. context—is that polarization increases the risk that government changeover means policy changeover. When coalitions' policy platforms starkly diverge, a new ruling coalition likely will have not only different but opposite policy priorities than the previous ruling coalition. Indeed, coalitions may be rewarded by electorates/selectorates for having such priorities (Myrick 2022, 7–8; Schultz 2017, 19–21).

A second institutional impact I propose is lessened leader accountability.[73] A core theoretical curb on leader defection is their incentive to continue in power (Aldrich et al. 2006; Fearon 1994; Fenno 1973; Putnam 1988; J. L. Weeks 2008). If a leader reneges on a commitment that some of their coalition finds important, that part of their coalition

[73] Schultz may have referenced this dynamic: "[Affective polarization] makes it harder for people to embrace policy proposals from the other side and makes it harder for elected officials to compromise across the aisle" (2017, 9). However, as the idea is not more discussed the reference could also simply refer to the separate implications of affective polarization for the public and elites without reference to accountability mechanisms. Schultz does not develop this inference further. He does argue that polarization encourages partisans to assume party-consistent positions, but does so to further argue that polarization reduces ability to learn from foreign policy mistakes. The argument does not engage with accountability/responsiveness dynamics.

might realign to support a domestic competitor, reducing the leader's chances of continued power. Defection is disincentivized. Polarization, though, short-circuits that accountability dynamic (Gallop and Greene 2021; Orhan 2022). Divergent policy platforms can make it easier for voters to engage in retrospective voting (Jones 2010; Stiers and Dassonneville 2020); but, by contributing to group distinction, policy polarization also creates social conditions that render supporters more responsive to their parties on uncrystallized issues or opinions of only moderate strength (Carmines, Ensley, and Wagner 2012; Rogowski 2018). As supporters' positions on such issues increasingly dovetail with those of their leader(s), likelihood decreases for co-partisan leader accountability on those issues.[74]

Evidence of this polarization-accountability dynamic in the context of foreign policy can be found in Gallop and Greene (2021). They theorized that polarized elites experienced less constraint on low-salience issues like foreign policy, and found cross-national evidence that higher levels of elite polarization correlated with the risky behavior of MID initiation (militarized interstate dispute). Orhan similarly found in a cross-national study that more affectively polarized states have lower levels of government accountability to their publics (2022, 726–27).

Broockman, Kalla, and Westwood provide experimental evidence to the contrary in the U.S. context (2023). They found that increasing partisan affective polarization

[74] Little et al. (2022) find on game theoretical grounds that voters' desensitization to incumbent performance is more responsible for variations in retrospective voting than diverging beliefs about incumbent performance. However, in practice, political polarization contains both of the theoretically distinguishable dynamics, and so it does not follow from their findings that polarization does not decrease accountability.

during a trust game did not reduce participants' willingness to vote for or against their congressional representatives (regardless of whether they knew the representative had voted congruently or incongruently with the participants' policy positions). They also found no evidence that more affectively-polarized people changed their policy positions to match those of their co-partisan congressman (or to oppose those of their out-partisan congressman) (2023, 19-23).

However, the authors acknowledge their study addressed only short-term effects of affective polarization. It did not test potential long term effects, such as theoretically relevant impacts to media choice or indirect impacts on policy positions through social interaction (2023, 38). Further, the key outcomes' operationalizations could have suppressed affective polarization's impacts. For vote choice, participants were given no co-partisan alternative to their current congressmen (2023, 66). Even a highly affectively-polarized person seems likely to vote for a co-partisan they disagree with if they have no other in-party alternative. Similarly, regarding policy position change: participants' policy positions were captured with a dichotomous support/oppose measure (2023, 58-59). That meant affective polarization would only register impacts on an issue position if the participant wholly inverted their initial issue position. That could be a tall order when the party cue comes from a legislator they might not know well enough to trust as a source for party cues. Moreover, and as the authors point out, a flash of interpersonal animus against an anonymous stranger or small group of strangers (as in a trust game) may have null impacts on accountability and issue positions because affective polarization could be a multidimensional construct (35-36). In sum, Broockman et al.'s

findings do not critically undermine the inference that polarization, to include its affective aspects, decreases leader accountability.

Polarization's International Audience: Elite and Public Cooperation Hesitation

Polarization's potential impact on voluntary defection matters more broadly because states interact with each other partially based on "learned" reputations.[75] That is, states forecast the probability of other states' future actions based on their past ones (Crescenzi 2007; Tomz 2012; Weisiger and Yarhi-Milo 2015; Renshon, Dafoe, and Huth 2018; Miller 2012). On the basis of such reputations and other assessments (such as motives and interests), states can adopt policies that are more or less trusting of another state, more or less accepting of risk (A. M. Hoffman 2002; Larson 1997, 701; Stiles 2018).[76] For example, a state that trusts its bargaining partner in a given context would be more willing to accept agreements with broad rather than specific language and weak rather than strong oversight provisions.

This learning incorporates information about states' past international behavior, but also domestic political factors that could impact international behavior. Several studies on international crises found that leader turnover impacted state reputations for resolve as perceived by other states or leaders (Kertzer, Renshon, and Yarhi-Milo 2021;

[75] Myrick discusses polarization's impacts on U.S. reputation, but only does so in terms of its international standing and international favorability toward the United States. She also assumes the learning chain I discuss, but does not explicate it or its conditionalities beyond a potential allusion to issue salience. My argument also expands on her argument regarding the importance of public opinion, particularly highlighting how publics can not only influence foreign policy but also learn from domestic contexts in other states (2022, 9-11).

[76] Scholars competingly treat trust as a rational assessment, an emotion, or an emotional belief (Brugger, Hasenclever, and Kasten 2013; Considine 2015; Haukkala, Van de Wetering, and Vuorelma 2018).

Schub 2020; Smith and Spaniel 2019; Wolford 2007). These impacts varied across regime type depending on the new leader's relationship to their predecessor's coalition (Wu, Licht, and Wolford 2021), and also varied depending on the leader's domestic constraints (Renshon, Dafoe, and Huth 2018). It mattered not only what a state had internationally done in the past, but what its domestic political context was.

Importantly, elites are not the only actors in a state who assess reputation and domestic political context. The public does, too. Kertzer, Renshon, and Yarhi-Milo's conjoint experiment found that both U.S. respondents and Israeli Knesset members used reputation, leader time in office, and regime type to calculate another state's resolve (2019). This is consistent with democratic peace survey experiments in China, the United States, and the United Kingdom (Bell and Quek 2018; Johns and Davies 2012; Tomz and Weeks 2013). They found that both authoritarian and democratic mass publics were less likely to prefer use of force against a democracy than an autocracy. These experiments do not focus on reputation or particular dynamics within a given state's domestic context, but they do suggest a state's domestic political context is not exogeneous to a foreign public's policy preferences regarding that state.

But, does it matter if publics assess other states and form foreign policy preferences based on those assessments? Responsiveness literature suggests the answer is "to some extent" (Bowler 2017). The exemplar of this is Goldsmith and Horiuchi (2012). They found that foreign publics' opinions about U.S. foreign policy positively correlated with their states' later policies toward the United States, an effect conditioned on issue

salience.[77] The operation of such responsiveness is somewhat though not wholly attenuated by many factors, such as elites conceiving of themselves as more of trustees than delegates (Hill, Jordan, and Hurley 2015; Pitkin 1967; Soontjens and Walgrave 2021). Regardless of perceived role, though, the threat of retrospective accountability incentivizes officeholders to consider the shadow of future accountability when making decisions (Aldrich et al. 2006; Colaresi 2012).

Another attenuating factor to consider is regime type. Official responsiveness to the public is not as strong a mechanism in autocracies as in democracies, as autocracies are definitionally possess weaker accountability to their publics. However, emerging literature suggests they still experience domestic constraint (Lueders 2021; Miller 2015). This constraint is most strongly tied to selectorate preferences but also is tied to public preferences via the specter of regime-toppling (Hyde and Saunders 2020; Weeks 2012; Weiss 2013). Key evidence here is how authoritarian states typically seek strong control of media in their borders, censoring opposing opinions and generating pro-regime content (Dukalskis and Gerschewski 2017; Frantz, Kendall-Taylor, and Wright 2020; Gunitsky 2015; Han 2015; Keremoğlu and Weidmann 2020; King, Pan, and Roberts 2014; Rød and Weidmann 2015). Contemporaneously, Russia's current smothering of domestic dissent regarding its invasion of Ukraine suggests regime continuation and public opinion on foreign policy are not wholly extricable (McMahon 2022).

The final attenuating factor of salience poses a greater challenge to foreign policy responsiveness. Many publics lack high interest in politics, and the subsamples that are

[77] The authors use a robustness check to rule out elite effects on the public.

interested tend to focus on domestic more than foreign affairs (Miller and Stokes 1963). Thus, one could argue that even if the public express an opinion about a given foreign policy on a survey, that opinion likely is not highly salient to them and by extension, their elected officials. However, Goldsmith and Horiuchi (2012) found that various publics' general foreign policy evaluations correlated with their states' low-salience foreign policy decisions, albeit to a lesser extent than high-salience foreign policy decisions like involvement in wars. This suggests the conditioning effect of salience reduces rather than negates the relationship between public opinion and foreign policy.

Open Questions: Unreliability, Affective Polarization, and Cross-National Applicability

Thus, there is a theoretical case that polarization can encourage foreign audience assessment of the polarized state's international unreliability and therefore reduce their preference to cooperate with that state. This is true for elite and public audiences, and in terms of scope, perhaps operative in both democratic and autocratic states. In terms of empirical support, Myrick (2022) experimentally tests a portion of the theoretical case but leaves untested a core element: the causal role of perceived unreliability. The study identified a positive causal effect of (1) information about U.S. policy polarization on U.K. respondents' (2a) perceptions of future U.S. unreliability and (2b) preference against future partnerships with the United States, but did not establish a causal chain from polarization to unreliability to cooperation hesitation.

Indeed, it was not designed to. Its design captured unreliability as a post-treatment outcome rather than including it as a manipulated factor in the treatment. This precludes contemporary methods of mediation analysis, because post-treatment measurement renders the unreliability variable non-randomized, and therefore potentially subject to

bias (Acharya, Blackwell, and Sen 2018, 363; Imai, Tingley, and Yamamoto 2013). Thus, what Myrick found is consistent with a polarization-encouraged commitment problem but does not establish it.

The untested unreliability assumption leaves open the possibility that polarization influences cooperation by channels other than perceived unreliability. For instance, one could infer from a state's polarization not that the state is unreliable, but that its people are unlikely to be coethnic with the observer (i.e. "people like me wouldn't behave like that"). This would result in unequal inferences regarding coethnicity across treatment groups that could impact cooperation outcomes (Dafoe, Zhang, and Caughey 2018), as perceived ethnic or cultural similarity positively correlates with trusting attitudes and behaviors (Dinesen, Schaeffer, and Sønderskov 2020). Another potential alternate explanation, as Myrick points out, could simply be that people dislike a state more if it is polarized, and want to cooperate with it less as a result (2022. 9). An exploratory observational analysis I conducted of Pew Global Attitudes data is consistent with the unreliability explanation (Appendix P; Pew 2021b). It found increased perceptions of U.S. political dysfunction negatively correlated with perceptions of its reliability, and the latter correlated with decreased cross-national preference for cooperation with the United States. However, the cross-sectional analysis cannot establish the point causally.

Beyond unreliability point, Myrick's study points to other open questions. She identifies several regarding contemporary U.S. polarization, and to those I would add questions of broader generalizability. Is there a relationship between polarization and cooperation hesitation when the given polarized state is not the United States? U.S. polarization is widely commented upon in the U.K. press, with such coverage likely

including its implications for the United Kingdom. The polarization-cooperation link could be less pronounced when the polarization in question is less salient. A further question regards the potential difference between general cooperation preference and specific policy support.[78] Will the findings hold when one presents respondents a concrete cooperation scenario with discrete policy options? This question matters because states ultimately face discrete choices with regards to cooperation rather than a five-point scale of general cooperation preference strength.

A further open question is whether polarization actually impacts perceptions of leader accountability and government changeover risk. Myrick's study did not test the government changeover facet of her theory, and to my knowledge, no study has tested this paper's proposed leader accountability element. Finally, a remaining outstanding question is whether a scope condition Myrick identified was an artifact of the study's research design. The study demonstrated not only that U.S. policy polarization impacted U.K. respondents, but also that U.S. affective polarization did not. This is surprising, given the two phenomena's co-occurrence and mutual reinforcement (Hetherington 2015; Hetherington and Rudolph 2015; Rogowski and Sutherland 2016; Webster and Abramowitz 2017).

The unexpected finding could be the result of a strength asymmetry in the study's treatments (Table 21). The policy polarization treatment noted Americans have different attitudes on both social and economic policies and think the parties "cannot agree on

[78] E.g., in the Myrick study, respondents were asked their agreement with the broad statement "My country should partner with the United States in future international agreements." (2022, 13).

basic facts." It also noted both parties' politicians disagree on many "basic policy issues" and vote according to party lines (2022, 12). On the other hand, the affective polarization treatment noted Americans are against their children marrying across party lines, have few friendships across party lines, and "strongly dislike" or "hate" opponent party members. It also notes that both parties' politicians "[u]se extreme, negative language to taunt politicians of the other party" and "post angry or hateful posts on social media about members of the other party" (2022, 12). The policy polarization treatment, with its emphasis on public/politicians' disagreement on many basic issues along party lines,

Table 21. Polarization Treatment Operationalizations (Myrick 2022)

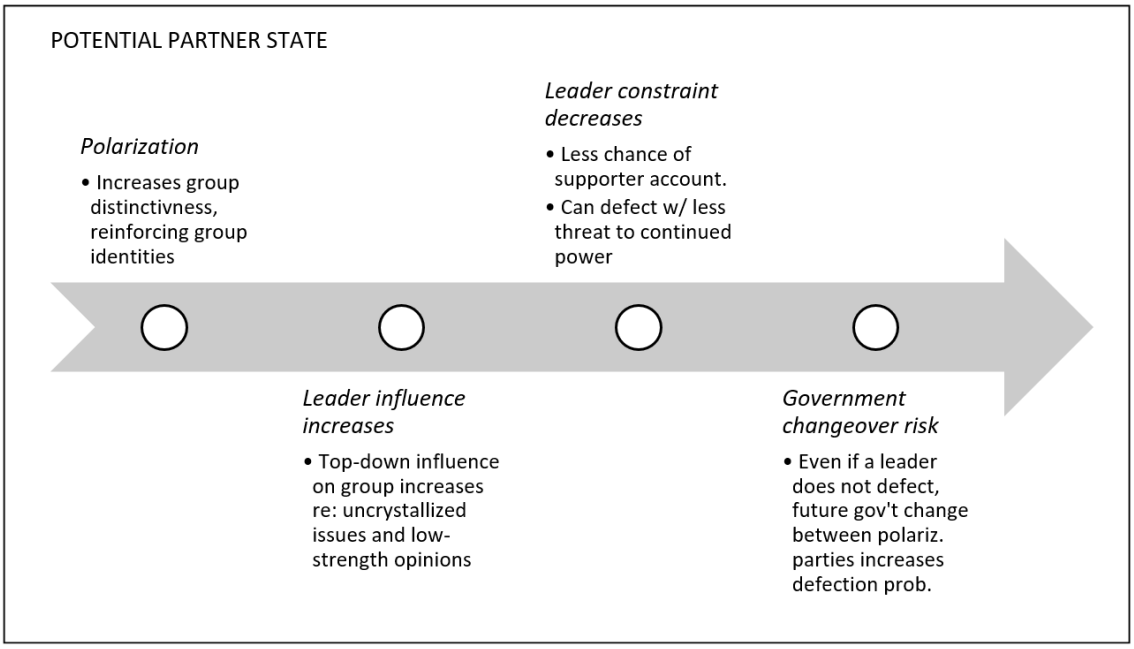
Introduction Language	**Policy Polarization Treatment**	**Affective Polarization Treatment**
Surveys from the United States show that, more than ever, Americans:	• Have different attitudes about social issues, such as abortion rights and gun laws. • Have different preferences over economic policies, such as tax rates and welfare spending. • Think their political parties cannot agree on basic facts.	• Oppose the idea of their child marrying someone from the other political party. • Have 'just a few' or 'no' close friends from the other political party. • 'Strongly dislike' or even 'hate' members of the other political party.
These differences are reflected in the US government. More than ever, Republican and Democratic politicians:	• Disagree on a wide range of basic policy issues. • Vote the same way as members of their own political party.	• Use extreme, negative language to taunt politicians of the other party. • Post angry or hateful posts on social media about members of the other party.

likely creates a clearer impression of policy unreliability than the other's suggestion that the public is insular and politicians are publicly mean. In the former, government changeover likely will be accompanied by policy change; in the latter, politicians may just be posturing publicly for political points while the business of government continues.

Theoretical Expectations

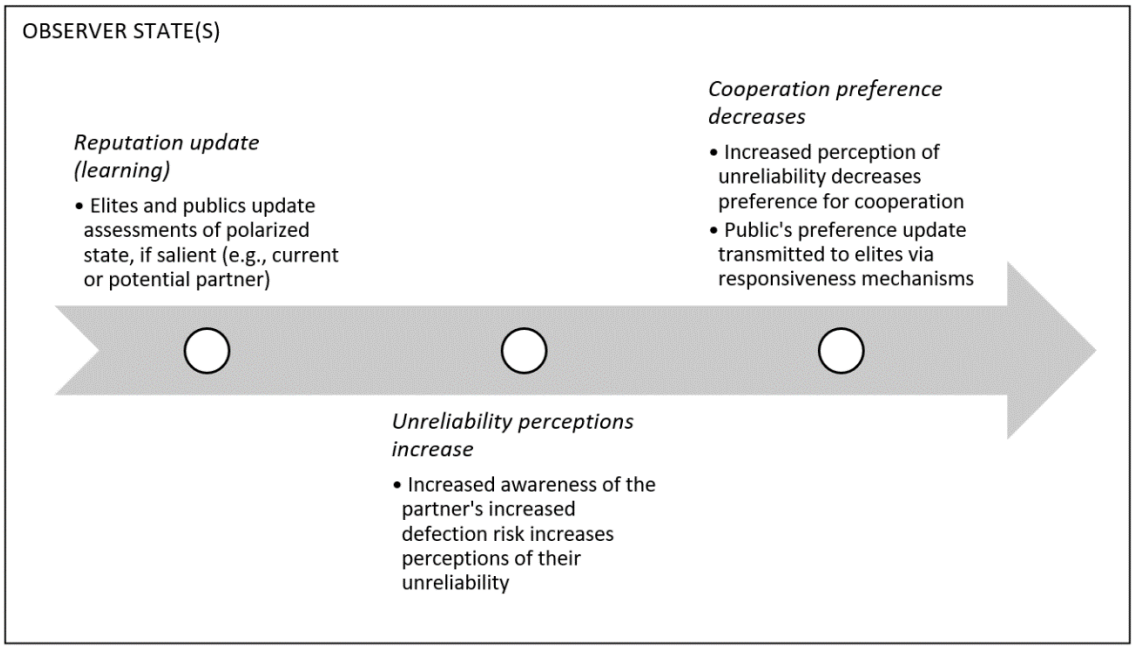
In sum, polarization in a potential partner state may reduce willingness to cooperate with that state due to increased risk of defection (Figure 12). Polarization—entangled policy and affective polarization—decreases accountability constraints on leaders in that they have more ability to lead supporters' opinions on low-salience topics like foreign policy. This decreased constraint is the result of the polarization's long-term social implications: increasing group distinctiveness and therefore the influence thereof on policy positions. Less-constrained leaders have more latitude to defect from commitments because less accountability means less risk to their continuance in power. Polarization also increases long-term defection risk because of the parties' diverging policy platforms. These policy divides mean that when one side assumes control of the government, its policies likely will diverge from those of the previous government. As this is what the new government's coalition wants or at least was willing to accept, this defection also is unlikely to be punished through retrospective mechanisms.

Figure 12. Theory: Polarization and Defection Risk



Such changes in defection probability are not opaque to other states, as they and their publics learn from the polarized state's domestic context if the state is salient to them (Figure 13). In turn, those elites and/or publics adjust their assessment of the target state in terms of reliability. These assessments shift elite/public foreign policy preferences regarding that state and influence their state's foreign policy via the mechanism of responsiveness.

Figure 13. Theory: Observer State Response to Polarization



Below are those expectations restated in hypotheses.

Hypothesis 1 (Polarization): Receiving information that a state is polarized will decrease individuals' preference to cooperate with that state, provided the state is salient to them.

Hypothesis 2 (Unreliability): Assessments of the polarized state's unreliability will account for a portion of polarization's negative effect on cooperation preference.

Hypothesis 3a-b (Causal Logic Tracing): Receiving information that a state is polarized will (a) decrease perceptions that the potential partner state's leader will be held accountable by supporters in case of defection;

and (b) increase perceived probability that government changeover will lead to defection.

I also explore the role of policy v. affective polarization. I anticipate that changing Myrick's affective polarization treatment so it communicates the policy implications of affective polarization will result in it impacting unreliability assessments and cooperation preference:

Hypothesis 4 (Affective Polarization): A partner state's affective polarization will decrease assessments of that state's reliability and preference to cooperate with that state.

Empirical Approach Overview

I tested Hypotheses 1-3 with a between-subjects, 2x2 survey experiment (polarized × unreliable). The experiment used an international politics vignette to deliver the polarization/unreliability information regarding a potential partner state. The use of a scenario also made it possible for the cooperation outcome to be discrete, concrete policy choices rather than general cooperation preference strength.

Advertised as being about "policy preferences", the experiment was fielded twice. First was a pilot conducted with 300 undergraduates enrolled in politics/global studies courses at a major southwestern research university. Based on the pilot results, I refined the study protocol (Appendices P). A large majority of pilot respondents preferred cooperation even when they were told the potential partner was unreliable, which

suggested the scenario was strongly slanted in favor of cooperation. In response, I modified the scenario to reduce the level of threat associated with non-cooperation. Second, survey firm CloudResearch Connect provided a quota sample demographically matched to the 2020 U.S. census (n = 750).[79] The student sample was compensated with research credit for their course and the CloudResearch sample received $1.06 each.

The undergraduate sample also participated in a separate survey experiment that tested Hypothesis 4 (affective polarization operationalization). The experiment, conducted sequentially with the other experiment, was a simple two-condition experiment that largely replicated Myrick's original design. A few deviations were necessary as my sample were largely Americans, while Myrick's sample from the United Kingdom and her subject was opinions about the United States. I discuss this second experiment's design and results after that of the 2x2 survey experiment. All studies were IRB-approved (Appendix R), and their question wordings are in Appendices W-X.

Method: Experiment 1 (Polarization and Cooperation Hesitation)

Design Choice: Acharya, Blackwell, and Sen (2018)

The 2x2 experiment's design is from Acharya, Blackwell, and Sen (2018).[80] Their factorial approach allows the identification of causal mechanisms via experimentation without tightly restrictive assumptions (Baron and Kenny 1986; Imai, Keele, and

[79] The census-matched quotas were as follows: *gender* (50%/50%), *age* (21.8%, 18-29; 25.8%, 30-44; 26.4%, 45-49; and 26% 60+); *ethnicity* (16% of Hispanic, Latino, or Spanish origin; 84% not); and *race* (78.2% white; 13.8% Black or African American; and 8% other).

[80] They build on Imai, Tingley, and Yamamoto (2013), Gerber and Green, and VanderWeele (2015). The first proposed the parallel experiment design; the second introduced "implicit mediation analysis"; and the third identified that the difference between ATE and ACDE is the combination of the mediated effect and interaction of the mediator and treatment.

Yamamoto 2010; Imai, Tingley, and Yamamoto 2013) and avoids the bias introduced by post-treatment, observational measurement of mediators.[81] It does this by defining causal mechanism broadly to include both mediation and interaction effects (VanderWeele 2015), and by including the mediator as a dichotomous factor in the experiment (Imai, Tingley, and Yamamoto 2013). The factor is set at 0 (no information provided regarding the mediator—in this experiment, the partner state's unreliability) and 1 (the value of the mediator is provided—"unreliable"). The other factor is the treatment, set at 0 (treatment "off", or "unpolarized") and 1 (treatment "on", or "polarized").

The result is a "natural mediator arm" (NMA) and "manipulated mediator arm" (MMA) of two treatment groups each. The NMA groups (00, 10) receive randomized information about potential partner state polarization but information about the unreliability mediator was omitted. In these groups, the mediation that occurs is "natural"—via respondents' inferences.[82] The MMA's groups (01, 11) received randomized information about polarization but fixed information that the potential partner state was sometimes unreliable.

This design allowed me to calculate unreliability's indirect contribution to cooperation preference.[83] Specifically, the NMA groups (00, 10) provide the average treatment effect (ATE): the average direct, indirect, and interaction effects of the polarization treatment. The MMA groups (01, 11), on the other hand, allow calculation of

[81] The bias is introduced because the mediator is not randomized, and therefore the relationship of the treatment via the mediator could be impacted by confounding variables (Acharya, Blackwell, and Sen 2018, 363).

[82] Acharya, Blackwell, and Sen draw this from Gerber and Green (2012, 333–36).

[83] I am indebted to Michael Bechtel and Kenneth Scheve for their Stata files on Harvard Dataverse that showed examples of how to code Acharya et. al's approach (2013)

the average controlled direct effect (ACDE) of polarization on the outcome. This is possible because they hold the anticipated unreliability mediator constant. The difference between the ATE and ACDE represents the quantity of interest for Hypothesis 2: unreliability's impact on cooperation preference via indirect effect and interaction.

This design is appropriate because I anticipate polarization to have an indirect effect on cooperation preference via perceptions of unreliability, and I cannot rule out an interaction between polarization and unreliability. Its limitation is it cannot disaggregate polarization's indirect and interaction effects via unreliability. This limitation is preferable to violating the core assumptions of other causal mediation analyses—largely, that no treatment-mediator interaction exists and/or that mediators measured post-treatment are as good as randomized (Acharya, Blackwell, and Sen 2018, 361, 367–68; Baron and Kenny 1986; Imai, Keele, and Yamamoto 2010; Imai, Tingley, and Yamamoto 2013). Better an accurate though imprecise measurement than one that could be biased in unknown directions and magnitudes. The limitation does mean, though, that a positive interaction effect could mask the hypothesized negative indirect effect.

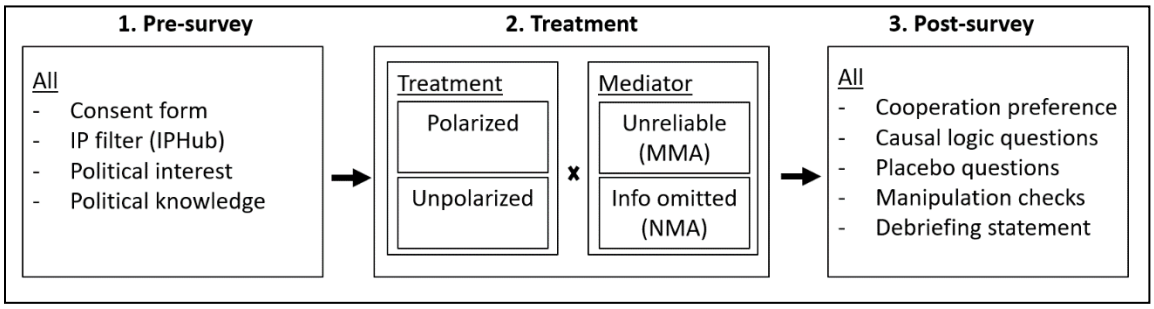
Scenario

The survey presents the respondents' state (unnamed) with a choice to cooperate with another state to discourage the aggression of a third state. The scenario is drawn from the NATO alignment choice some countries have encountered since the 2014 invasion of Crimea. In terms of detail, it provides only necessary context. The more information provided, the more realistic the scenario; but the more information provided, the less chance respondents will encounter and recall the treatment information. As this is an

initial analysis, the goal is to identify whether the hypothesized relationships exist at all, and so I err on the side of finding treatment effects (Brutger et al. 2022).

Below is the design (Figure 14), and a description of its contents for the full fielding. The pilot study had a slightly different version of the protocol, as I refined the treatment/questions based on its results (Appendix O).

Figure 14. Design: Experiment 1 (Aurl/Zerm)



Respondents initially encountered the consent form, a hidden "trap" question to identify bots, and a location-screening by free service IPHub to confirm respondents were actually in the United States (Winter et al. 2019).[84] Then, the survey opened by asking respondents' general level of political interest, interest in campaigns this year, and interest in foreign policy (scale 1-5). Five political knowledge questions followed, using an ANES protocol to nudge respondents to not look up responses.[85] The questions queried political knowledge associated with partisan competition in the United States

[84] The "trap" question consisted of a question hidden by JavaScript. This meant that an answer to the question indicated a respondent was a bot, because human respondents could not see the question, much less interact with it to answer it.

[85] The ANES protocol informs respondents that the researchers are interested in the guesses they make when they don't know question answers and asks them whether they are willing to promise to not look up answers. It then asks respondents a question they almost certainly do not know the answer to (the year of an obscure Supreme Court case) to permit identification of respondents who probably looked up the answer. Respondents who correctly identify the year receive an additional nudge in the form of a question that asks whether they had known the answer already or had looked it up (ANES 2021, 311–12).

(House majority party, Congressional majority to override presidential veto) and foreign policy (prime minister of the United Kingdom, whether China and Taiwan are allies, what five states are permanent members of the United Nations Security Council). I included the political interest/knowledge questions based on pilot study results that indicated political sophistication was an important moderator. Specifically, the pilot analysis indicated that undergraduates who majored in/were considering majoring in politics-related majors were closer to displaying the hypothesized dynamics than the other undergraduates.[86]

> After the political sophistication measures, the survey introduced the scenario:[87]
>
> *You will be asked to read a political scenario and provide your opinions on it.*
>
> *Please read it carefully. Some parts of it may strike you as important; other parts may seem unimportant.*

The survey then presented background information about the three states, to include the polarization/unreliability treatments. The state identities (Aurl/Zerm) were randomly assigned to mitigate bias, but for ease of reading, I refer to the potential ally as Aurl and

[86] Political interest and political knowledge are two of three general ways to proxy political sophistication, the other being education. For the political knowledge questions, I omitted "don't know" as a response option in order to avoid measuring respondents' propensity for guessing (Barabas et al. 2014; Gallina 2023; Highton 2009; Jessee 2017; Lupton, Myers, and Thornton 2015; Luskin 1987; Luskin and Bullock 2011; Milesi 2016; Mondak 2000; Solvak 2009; Weith and Krouwel 2013)

[87] This language is a modification of that from Tomz and Weeks (2013).

potential enemy as Zerm.

> *You are a citizen of a given country. You follow the news about a major rivalry between two other nearby countries: Aurl and Zerm.*
>
> - *Aurl and Zerm possess roughly equal strength, and both are stronger than your country*
> - *Your country is somewhat friendlier with Aurl than Zerm*
> - *Aurl is very politically divided, with major disagreements over policies and mistrust/disrespect between opposing sides* [OR *Aurl is not very politically divided, with minor disagreement over policies and trust/respect between opposing sides*]
> - *Aurl sometimes does not follow through on its international promises* [OR no information provided here]

The respondents then encountered the scenario, which seeks to establish a situation in which cooperation with Aurl is preferable to non-cooperation if one trusts Aurl to not defect. Its wording was refined from the pilot study, whose results indicated the pilot scenario may have been slanted in favor of cooperation with Aurl.

> *Zerm invades a country near you that was increasingly friendly with Aurl. It threatens that other countries friendly with Aurl may be next unless they demonstrate that they are not threats to Zerm.*

Aurl offers to immediately station powerful defense systems in your country to protect you from Zerm as long as necessary.

- *The systems would certainly repel an attack by Zerm.*
- *Aurl says it will not withdraw the systems until your country is out of danger.*
- *Zerm says your country's acceptance of Aurl's offer would demonstrate that your country is a threat to Zerm. It also says your country's rejection of Aurl's offer would demonstrate that your country is not a threat to Zerm.*

The survey then asked, "Which policy option do you prefer your country follow?", with a dichotomous choice to accept or reject Aurl's offer. It concluded with three blocks of questions that contain manipulation checks and permit me to test Hypothesis 3 and assumptions regarding information equivalence. The manipulation checks are three-fold. The first, a subjective manipulation check, tests whether the scenario encouraged respondents to think Zerm could be trusted to not invade their country if their country rejected Aurl's offer. If respondents thought Zerm would invade if they rejected Aurl's offer, then Aurl's domestic conditions likely would have little impact on respondents' cooperation preference. The final two are factual manipulation checks testing respondents' ability to correctly identify the information they were provided regarding Aurl's polarization and unreliability.

The causal logic questions test whether the treatment impacted concern that Aurl would defect (Hypothesis 3a-b). Particularly, they asked how probable Aurl is to keep its

promise in general and in case of a government handover in Aurl. They also asked how likely Aurl's leader is to be held accountable by their supporters and by their opponents—a bifurcation recommended by cognitive interviewing after the pilot study.[88] As these questions are asked post-treatment, they can only help establish if polarization impacts perceived defection chances as expected; it cannot establish those chances act as a mediator or moderator.

The question regarding government changeover, presented last of the three causal logic questions, was unique in its presentation. It was embedded in an update to the scenario rather than a hypothetical. I selected this approach because a hypothetical version of the question seemed excessively cognitively taxing, while a narrative approach seemed less so.[89]

Final section of the scenario:

Your country chooses to accept Aurl's offer.

- *Just one week later, the defense systems are stationed and operational in your country*
- *Zerm says the defense systems demonstrate that your country is a threat to Zerm*

[88] A cognitive interviewee explained that they had a hard time answering a general accountability question because polarization meant on one hand that a leader's supporters were less likely to hold them accountable and on the other hand that the leader's opponents were more likely to hold them accountable.

[89] If your country accepts Aurl's offer, and Aurl's government later changes hands from one side to the other, how likely is Aurl to keep its promise to keep its defense systems in your country as long as necessary?

After some time passes, Aurl's government changes hands from its current administration to the opposing side. Everything else in the scenario remains the same.

In this scenario, how likely is Aurl to keep its promise to keep its defense systems in your country as long as necessary?

The general defection question is important because provides a theoretical backstop: if Hypotheses 2-3 produce null results, I can test to see whether respondents infer defection chances from polarization at all. Importantly, I varied these three questions' wording so some respondents were asked about the probability of Aurl breaking its promise, and others were asked about Aurl keeping its promise. The wording variation permits identification of potential framing effects from using positive v. negative language.

Finally, placebo questions tested information equivalence between the four groups. While the only information I varied was the polarization and unreliability factors, those manipulations could have asymmetrically triggered background assumptions across treatment groups, with resultant potential bias to treatment effects (Dafoe, Zhang, and Caughey 2018). To test for this possibility, the placebo questions queried how likely Aurl and Zerm's citizens were to be "people like you overall" and "people like you" in terms of their values/beliefs, appearance, language, and goals.[90] They also asked how likely the

[90] In the pilot study, I also randomly asked some respondents how likely Aurl/Zerm's citizens were to be "unlike" them in general and those particular ways. However, the "unlike" wording was more cognitively taxing than the "like" wording due to its creation of a double-negative (e.g., "they are somewhat unlikely to be unlike me" v "they are somewhat unlikely to be like me").

two states were to be majority Christian, majority Caucasian, a former communist country, or located on a particular continent. Finally, they also asked if Aurl and Zerm reminded respondents of any countries, and if so, which one.

The placebo questions matter because respondents may be more/less likely to cooperate with Aurl if they associate a particular scenario more with real-world factors than another scenario.[91] For example, respondents in the two polarized conditions may be more likely to associate Aurl with the United States than those in the two unpolarized conditions, and by extension to associate the scenario with current events involving the United States and Russia. Similarly, treatment effects may be biased if the treatments encourage respondents to perceive Aurl's citizens as being their co-ethnics (or Zerm's citizens as not being their co-ethnics). Overall, these placebo questions will not prevent violations of information equivalence but will allow me to identify violations.

A summary of the included variables and their operationalizations is below (Table 22). One of the variables (education) was provided by the polling firm.

Table 22. Variable Summary

Variable	**Category**	**Operationalization**
Cooperation preference	DV	Dichotomous (reject/accept offer)
Polarization	IV (manipulated)	Dichotomous (unpol./polarized)
Unreliability	IV (manipulated)	Dichotomous (omitted/unreliable)
Defect. prob. General Gov't handover	IV/Causal logic check	Ordinal (5pt, unlikely/likely)

[91] To further mitigate this risk, I considered including an embedded natural experiment, or plausibly random genesis narrative of Zerm's polarization and unreliability (Dafoe, Zhang, and Caughey 2018). However, I have not been able to identify a scenario that has no apparent risk of communicating unintended information about the explanatory variables. E.g., I could use a scenario in which Zerm's current political conditions arose by chance after a lone gunman succeeded in killing Zerm's popular leader twenty years before. But, what does that say about Zerm's political stability (and therefore reliability) if one lone gunman can so radically impact the functioning of the state?

Leader accountability Supporters Opponents	IV/Causal logic check	Ordinal (5pt, unlikely/likely)
Coethnicity (both Aurl/Zerm) General Values/beliefs Appearance Language Goals Index	Placebo	Varies Ordinal (5pt, unlikely/likely) Ordinal (5pt, unlikely/likely) Ordinal (5pt, unlikely/likely) Ordinal (5pt, unlikely/likely) Ordinal (5pt, unlikely/likely) Continuous (average of above)
Real-world ID/traits (both Aurl/Zerm) Majority Christian Majority Caucasian Former communist Reminds of Russia and U.S	Placebo	Varies Ordinal (5pt, unlikely/likely) Ordinal (5pt, unlikely/likely) Ordinal (5pt, unlikely/likely) Nominal, recoded dichotomous
Political interest General Campaigns Foreign policy	Potential moderator element	Varies Ordinal (5pt, never/always) Ordinal (3pt, not much/v. much) Ordinal (3pt, not much/v. much)
Political knowledge Current House majority Congressional veto override Taiwan-China relationship U.K. Prime Minister UNSC permanent members	Potential moderator element	Varies Ordinal (5pt), recoded dichot. Ordinal (5pt), recoded dichot. Dichotomous Ordinal (5pt), recoded dichot. Nominal, recoded ordinal (6pt)
Education	Potential moderator element	8pt, no formal education or high school incomplete/professional degree or doctorate
Political sophistication	Potential moderator	Standardized index of the three political interest variables and an index of the five political knowledge variables
Prob. Zerm invade if reject Aurl's offer	Subjective manipulation check	Ordinal (5pt, unlikely/likely)
Attentiveness: correctly IDing Polarization treatm. info Unreliability treatm. info	Factual manipulation check	Dichotomous (incorrect/correct)
Placebo positive framing	Design check	Dichotomous (break/keep)

Results: Experiment 1

Sample Quality

Before testing the hypotheses, I first address sample quality. The two sample quality metrics suggest the sample was human but also was only moderately attentive. None of the respondents answered the trap question, and 82.93% and 76.40% of them respectively identified the information they were given about polarization and unreliability. However, only 68.0% of the sample answered both factual manipulation checks correctly (Table 23).

Table 23. Performance on Factual Manipulation Checks

	Overall	**Group 00 (Unpol/Omit)**	**Group 01 (Unpol/Unrel)**	**Group 10 (Pol/Omit)**	**Group 11 (Pol/Unrel)**
Polarization	82.93%	82.04%	78.92%	85.71%	85.34%
Unreliability	76.40%	80.58%	73.51%	83.93%	68.06%
Both	68.0%	71.36%	62.16%	73.81%	64.92%

Further, while performance on the polarization question did not vary significantly by treatment group, performance on the unreliability question (and therefore overall performance) did. Respondents told nothing about Aurl's unreliability performed the best (80.58-83.93% answered the unreliability question correctly). But those told Aurl was unreliable performed worse. The unpolarized group (Group 01) had only 73.51% answer correctly, and the polarized group had the least (Group 11, 68.6%).

Of note, the factual checks were post-treatment, and therefore could be subject to bias via a confounding relationship between the treatment and factual manipulation check performance. This means that sub-setting to focus on "attentive" respondents alone may effectively be imposing a selection bias "of unknown sign and magnitude" (Aronow, Baron, and Pinson 2019, 573). With that enormous caveat: taken at face value, the

asymmetrical attention results suggest that polarization's ACDE may be suppressed in magnitude and significance more than its ATE (which may already be noisy due to around 30% of inattentive respondents in Groups 00-10). In turn, both would mean a noisy measure for EE. Unless the effect sizes are large, they may be difficult to find.

Another potential attentiveness issue is that nearly 10% of the sample not only answered the unreliability check incorrectly, but in the wrong direction. Some of the groups were told Aurl sometimes does not follow through on its international promises, some were told nothing about Aurl's unreliability, and none were told Aurl "almost always follows through on its international promises." Yet, 8.93% of the sample either guessed, mistakenly remembered, or mis-selected the third option.

These responses appear to have been in part sincere misperceptions rather than mistakes or guesses. In Table 24, I compare the distribution of perceived general Aurl defection chance among those who overestimated Aurl's reliability and who incorrectly answered they were told nothing about Aurl's reliability. In this subsample, the reliability over-estimators had correspondingly lower assessments of Aurl's general defection chances (*n = 110* from Groups 01 and 11). These assessments are only moderately lower, with 69.23% v. 57.14% judging Aurl defection "unlikely" or "somewhat unlikely." But, they would work to amplify the measurement noise that the inattentive third of the sample will already introduce. The significance of all three of my main measures is likely suppressed.

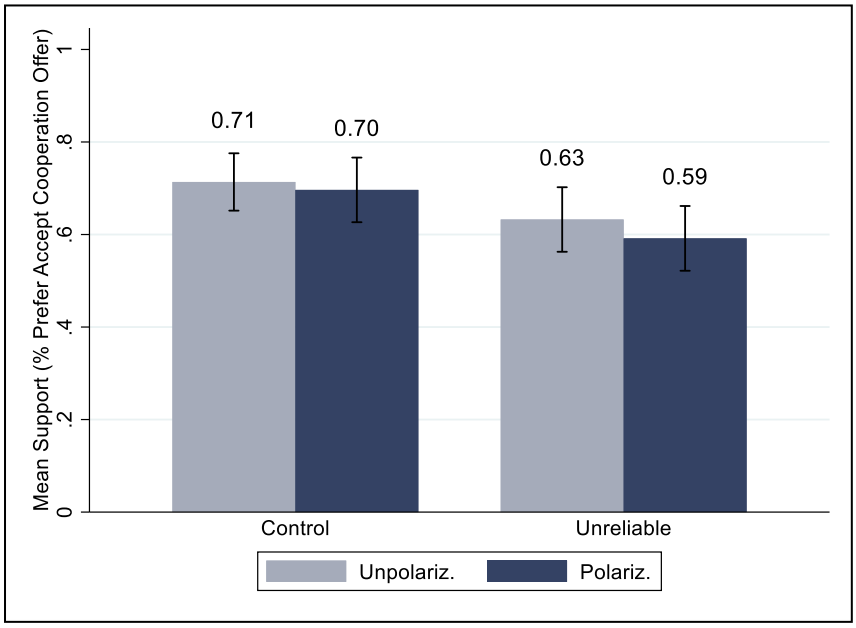
Table 24. Assessed Aurl Defection Chances by Incorrect Unreliability Answer

	Unlikely	**Somewhat unlikely**	**Neither likely nor unlikely**	**Somewhat likely**	**Likely**
"Almost always follows through"	7 (26.92%)	11 (42.31%)	4 (15.38%)	4 (15.38%)	0 (0%)
"It did not say anything about it"	17 (20.24%)	31 (36.90%)	13 (15.48%)	22 (26.19%)	1 (1.19%)

Polarization and Unreliability (Hypotheses 1-2)

Taken as a whole, the treatment group means do not display the expected pattern (Figure 15). I anticipated that Aurl offer acceptance would be lower among "control" condition respondents told Aurl was polarized rather than unpolarized. Lacking any information about Aurl's unreliability, I expected they would infer unreliability from its polarization. No such reasoning is evidenced in the groups' means (0.71 v. 0.70, $p = 0.72$). Further, among respondents told that Aurl is polarized (dark blue columns), also being told Aurl was unreliable decreased preference for cooperation relative to those told nothing of Aurl's unreliability (0.70 v 0.59, $p = 0.04$). This again suggests that those not told of Aurl's unreliability did not infer it from Aurl's polarization. Or, at least, it suggests that whatever unreliability level they inferred was less than the given amount of: "sometimes does not follow through on… international promises." In sum, if a state's domestic polarization increases concerns about the state's international reliability, it does not do so at levels that register in this study.

Figure 15. Means: Offer Acceptance by Treatment Group



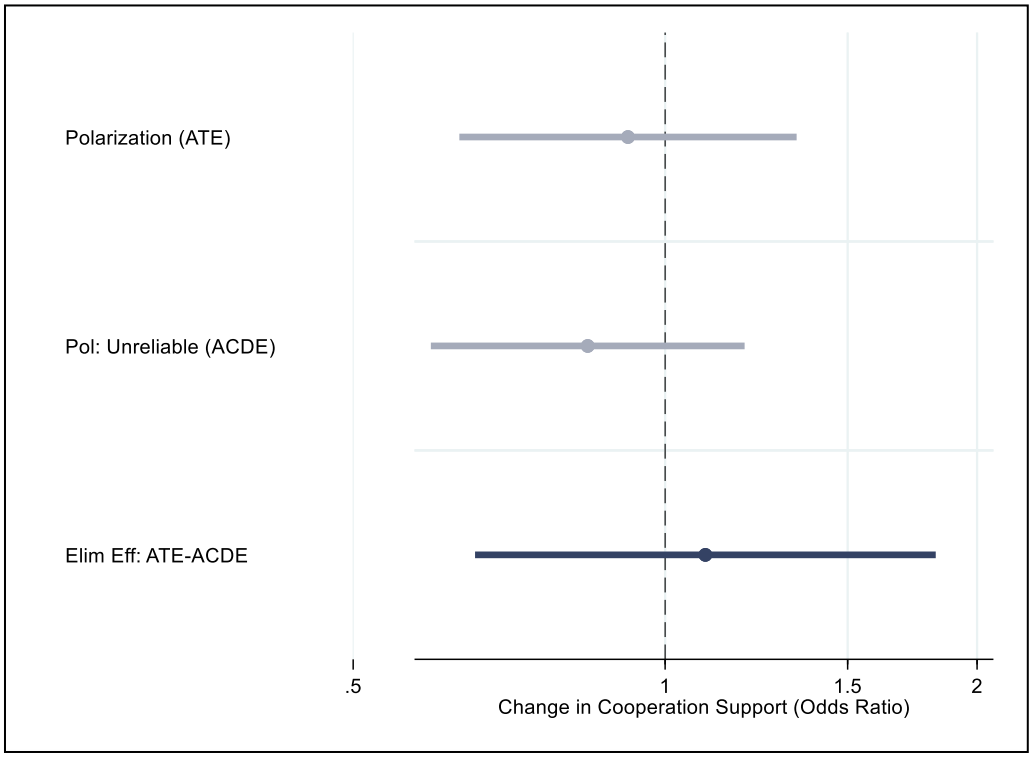
The absence of an apparent relationship between polarization and unreliability could be because the scenario remains slanted toward Aurl cooperation. As shown in Figure 15, support for cooperation is relatively high across all groups. To be sure, the information about Aurl sometimes not following through on international promises mattered, as it decreased means from 0.71 to 0.63 in the unpolarized condition and 0.70 to 0.59 in the polarized condition ($p = 0.09$ and 0.04). However, even those told Aurl was unreliable had a majority prefer to accept Aurl's offer (59-63%). The scenario thus appears to have been slanted toward encouraging cooperation, somehow creating the impression that rejecting the offer was more risky than accepting it, even though an Aurl defection would be very costly.

Particularly, the scenario's assertion of an international Aurl-Zerm rivalry likely communicated that Aurl would probably follow-through on its promise, regardless of past unreliability on unspecified issues. Security issues are important, especially ones

involving one's adversaries. Moreover, the rivalry could have appeared to have been one issue on which Aurl's opposing sides actually agreed. Respondents may thus have been left with a choice between (1) non-cooperation and taking Zerm at their word and (2) cooperating with a polarized Aurl whose modestly higher level of general unreliability is washed out by their probably quite-strong security considerations.

Turning to the three estimates of interest (ATE, ACDE, EE), they offer no more support to Hypotheses 1-2 than the difference-in-mean results (Figure 16). Contra Hypothesis 1, polarization's ATE was statistically insignificant ($p = 0.717$). This means Aurl's polarization had no net direct, indirect, and interaction effects on respondent willingness to accept Aurl's offer. Polarization's ACDE was similarly insignificant, meaning Aurl's polarization exerts no direct influence on respondent willingness to accept their offer of help ($p = 0.417$). As a result, polarization's EE—or indirect and interaction effects on offer acceptance via unreliability—is also insignificant ($p = 0.774$). That result is contra Hypothesis 2, which expected polarization to have an indirect effect on offer acceptance via perceptions of Aurlian unreliability.

Figure 16. Estimated Impact of Polarization on Offer Acceptance (Odds Ratios)



ATE: avg. treatm. effect; ACDE: average direct controlled effect; EE: eliminated effect
Coefficients are logarithmically transformed to facilitate ease of visual interpretation
Model: logistic

These null result echo those of the pilot study, which also found no significant impact of polarization's ATE, ACDE, or EE on Aurl offer acceptance. However, the pilot study found that the hypothesized effects of polarization and unreliability were evident at nearly statistically significant levels among the subsample of undergraduates who were majoring or considering majoring in politics-related disciplines (Appendix O). On the inference their major may have proxied for political sophistication, I repeated the regression analysis on the subsample of respondents that were above the 75th percentile of political sophistication.[92] Of note: an even more impactful moderator in the pilot study

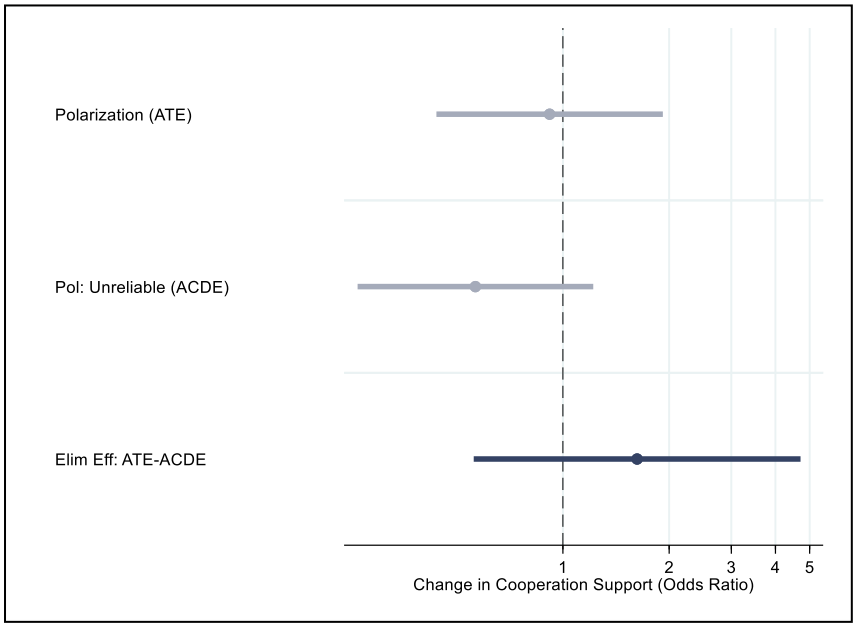
[92] Political sophistication was an index variable of the standardized political interest variables and political knowledge index (alpha = 0.7927). The political knowledge index had questionable inter-

was attentiveness, as it made some of the results achieve statistical significance despite the sample being remarkably small. However, the attentiveness measure was pre-treatment in the pilot: something impossible for this study's factual manipulation checks regarding treatment information (Montgomery, Nyhan, and Torres 2018). So, I omit the attentiveness as a moderator in the main analysis. Appendix S includes an analysis that included attentiveness as a moderator, while acknowledging the inherent limitations thereof.

Unlike in the pilot study, the political sophistication analysis offers no more support to Hypotheses 1-2 than the whole-sample analysis did (Figure 17). All three impacts become more significant, with ATE and ACDE moving to the right and EE to the left; but they ultimately remain insignificant, as the lowest p-value among them is 0.22 (ACDE). The subsample sizes for all the constituent regressions are small, with 98, 82, and 180 respondents respectively. But, as the results stand, they evidence no impact of Aurl's polarization on respondent willingness to accept Aurl's offer—neither directly nor via perceptions of their unreliability. This raises a question the next section addresses: whether polarization increased either of the two anticipated causal pathways (reduced leader accountability, increased chance of defection via government changeover).

reliability, as its Cronbach's alpha was 0.5247. Separating the political knowledge questions into the domestic and foreign policy domains (e.g., House majority and veto override; U.K. prime minister, Taiwan, and UNSC members) did not improve the inter-reliability, nor did various other permutations of the five questions. I did not include education in the political sophistication measure because it dropped Cronbach's alpha to 0.7107.

Figure 17. Est. Impacts of Polariz. on Offer Acceptance (Odds Ratios, Subsample)



ATE: avg. treatm. effect; ACDE: average direct controlled effect; EE: eliminated effect
Coefficients are logarithmically transformed to facilitate ease of visual interpretation
Model: logistic

The whole-sample and sophisticated subsample results may also suggest differences between the full fielding sample and pilot study's sample of undergraduates. In both studies, focusing on political sophistication moved the whole-sample's results in the direction of greater significance and even near-significance (in the case of the pilot study). However, those directions differed between the studies. ACDE was positive in the pilot study and negative in the full fielding. Similarly, EE in the pilot study was nearly significant and negative, while it was positive and less significant in the full fielding.

Part of this difference could be due to the scenario alterations I made after the pilot study to decrease its slant toward Aurl offer acceptance. Another part of it could be that the pilot sample's understanding of polarization was different than the full fielding sample's. Undergraduates enrolled in politics/politics-adjacent courses are more likely to

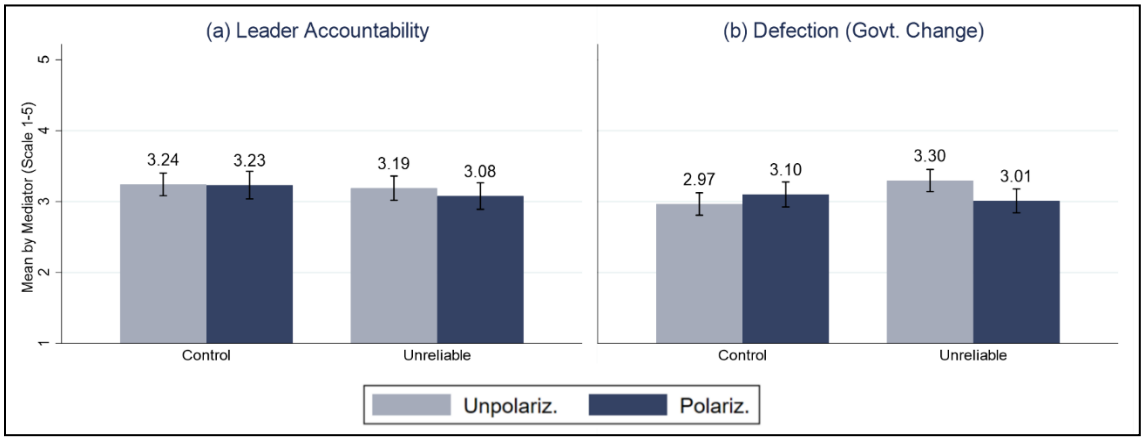
be exposed to contemporary scholarly theoretical discussions of American polarization and its impacts. This seems particularly true for those students interested enough to major or consider majoring in related fields, that is, for the pilot study subsample that showed impacts most consistent with my expectations. If that is the case, the pilot study's stronger support for my expectations could be an echo of classroom discussions about the theoretical dynamics that motivated this study in the first place. The full fielding sample, in contrast, likely was less exposed to theories about how and why polarization should impact politics.

Causal Logic: Leader Accountability and Government Changeover

The previous section raised the question of whether polarization impacts either of my proposed causal logic pathways. It does not, though some of the near-significant results could have had their significance suppressed by the third of inattentive respondents and the scenario's slant toward Aurl cooperation. Interestingly, polarization does impact perceptions of Aurl's defection probability in general at statistically significant levels in a whole-sample analysis.

In my argument, a polarized state has higher defection probability because of weakened accountability dynamics and increased probability that government changeover means policy changeover. Figure 18 supports neither hypothesis. Respondents' mean perceived probability of a defecting leader being held accountable by their supporters does not differ by polarization treatment in the group told nothing about Aurl's unreliability (Figure 18a, 3.24 v. 3.23). That is, respondents do not appear to have inferred that polarization means lower chance a leader will be held accountable for defecting from Aurl's promise.

Figure 18. Means: Gov't Changeover, Accountability by Treatment Group



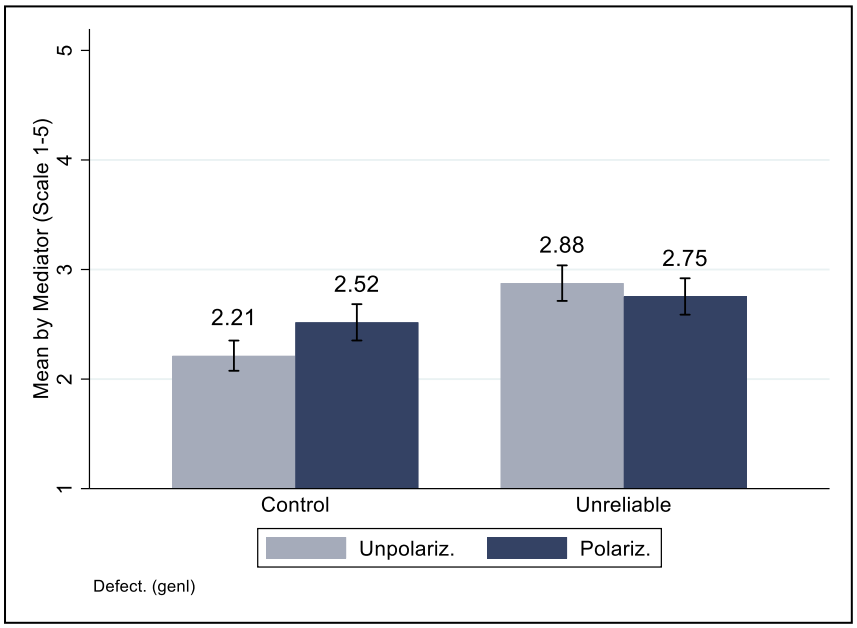
The whole-sample means are similarly unsupportive of Hypothesis 3b (Figure 18b). In the "control" group for unreliability, respondents told Aurl was polarized assigned higher mean probability of Aurl defecting in case of government changeover than respondents told Aurl was unpolarized. (3.10 v. 2.97). However, the difference did not reach standard levels of significance ($p = 0.26$). Unsurprisingly, there was a significant difference-in-means among respondents told that Aurl was unpolarized, with the "control" group assigning lower mean defection likelihood than the group told that Aurl was unreliable (2.97 v. 3.30, $p = 0.00$).

Somewhat surprisingly, that pattern was not mirrored among respondents told Aurl was polarized. The means of those told/not told that Aurl was unreliable did not differ significantly, (3.01 v. 3.10, $p = 0.47$). In isolation, one could interpret this as evidence that polarization increases concern of defection via government changeover, as respondents in the polarized condition had the same offer acceptance rates regardless of if they were told about Aurl's unreliability. However, this is improbable given the reality

there was no difference in offer acceptance by polarization status among those told nothing of Aurl's unreliability.

If polarization has no impact on offer acceptance via unreliability (contra Hypotheses 1-2), and no impact on the two anticipated causal elements within unreliability (contra Hypothesis 3a-b), then polarization may have no impact on defection perceptions at all. Figure 19, though, indicates this is not the case. In the "control" condition, respondents told Aurl was polarized had a higher mean perception of Aurl defection probability than those told Aurl was unpolarized (2.51 v. 2.24, p = 0.01). This suggests that respondents inferred from Aurl's polarization that Aurl had higher chances of defection, albeit only modestly higher.

Figure 19. Means: General Defection Probability by Treatment Group

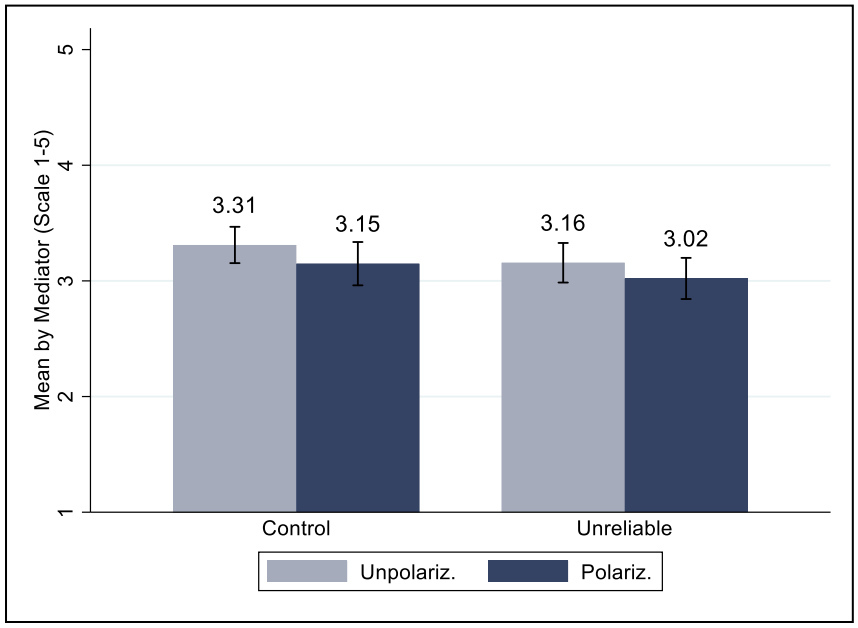


Importantly, this modest gain was in the context of Aurl's overall perceived defection chances being low. This was true even in the groups told Aurl was unreliable. The highest mean probability of defection (2.88) falls on the scale between "somewhat

unlikely" and "neither likely nor unlikely": hardly a high level of concern. That, combined with the sample's high rate of offer acceptance even when told Aurl was sometimes unreliable, strongly suggests the scenario remained slanted toward Aurl offer acceptance.

The subjective manipulation check results permit me to test that inference. The check asked how likely respondents thought a Zerm invasion was if their country rejected Aurl's offer. As shown in Figure 20, all four groups assigned a mean somewhere between "neither likely nor unlikely" (3) and "somewhat likely to invade" (4). These means are not remarkably high, but they are higher than respondents' mean perception that Aurl would defect. That could account for why most respondents preferred acceptance. They judged a Zerm invasion in case of offer rejection was more probable than Aurl defection in case of offer acceptance.

Figure 20. Means: Prob. Zerm Invades if Reject Aurl Offer



A remaining implication to be explored is what accounts for defection concern if not Hypothesis 3a-b. If polarization has null results on leader accountability and defection via government changeover, then what accounts for the increased concern of defection in Figure 19? There are two explanations I can test. The first is that polarization perhaps influences perceptions of coethnicity. Coethnicity theoretically could impact defection concern, as one is more likely to trust coethnics than non-coethnics. The second explanation is that the scenario unequally prompted real-world connections across the treatment groups. If that occurred, respondents may have asymmetrically imported real-world background information across the groups, which in turn could impact Aurl defection probabilities or Zerm invasion probabilities.

Ordinal logistic regressions analyses do not strongly support the coethnicity explanation (Table 25). The analyses—conducted on the five coethnicity variables and an index variable thereof (alpha = 0.8227)—show an asymmetry in perceived coethnicity by treatment group but not in a way that mirrors the groups' relative defection concern. To be sure, the three groups in Figure 19 with higher defection chances than group 00 also show lower perceived coethnicity by several measures in Table 25. And, the more particularized of those measures (likeness in values/beliefs and goals) plausibly could impact defection concerns. However, the relative magnitude of the odds ratios by group do not match the relative magnitude of general defection perceptions in Figure 19.

In Figure 19, the group with the *smallest* difference in defection concern from Group 00 is Group 10, those in the "control" condition told that Aurl was polarized. In Table 25, however, Group 10 is the group that shows the *greatest* impact for two of the four variables (index, values/beliefs) relative to Group 00. Further, its odds for the other

Table 25. Est. Impact of Treatm. Group on Aurl Coethnicity Perceptions (Odds Ratios)

VARIABLES	(1) Index	(2) Overall	(3) Values/beliefs	(4) Appearance	(5) Language	(6) Goals
01 (unpol, unrel)	0.864	0.731*	0.691*	1.058	0.960	0.705*
	(0.153)	(0.139)	(0.132)	(0.196)	(0.177)	(0.134)
10 (pol, om)	0.677**	0.702*	0.532***	1.000	0.748	0.807
	(0.122)	(0.138)	(0.104)	(0.191)	(0.141)	(0.159)
11 (pol, unrel)	0.708**	0.647**	0.580***	1.077	0.831	0.619**
	(0.124)	(0.122)	(0.109)	(0.197)	(0.150)	(0.116)
Observations	750	750	750	750	750	750

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

two variables are quite similar in magnitude to those of the other two groups. This relative similarity in co-ethnicity perceptions among the three groups strongly contrasts with Figure 19, where Groups 01 and 11 had much stronger perceptions of Aurl defection probability than Group 10. Overall, this suggests that if perceived coethnicity accounts for general defection chances instead of government changeover or leader accountability, it only accounts for a minority of it. These results are robust to an alternate specification that uses a *polarization* × *unreliability* interaction (Appendix T).

The second alternative explanation of polarization's impact on general defection chances is unsupported (Table 26). Only twice was information equivalence violated, and neither violation suggests major importation of background information that would alter respondents' perceptions regarding Aurl defection. Group 10 had increased odds of thinking Aurl was a former communist country (68.4%, $p < 0.01$) and Group 11 had increased odds of thinking Aurl was in North America (98.1%, $p < 0.05$). However, Group 10's association of Aurl with a former communist country was not associated with a particular country, as they were no more likely to think Aurl was on a particular

continent than the reference group. Similarly, Group 11's association of Aurl with North America suggests they thought of the United States, but their non-identification of Aurl as majority Caucasian or the United States later in the survey suggests the U.S.-Aurl association was weak. Finally, the results for Group 01 also suggest information equivalence violations are not responsible for polarization's apparent impact on general defection chances. Group 01 had the highest mean perceived probability of Aurl defection, and yet were no more likely than any of the other groups to think Aurl was more likely to be majority Christian, majority Caucasian, a former communist country, or in a particular continent. It also was no more likely to say Aurl and Zerm reminded them of actors closely associated with the Russian/Ukraine war. As before, all these results are robust to an alternate specification that uses a *polarization* × *unreliability* interaction (Appendix T).

In sum, this study finds no support for Hypotheses 3a-b, as polarization neither decreased perceptions that Aurl's leader would be held accountable in case of defection, nor increased perception that an Aurl government changeover increased risk of defection.

Table 26. Est. Impact of Treatm. Group on Aurl's Real-World Traits (Odds Ratios)

VARIABLES	(1) Maj. Christ.	(2) Maj. Cauc.	(3) Fmr. comm.	(4) Remind U.S.	(5) Remind Rus./ U.S.	--
01 (unpol/unrel)	0.881	0.773	1.187	1.075	1.237	--
	(0.165)	(0.142)	(0.217)	(0.372)	(0.369)	
10 (pol/om)	1.041	0.804	1.684***	1.368	1.386	--
	(0.200)	(0.152)	(0.319)	(0.454)	(0.415)	
11 (pol/unrel)	0.888	0.794	1.294	1.486	1.403	--
	(0.168)	(0.145)	(0.234)	(0.477)	(0.407)	
Observations	750	750	750	750	750	--

MODELS CONT'D (Table 26)	(6) Africa	(7) Asia	(8) Austr.	(9) Eur.	(10) N. Am.	(11) S. Am.
01 (unpol/unrel)	1.251	1.191	2.254	0.986	1.127	0.547
	(0.589)	(0.393)	(1.966)	(0.204)	(0.385)	(0.340)
10 (pol/om)	1.836	1.329	1.229	0.907	1.036	0.298
	(0.820)	(0.440)	(1.236)	(0.194)	(0.369)	(0.238)
11 (pol/unrel)	2.001	1.337	0.537	0.762	1.981**	0.529
	(0.859)	(0.429)	(0.660)	(0.159)	(0.612)	(0.329)
Observations	750	750	750	750	750	750

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

However, it also shows that polarization did increase respondents' perceived probability of Aurl defecting in general. The higher perceived risk of a Zerm invasion appears to have overwhelmed that modest increase in defection risk. As polarization thus had no impact on the two anticipated causal pathways for voluntary defection concern, this raised the question of what else could account for polarization's positive impact on general defection risk. I tested the possibility that polarization could perhaps impact coethnicity perceptions, which theoretically could impact defection concern, but found little support. Information equivalence was similarly unresponsible, as the violations of information equivalence among the groups were infrequent and theoretically insignificant.

I am thus left with the question of what accounts for polarization's impact on perceived defection probability in general. The theoretical smoking gun I cannot account for is whether Aurl being polarized increased perceived probability of involuntary defection due to gridlock. I focused on voluntary defection, but as both my anticipated

causal pathways were ruled out, gridlock is a leading candidate. My results, however, come with three caveats.

The first is significance suppression due to respondent inattentiveness, as approximately 30% of the sample missed at least one of the treatment items. I cannot assess with high confidence whether this inattention had asymmetrical effects across sample groups, as it was a post-treatment measure and therefore subject to post-treatment bias. However, if post-treatment bias was inoperative, the distribution of inattention by treatment group suggests that inattention's effects would most impact the ACDE measure. That in turn would also render the EE measure more error-laden and thus less significant. At the very least, I can say with confidence that 30% of the respondents engaging with a different scenario than the one they were presented would introduce error into the outcome measurements.

The second caveat is that question framing conditions the results for the three causal logic questions (defection/accountability) (Appendix T). Respondents asked the probability that Aurl would "keep" its promise had means akin to those in this analysis, though the leader accountability/government changeover outcomes approached standard levels of significance. However, those asked the probability that Aurl would "break" its promise had no statistically significant impact of polarization on any of the three causal logic outcomes. It approached significance for general defection concern and leader accountability ($p = 0.22$-0.23); however, leader accountability was in the wrong direction. Overall, though, these differences could be due to some underlying imbalance in the two framing groups. Despite random assignment, the two groups have statistically significant differences-in-means for Aurl offer acceptance, a question asked before they encountered

the keep/break framing (+7%, p = 0.04). The framing effect results must be taken with a grain of salt.

The third caveat is that my results are not robust to subsample analysis by political sophistication. Unlike the whole sample, the subsample of those above the 75th percentile in political sophistication had no impacts from polarization to perceived defection probability in general. Also unlike the sample on average, polarization impacted their perceived probability of defection-by-government-changeover and leader accountability. The difference was in the expected direction for only government changeover, and both differences were statistically insignificant. These mixed results raise the continuing question of what role political sophistication plays in one's assessments of polarization's implications, but again were small-*n* and thus only suggestive.

Method: Experiment 2 (Affective v. Policy Polarization)

Design

The survey experiment to test Hypothesis 4 was a partial replication of Myrick's (2022) study. Thus, it had two treatment groups: a control group exposed to Myrick's original affective polarization treatment and a treatment group exposed to a modification thereof (Table 27). The modification was threefold. First, I abbreviated the survey, omitting all portions save those of theoretical relevance to this paper (affective polarization, unreliability, and cooperation preference). Second, I replaced two of the six items Myrick used to describe affective polarization with two of the items from her operationalization of policy polarization. The two new items were that Americans "[t]hink their political parties cannot agree on basic facts" and that American politicians "[v]ote the same way

as members of their own parties." I selected them because they include social/group dimensions that make them suitable to affective polarization and I qualitatively judged them to convey policy implications more strongly than the replaced two affective polarization items.

Table 27. Affective Polarization Operationalizations by Group

Modified Survey	**Original Survey (Myrick 2022)**
Surveys from *Country X* show that, more than ever, people in *Country X*: • Oppose the idea of their child marrying someone from the other political party. • Have 'just a few' or 'no' close friends from the other political party. • *Think their political parties cannot agree on basic facts.*	Surveys from the *United States* show that, more than ever, *Americans*: • Oppose the idea of their child marrying someone from the other political party. • Have 'just a few' or 'no' close friends from the other political party. • '*Strongly dislike' or even 'hate' members of the other political party.*
These differences are reflected in *Country X's* government. More than ever, politicians *from Country X's two major political parties*: • Use extreme, negative language to taunt politicians of the other party. • *Vote the same way as members of their own political party.*	These differences are reflected in *the US* government. More than ever, *Republican and Democratic* politicians: • Use extreme, negative language to taunt politicians of the other party. • *Post angry or hateful posts on social media about members of the other party.*

The third modification was anonymizing the partner state as "Country X." This change was necessary because my respondents were largely American rather than from the United Kingdom, so I could not query their opinions about their country's relationship with the United States. I used "Country X" as the partner state instead of a

named U.S. partner as the latter would import background context into the scenario that was different from the original. It also would require more extensive modification of the treatment's language, as what is true of polarization in the United States is not true everywhere. An extended discussion of the costs and benefits of the Country X approach and a table showing the original/modified wordings are in Appendix U.

Fielding

The Country X experiment was conducted at the same time and with the same undergraduate sample as the Aurl/Zerm pilot study. The order of the two experiments was randomized, with a distraction task between them. The random ordering of the experiments provided me a pre-treatment proxy for attentiveness, as presumably the respondents who took the Country X scenario first would be less cognitively taxed than those who encountered it after the Aurl/Zerm scenario.[93] The distraction task consisted of demographic data: international student status, gender, race/ethnicity, and whether respondents were majoring or thinking about majoring in politics/global studies. Whether the distraction task was before the survey or not, the survey introduced the experiment with the following language:

> *We are interested in understanding your attitudes toward Country X, a historically close military and economic ally of your country that we have*

[93] One could also wonder if respondents would be more attentive if they took the Country X scenario second rather than first, as one's attention level could increase as one "warms up" in a survey. However, I did not find that to be the case in testing the survey.

anonymized. We will first provide you with some information about its politics and then ask you for your opinions.

Please read this information carefully.

It then presented the treatment information in Table 27, column 1. Interspersed with that information were two factual manipulation checks, identical in location and language to those of the original study save for the Country X name swap. After treatment exposure and manipulation checks, the survey asked respondents' agreement with four statements (7-point scale, "strongly disagree" to "strongly agree"). The first two were Myrick's indicators for current perceived reliability and the latter two for future perceived reliability:

Country X would come to the aid of my country in the event our security is threatened

Country X no longer maintains its commitments to foreign countries

My country should partner with Country X in future international agreements

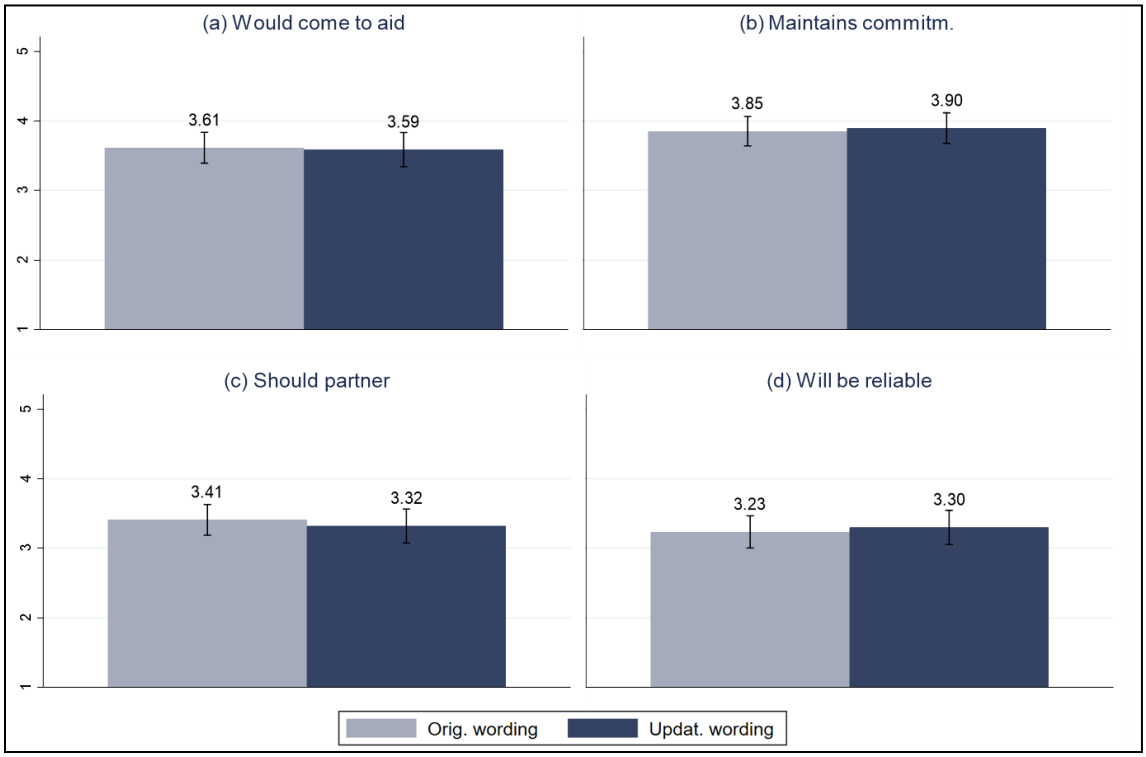
Country X will not be a reliable future partner for my country

After capturing the four outcomes, I asked Myrick's final manipulation check, a subjective one querying respondents' perceptions regarding how often Country X's political parties agree (4-point scale, from "almost never" to "almost always", as well as a fifth, "don't know" option). The experiment concluded with the placebo question batteries used for the Aurl/Zerm pilot study. Respondents either then proceeded to the Aurl/Zerm experiment via the distraction task questions or were shown a debriefing statement after a series of unrelated questions for a separate study.

Results: Experiment 2

Neither descriptive nor regression results support Hypothesis 4. As shown in Figure 21, none of the mean reliability and cooperation preferences differ between the control and treatment groups. For all four outcomes, the means are tightly-ranged, with 0.09 on a scale of 1-5 being the largest difference (Figure 21c). The updated wording did not change the outcomes.

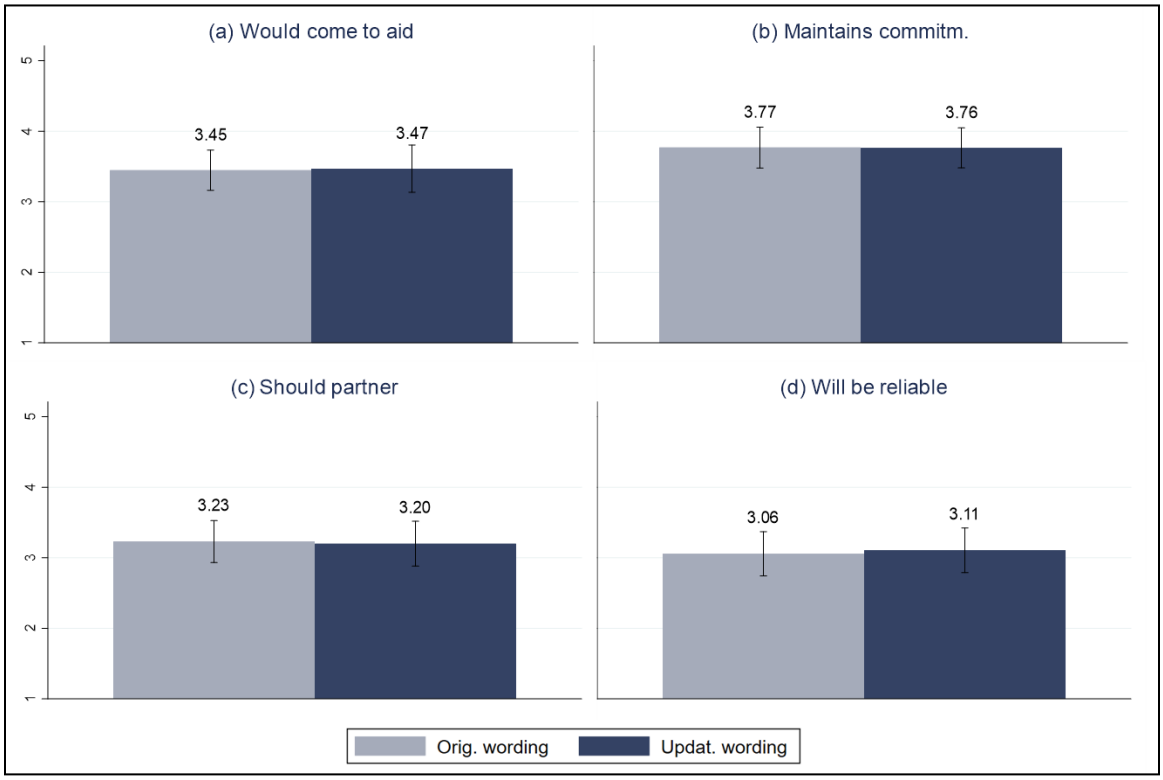
Figure 21. Means: Reliability and Cooperation Preference Outcomes



Unlike in the Aurl/Zerm experiment, focusing the analysis on the subsample of 172 respondents who were most attentive and most likely to have greater understanding of politics (politics/politics-adjacent majors) did not reveal significant effects (Figure 22).[94] The subsample's means for the four outcomes are even more tightly-arrayed, with the greatest difference being only 0.05 (Figure 22d). Indeed, the updated treatment's non-impact across all four outcomes is robust to three different subsample permutations involving academic major and scenario order (Appendix V). The result is null, and highly likely not due to some hidden theoretical or design-induced conditionality that masks the anticipated result.

[94] "Most attentive" was operationalized as respondents who correctly identified all treatment information/control group equivalent information items in the two factual manipulation checks.

Figure 22. Means: Reliability and Cooperation Preference Outcomes (Subsample)



The subjective manipulation check perhaps explains why: the updated treatment I had anticipated would be stronger than the original treatment was actually weaker (Figure 23). Both groups assigned an average rating of Country X's two parties agreeing somewhere between "almost never" and "rarely," but the treatment group had a marginally higher perception of party agreement frequency in Country X (1.68 v. 1.50, p = 0.05). Unlike the four main outcomes, however, robustness checks by subsample indicate that political sophistication and attentiveness could moderate these results (Table 28).

Figure 23. Means: Perception of Party Agreement Frequency

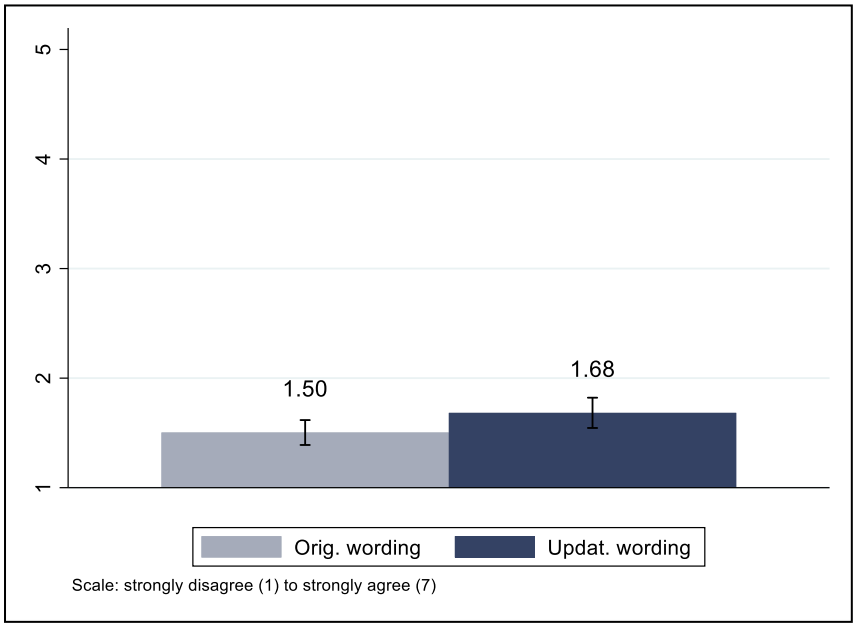


Table 28: Estimated Impact of Treatment on Perceived Partisan Agreement (OLS)

	(1) Whole Sample	(2) X Scen. First	(3) Major	(4) X First & Major
Treatment	0.180	0.390	0.064	-0.193
	(0.091)*	(3.17)**	(0.64)	(1.29)
Constant	1.503	1.418	1.550	1.711
	(0.064)**	(16.59)**	(21.83)**	(15.37)**
R^2	0.01	0.06	0.00	0.02
N	285	152	218	101

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

Attentiveness by itself unexpectedly amplified the results found in the whole-sample scenario: respondents who encountered the Country X experiment before the Aurl/Zerm pilot had higher perceptions of Country X partisan agreement than the whole-sample did (column 2). However, among the even smaller subsample of politics-related majors who encountered the Country X

scenario first, the treatment registered a negative impact on perception of partisan agreement (-0.193, p = 0.199, column 4). This impact only approaches statistical significance but is also a sample of only 101 respondents. A larger sample analysis could test whether political sophistication is a moderator (provided attentiveness).

Discussion

On one hand, this study suggests that Myrick's finding about affective polarization stands. Affective polarization does not decrease partner perceptions of a state's future reliability or willingness to cooperate with that state, as my alternate affective polarization operationalization had no stronger an impact on those outcomes than the previous operationalization. On the other hand, this study also suggests that Myrick's finding about policy polarization did not transfer to my scenario study using a U.S. sample: the fictional state of Aurl being policy and affectively polarized had no net, direct, or indirect/interaction effect on respondent willingness to cooperate with them. These two strands could be taken to mean that polarization does not encourage commitment problems—regardless of its affective or policy aspects. However, the study's implications are more modest, and do not support so significant a takeaway.

First, despite its null results, the study suggests that Myrick's findings regarding policy polarization's impacts on unreliability perceptions are generalizable beyond the U.S. context. To be sure, polarization had no statistically significant impact on offer acceptance, and respondents did not seem to infer high levels of unreliability from knowledge that Aurl was polarized. However, polarization did impact respondents' perceptions of Aurl's general chances of defection. And, it did so among U.S. rather than

U.K. respondents about polarization that was not that of the United States. Myrick's other finding—U.S. polarization decreasing U.K. respondents' preference to cooperate with the United States in the future—is not evidenced in this study. But, that is less likely due to some unique property of U.S. polarization and more likely because of the issue with inattentiveness combined with the much larger issue of scenario slant toward accepting Aurl's offer.

Second and relatedly, the study's results highlight that unreliability is not a sufficient condition to induce a commitment problem. Like in Myrick's study, respondents updated at least general unreliability assessments, as they assessed that Aurl's general defection chances increased. And they also may have updated their general cooperation preferences, as Myrick's respondents did. Non-cooperation, though, did not follow from those general assessments, as a majority chose to accept Aurl's offer anyway. It was a matter of magnitude. For non-cooperation to follow from unreliability, the magnitude of polarization's unreliability impacts must be large enough that the risk posed by non-cooperation seems relatively lower. As respondents assessed a Zerm invasion in case of offer rejection was more likely than an Aurl defection in case of offer acceptance, the scenario did not meet that threshold.

Third, the study suggests the general public links polarization with unreliability but raises the question of why. The two proposed institutional reasons—decreased leader accountability and government changeover risk—were statistically unconnected with polarization in this study. One reason for this could be that those two proposed causal pathways are indeed inoperative. The non-existence of weakened leader accountability is consistent with Broockman et al.'s findings regarding affective polarization's null

impacts on Americans' vote choice. The non-existence of a link between government changeover and policy changeover in a polarized state is less plausible, but perhaps could indicate that polarization most impacts domestic and not foreign policy. Another explanation (and the most I can say with these results) is that my full fielding respondents did not seem to *think* polarization did either of those things.

This lack of an apparent causal mechanism—especially considering the insignificance of the coethnicity results—raises the question of why my respondents inferred that polarization meant increased general defection chances. I had theoretically dismissed the role of polarization's impact on involuntary defection chances, particularly through legislative gridlock. That dismissal appears premature. Further, soft power considerations could be in play. Myrick discussed potential impacts to international favorability toward the United States as potentially impacting its international attractiveness. The coethnicity measures roughly proxied for favorability, as it tested for coincidence in values/beliefs and goals and most people presumably do not have low opinions of their own values/beliefs and goals. However, a more tailored measure of favorability would be appropriate.

Another factor potentially at play in the study's silence on causal logic is difference between general public and foreign policy practitioner reasoning about polarization's impacts. The full fielding sample had null effects of polarization on prospects of leader accountability or government changeover defection chances. The pilot study, however, diverged with regard to leader accountability among its subsample of students majoring/interested in majoring in politics-related subjects. This divergence could be due to the scenario changes I made to slant the scenario less toward Aurl

cooperation. However, it also could be because undergraduate students enrolled in politics-related courses are more aware of polarization's theoretical impacts, especially those with enough interest in politics-related subjects to major in them. If that is the case, future research may find that the causal pathway from polarization perception to increased general defection chances may have different stops along the way for the public and for policy practitioners.

A third dynamic at play in the causal logic silence is a potential confounding effect of my design on the government changeover measure. I presented the government changeover in narrative form, rather than hypothetical, to reduce cognitive load on participants. However, in doing so, I may have provided an additional data point that bolstered Aurl's reputation for reliability in the polarization condition. The scenario noted that "After some time passes, Aurl's government changes hands from its current administration to the opposing side. Everything else in the scenario remains the same." The reference to time could have created the perception in people's minds that Aurl's existing government had managed to maintain the policy in the face of partisan opposition, and so Aurl likely would maintain the policy even through the government changeover, particularly given the overarching policy area is one that Aurl's parties could agree on: Zerm is bad.

A final implication of the study is that people may indeed have less concern with affective polarization than policy polarization with regard to its impact on partner defection. That suggests that if a foreign audience primarily hears via elites and media of a state's affective vice policy polarization, there likely will be no net impact on perceptions of unreliability and cooperation preference. On the other hand, if foreign

audiences hear of another state's policy polarization, to include a general state of high politicization that even touches foreign policy domains, they will be more skeptical of that state's reliability and more cautious of partnering with that state. However, as discussed at the beginning of this section, potential unreliability does not a commitment problem make.

Two caveats on this finding are that I did not test the relative magnitude of affective polarization overall against some other condition, and that my sample may meaningfully differ from Myrick's. Limited sample size precluded adding a third or even fourth treatment condition that would provide a true control group and/or replicate Myrick's policy polarization treatment. So, my discussion assumes my sample would have shown that policy polarization was more impactful than affective polarization, and that affective polarization's impacts were indistinguishable from that of a control group. If one or both of those assumptions were invalid, it would suggest my sample meaningfully differed from Myrick's. That is plausible, as undergraduate Americans enrolled in politics-related courses could differ from her quota sample of U.K. adults. The students perhaps were more likely to infer policy implications from the original treatment, which would decrease the anticipated strength difference between the new and old treatments.[95] My results thus suggest that affective polarization's null result was

[95] On the other hand, the subsample of attentive, politics-related majors found the updated treatment wording to more strongly convey partisan disagreement than the original treatment wording. As they are hypothetically the most sophisticated in the sample with respect to polarization, this suggests that sophistication did not increase the undergraduates' ability to infer policy implications from affective polarization. However, there also could be a positive interaction effect, in which being a major permits the undergraduates to infer even stronger implications from the updated information provided than the implications they inferred from the original treatment.

robust to a different operationalization; but greater confidence would demand I replicate the original policy polarization and control conditions alongside the two I tested.

A related limitation of my findings is that the full fielding sample also was not a representative sample. It was a quota sample, demographically matched in terms of age, gender, ethnicity, race, and age to the 2020 U.S. census. However, the participants were not randomly selected. They opted-in to take surveys online, and further opted-in to take a survey advertised as being about politics. The latter selection bias particularly could have resulted in an unrepresentative sample in terms of political knowledge, interest, and sophistication. If that is the case, the significance of the null results regarding polarization and unreliability could be overstated: not insignificant enough. This would underscore rather than change the results, though.

A final limitation of my study is that the design precludes disaggregation of polarization's interaction and indirect effect on cooperation preference via unreliability. As discussed, the scenario's evident slant toward Aurl cooperation may have overwhelmed considerations of Aurl's defection probability. However, even if it had not done so entirely, and polarization *did* minorly decrease cooperation preference via perceptions of unreliability, I may not have been able to demonstrate it due to the additive interference of an interaction effect. A sufficiently noisy or positive interaction effect would mask the negative direction and/or significance of the hypothesized indirect effect. A positive interaction effect is plausible—the unreliability of a polarized state could be less concerning than the unreliability of an unpolarized state, which presumably has consensus behind its unreliability. However, and to return to the beginning, I cannot test

that inference without making assumptions about confounding variables that I cannot support.

Conclusion

Do (1) perceptions of another state's domestic polarization influence (2) assessments of that state's international reliability and thereby (3) shape willingness to cooperate with that state? This study says yes to the first and no to the second. Polarization does impact public concern that the polarized state is unreliable in terms of general defection chances. However, and as shown in this study, general defection concern does not always translate to a commitment problem. Whatever level of unreliability is conveyed by polarization, the level is less than "sometimes does not follow through on its international promises." And, when in a context that partner defection seems less probable and less risky than non-cooperation's outcome, the modest increase in general defection concern is moot.

The study's "no" to the question of whether polarization shapes cooperation preference is thus highly conditional. Theoretically, polarization maintains a relationship with cooperation preference, as it impacts defection concern and such concern is hardly immaterial to cooperation. However, the conditions in which polarization's modest impact actually translates to "cooperate" versus "don't cooperate" will be limited. Polarization will only be the tipping point if the perceived likelihood of non-cooperation going poorly is quite close to the perceived likelihood of defection. The study therefore points to polarization as but one factor among others that impact cooperation choices.

The results also suggest that both the public and more politically-sophisticated populations factor polarization into their defection assessments. But they perhaps do so for different reasons. The study indicates with low confidence that political

sophisticates—particularly those with greater knowledge of polarization and its impacts—infer that polarization weakens leader accountability. But intriguingly, the study cannot account for why the most general of my samples inferred that polarization increased defection chances. They showed no evidence of the potential causal logics of weakened leader accountability, risk of government changeover, or perceived coethnicity.

In terms of the "missing" factors the public appears to be considering, government changeover concern cannot be dismissed despite its null effects. The study's narrative framing of the government changeover question may have induced spurious results by encouraging perceptions the partner state had demonstrated their reliability. Two other candidate explanations are gridlock or weakened soft power. Overall, though, these open questions highlight that how polarization *should* impact defection chances and how people *think* polarization impacts defection chances may be different things.

CHAPTER 5

CONCLUSION

This dissertation offers a novel theoretical accounting of why, how, and to what effect states use disinformation in their foreign policies. The why is the same "why" that two-level games suggest for any foreign policy tool: to "get more" or at least "get something", "keep what you have" and/or "keep the other guy from getting something." The "how" is using message and attributional disinformation to modify (1) opponent win-sets and uncertainty and (2) opponent perceptions of one's own win-set and uncertainty. More granular methods are plentiful, all involving the manipulation of preference change or preference mobilization/formation; politicization; international good will/popularity; and/or chief negotiator popularity, credibility, or autonomy.

Those manipulation points are the locus of disinformation's effects. Some of the effects can pertain directly to a discrete policy goal (e.g., masking a particular status quo), and others such as polarization or attacking soft power (good will/popularity) have more long-term effects and can be simultaneously pursued—"two birds with one stone." Importantly, the two-level contextualization highlights that general and more-focused effects can be pursued among a wider audience than just an opponent's Level I players. Level II impacts can reverberate into the Level I game through responsiveness mechanisms.

The two-level account of disinformation in foreign policy generates several testable implications, of which the dissertation tested two. One of the implications—that disinformation can pursue impacts through multiple pathways simultaneously—receives suggestive support. Though the analysis was majorly limited by probable floor effects, a

small sample size due to disaggregation, and the possibility of post-treatment bias, the results suggested that political disinformation may in fact impact not only those that believe it, but those that don't. Significant or near-significant impacts were seen on disinformation belief, and (depending on partisanship and the disinformation's partisan congeniality) generalized trust, anger, and fear. These findings are consistent with the inference that disinformation can encourage sociopolitical "ripple effects" beyond belief that could damage a state's soft power and potentially increase leader autonomy. While I found no direct evidence that political disinformation increases affective polarization, the possibility of floor effects suggests that disinformation in a polarized state is more likely to contributes to the maintenance of high polarization levels than raise them.

The second implication—that disinformation can encourage a commitment problem by encouraging polarization—also receives mixed support. As mentioned, I found no evidence that political disinformation increased affective polarization. However, that could be due to floor effects, and suggestive evidence was found of negative reasoning regarding one's outgroup party when one was exposed to and disbelieved disinformation congenial to the outgroup party. This suggests disinformation can impact group assessments—presumably to include the group like assessments inherent to affective polarization.

Moving to the later links in the disinformation-commitment problem chain: the dissertation also shows a partner state's polarization does modestly increase public concern of its defection. This is consistent with the two-level account's commitment problem implication. Contra that implication, though, the public does not seem to

consider the prospect of weakened leader accountability in its defection assessment (though political sophisticates might).

This suggests refinement in the area of elite v. public perceptions. How polarization actually increases defection chances may not be the reason the public thinks it increases defection chances. Future work can seek to identify the reasons why the public infers increased defection chances from another state's polarization. It also could test if those reasons differ from those of foreign policy elites, who may be more aware of polarization's institutional impacts.

REFERENCES

Abramowitz, Alan. 2010. *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy*. Yale University Press.

Abrams, Dominic, and Michael A. Hogg. 1990. *Social Identifications: A Social Psychology of Intergroup Relations and Group Processes*. Florence, UNITED STATES: Taylor & Francis Group. http://ebookcentral.proquest.com/lib/asulib-ebooks/detail.action?docID=178266.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2018. "Analyzing Causal Mechanisms in Survey Experiments." *Political Analysis* 26 (4): 357–78. https://doi.org/10.1017/pan.2018.19.

Aldrich, John H., Christopher Gelpi, Peter Feaver, Jason Reifler, and Kristin Thompson Sharp. 2006. "Foreign Policy and the Electoral Connection." *Annual Review of Political Science* 9 (1): 477–502. https://doi.org/10.1146/annurev.polisci.9.111605.105008.

Aldrich, John H, David W Rohde, Lawrence C. Dodd, and Bruce Oppenheimer. 2001. "The Logic of Conditional Party Government: Revisiting the Electoral Connection." In *Congress Reconsidered*, 7th ed. Washington DC: CQ Press.

Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2020. "Evaluating the Fake News Problem at the Scale of the Information Ecosystem." *Science Advances* 6 (14): eaay3539.

Al-Rawi, Ahmed, and Anis Rahman. 2020. "Manufacturing Rage: The Russian Internet Research Agency's Political Astroturfing on Social Media." *First Monday*.

Altay, Sacha, Emma de Araujo, and Hugo Mercier. 2021. "'If This Account Is True, It Is Most Enormously Wonderful': Interestingness-If-True and the Sharing of True and False News." *Digital Journalism*, 1–22.

ANES. 2021. "ANES 2020 Time Series Study Full Release User Guide and Codebook."

Anzera, Giuseppe, and Alexandra Massa. 2021. "Using International Relations Theories to Understand Disinformation: Soft Power, Narrative Turns, and New Wars." In *Politics of Disinformation*, 35–49. John Wiley & Sons.

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. 2019. "A Note on Dropping Experimental Subjects Who Fail a Manipulation Check." *Political Analysis* 27 (4): 572–89. https://doi.org/10.1017/pan.2019.5.

Bachleda, Sarah, Fabian G. Neuner, Stuart Soroka, Lauren Guggenheim, Patrick Fournier, and Elin Naurin. 2020. "Individual-Level Differences in Negativity Biases in News Selection." *Personality and Individual Differences* 155: 109675.

Bailey, Michael A., Anton Strezhnev, and Erik Voeten. 2017. “Estimating Dynamic State Preferences from United Nations Voting Data.” *Journal of Conflict Resolution* 61 (2): 430–56.

Barabas, Jason, Jennifer Jerit, William Pollock, and Carlisle Rainey. 2014. “The Question(s) of Political Knowledge.” *American Political Science Review* 108 (4): 840–55. https://doi.org/10.1017/S0003055414000392.

Barlow, Fiona Kate, Stefania Paolini, Anne Pedersen, Matthew J. Hornsey, Helena RM Radke, Jake Harwood, Mark Rubin, and Chris G. Sibley. 2012. “The Contact Caveat: Negative Contact Predicts Increased Prejudice More than Positive Contact Predicts Reduced Prejudice.” *Personality and Social Psychology Bulletin* 38 (12): 1629–43.

Baron, Reuben M., and David A. Kenny. 1986. “The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations.” *Journal of Personality and Social Psychology* 51 (6): 1173.

Bauer, Paul C., and Markus Freitag. 2018. “Measuring Trust.” *The Oxford Handbook of Social and Political Trust* 15.

Baum, K., S. Meissner, and H. Krasnova. 2021. “Partisan Self-Interest Is an Important Driver for People’s Support for the Regulation of Targeted Political Advertising.” *PLoS ONE* 16 (5): e0250506.

Baum, Matthew A. 2002. “Sex, Lies, and War: How Soft News Brings Foreign Policy to the Inattentive Public.” *American Political Science Review* 96 (1): 91–109. https://doi.org/10.1017/S0003055402004252.

Baum, Matthew A., and Tim Groeling. 2010. “Reality Asserts Itself: Public Opinion on Iraq and the Elasticity of Reality.” *International Organization* 64 (3): 443–79. https://doi.org/10.1017/S0020818310000172.

Bechtel, Michael M., and Kenneth F. Scheve. 2013. “Replication Data for: Mass Support for Climate Cooperation Depends on Institutional Design.” Harvard Dataverse.

Begg, Ian Maynard, Ann Anas, and Suzanne Farinacci. 1992. “Dissociation of Processes in Belief: Source Recollection, Statement Familiarity, and the Illusion of Truth.” *Journal of Experimental Psychology: General* 121 (4): 446.

Bell, Mark S., and Kai Quek. 2018. “Authoritarian Public Opinion and the Democratic Peace.” *International Organization* 72 (1): 227–42. https://doi.org/10.1017/S002081831700042X.

Bennett, W. Lance, and Steven Livingston. 2018. “The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions.” *European Journal of Communication* 33 (2): 122–39.

Bentley, Michelle, and Adam B. Lerner. 2022. *A Trump Doctrine?: Unpredictability and Foreign Policy*. Taylor & Francis.

Bertrand, M., and E. Duflo. 2017. “Field Experiments on Discrimination.” In *Handbook of Economic Field Experiments*, 1:309–93. Elsevier. https://doi.org/10.1016/bs.hefe.2016.08.004.

Beskow, David M., and Kathleen M. Carley. 2020. “Characterization and Comparison of Russian and Chinese Disinformation Campaigns.” *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, 63–81.

Bhadani, Saumya, Shun Yamaya, Alessandro Flammini, Filippo Menczer, Giovanni Luca Ciampaglia, and Brendan Nyhan. 2022. “Political Audience Diversity and News Reliability in Algorithmic Ranking.” *Nature Human Behaviour* 6 (4): 495–505. https://doi.org/10.1038/s41562-021-01276-5.

Bittman, Ladislav. 1981. “Soviet Bloc ‘Disinformation’and Other ‘Active Measures.’” *Intelligence Policy and National Security*, 212–28.

Bjola, Corneliu, and Ilan Manor. 2018. “Revisiting Putnam’s Two-Level Game Theory in the Digital Age: Domestic Digital Diplomacy and the Iran Nuclear Deal.” *Cambridge Review of International Affairs* 31 (1): 3–32. https://doi.org/10.1080/09557571.2018.1476836.

Bowler, Shaun. 2017. “Trustees, Delegates, and Responsiveness in Comparative Perspective.” *Comparative Political Studies* 50 (6): 766–93. https://doi.org/10.1177/0010414015626447.

Brader, Ted. 2005. “Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions.” *American Journal of Political Science* 49 (2): 388–405. https://doi.org/10.1111/j.0092-5853.2005.00130.x.

Broockman, David E., Joshua L. Kalla, and Sean J. Westwood. 2023. “Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not.” *American Journal of Political Science* 67 (3): 808–28. https://doi.org/10.1111/ajps.12719.

Brown, Jonathan N, Danielle L Lupton, and Alex Farrington. 2019. “Embedded Deception: Interpersonal Trust, Cooperative Expectations, and the Sharing of Fabricated Intelligence.” *Journal of Global Security Studies* 4 (2): 209–26. https://doi.org/10.1093/jogss/ogy026.

Brown, Nina I., and Jonathan Peters. 2018. “Say This, Not That: Government Regulation and Control of Social Media.” *Syracuse L. Rev.* 68: 521.

Brown, Rupert J., and G. F. Ross. 1982. “The Battle for Acceptance: An Exploration into the Dynamics of Intergroup Behaviour. H. Tajfel.” Social Identity and Itergroup Relations, Cambridge Univ. Press, London.

Brugger, Philipp, Andreas Hasenclever, and Lukas Kasten. 2013. “Vertrauen Lohnt Sich: Über Gegenstand Und Potential Eines Vernachlässigten Konzepts in Den Internationalen Beziehungen.” *Zeitschrift Für Internationale Beziehungen*, 65–104.

Brutger, Ryan, Joshua D. Kertzer, Jonathan Renshon, Dustin Tingley, and Chagai M. Weiss. 2022. “Abstraction and Detail in Experimental Design.” *American Journal of Political Science* n/a (n/a). https://doi.org/10.1111/ajps.12710.

Cairns, Ed, Dale Hunter, and Linda Herring. 1980. “Young Children’s Awareness of Violence in Northern Ireland: The Influence of Northern Irish Television in Scotland and Northern Ireland.” *British Journal of Social and Clinical Psychology* 19 (1): 3–6. https://doi.org/10.1111/j.2044-8260.1980.tb00920.x.

Campbell, James E. 2018. *Polarized: Making Sense of a Divided America*. Princeton University Press.

Capra, C Mónica, Kelli Lanier, and Shireen Meer. 2008. “Attitudinal and Behavioral Measures of Trust: A New Comparison.” *Department of Economics, Emory University*, 52.

Caquet, P. E. 2019. *The Bell of Treason: The 1938 Munich Agreement in Czechoslovakia*. E-Book. Other Press, LLC. https://www.google.com/books/edition/The_Bell_of_Treason/YOiBDwAAQBAJ?hl=en&gbpv=0.

Carmines, Edward G., Michael J. Ensley, and Michael W. Wagner. 2012. “Who Fits the Left-Right Divide? Partisan Polarization in the American Electorate.” *American Behavioral Scientist* 56 (12): 1631–53. https://doi.org/10.1177/0002764212463353.

Carr, E.H. 1939. *The Twenty Years’ Crisis, 1919-1939*. Edited by Michael Cox. London: Palgrave Macmillan. https://www.google.com/books/edition/The_Twenty_Years_Crisis_1919_1939/9ZSkDQAAQBAJ?hl=en&gbpv=0.

Charles, Douglas M. 2000. “American, British and Canadian Intelligence Links: A Critical Annotated Bibliography.” *Intelligence and National Security* 15 (2): 259–69. https://doi.org/10.1080/02684520008432610.

Chaudoin, Stephen, Helen V. Milner, and Dustin H. Tingley. 2010. “The Center Still Holds: Liberal Internationalism Survives.” *International Security* 35 (1): 75–94. https://doi.org/10.1162/ISEC_a_00003.

Chen, Meng, Weihua Yu, and Ke Liu. 2023. “A Meta-Analysis of Third-Person Perception Related to Distorted Information: Synthesizing the Effect, Antecedents, and Consequences.” *Information Processing & Management* 60 (5): 103425. https://doi.org/10.1016/j.ipm.2023.103425.

Cheng, Zicheng, Hugo Marcos-Marne, and Homero Gil de Zúñiga. 2023. “Birds of a Feather Get Angrier Together: Social Media News Use and Social Media Political Homophily as Antecedents of Political Anger.” *Political Behavior*, March. https://doi.org/10.1007/s11109-023-09864-z.

Choi, Charles Q. 2021. “Groovy Flat-Packed Pasta Could Help Revolutionize Food Production.” ABC News. March 9, 2021. https://abcnews.go.com/Technology/groovy-flat-packed-pasta-revolutionize-food-production/story?id=77561572.

Clayton, James, and Ben Derico. 2023. “Clearview AI Used Nearly 1m Times by US Police, It Tells the BBC.” BBC. March 27, 2023. https://www.bbc.com/news/technology-65057011.

Cock Buning, Madeleine de. 2018. “A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation.” European Commission. https://cadmus.eui.eu/bitstream/handle/1814/70297/DeCockB_2018?sequence=1&isAllowed=y.

Colaresi, Michael. 2012. “A Boom with Review: How Retrospective Oversight Increases the Foreign Policy Ability of Democracies.” *American Journal of Political Science* 56 (3): 671–89. https://doi.org/10.1111/j.1540-5907.2011.00567.x.

Considine, Laura. 2015. “‘Back to the Rough Ground!’A Grammatical Approach to Trust and International Relations.” *Millennium* 44 (1): 109–27.

Corbu, Nicoleta, Denisa-Adriana Oprea, Elena Negrea-Busuioc, and Loredana Radu. 2020. “‘They Can’t Fool Me, but They Can Fool the Others!’ Third Person Effect and Fake News Detection.” *European Journal of Communication* 35 (2): 165–80. https://doi.org/10.1177/0267323120903686.

Coventry Evening Telegraph. 1955. “Two Jews Executed in Cairo for Spying,” January 31, 1955.

Crescenzi, Mark J. C. 2007. “Reputation and Interstate Conflict.” *American Journal of Political Science* 51 (2): 382–96. https://doi.org/10.1111/j.1540-5907.2007.00257.x.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. “Information Equivalence in Survey Experiments.” *Political Analysis* 26 (4): 399–416. https://doi.org/10.1017/pan.2018.9.

Daoyenikye, Fredrick. 2023. “How Does Social Media Contribute to Affective Polarization? Mediating Role of Fake News Beliefs.” M.A., United States -- New Mexico: New Mexico State University. https://www.proquest.com/docview/2827900305/abstract/532C8D7B23694792PQ/1.

Davison, W. Phillips. 1983. “The Third-Person Effect in Communication.” *The Public Opinion Quarterly* 47 (1): 1–15.

Dawson, Andrew, and Martin Innes. 2019. “How Russia’s Internet Research Agency Built Its Disinformation Campaign.” *The Political Quarterly* 90 (2): 245–56.

De Vries, Catherine E., Sara B. Hobolt, and Stefanie Walter. 2021. “Politicizing International Cooperation: The Mass Public, Political Entrepreneurs, and Political Opportunity Structures.” *International Organization* 75 (2): 306–32. https://doi.org/10.1017/S0020818320000491.

Delhey, Jan, and Kenneth Newton. 2005. “Predicting Cross-National Levels of Social Trust: Global Pattern or Nordic Exceptionalism?” *European Sociological Review* 21 (4): 311–27. https://doi.org/10.1093/esr/jci022.

Delhey, Jan, Kenneth Newton, and Christian Welzel. 2011. “How General Is Trust in ‘Most People’? Solving the Radius of Trust Problem.” *American Sociological Review* 76 (5): 786–807. https://doi.org/10.1177/0003122411420817.

Demirjian, Karoun. 2023. “Russians Have Many Theories about the MH17 Crash. One Involves Fake Dead People.” *Washington Post*, April 15, 2023. https://www.washingtonpost.com/world/russians-have-many-theories-about-the-mh17-crash-one-involves-fake-dead-people/2014/07/22/9a1c5ec9-11b6-4384-b585-53fff62e5779_story.html.

Dinesen, Peter Thisted, Merlin Schaeffer, and Kim Mannemar Sønderskov. 2020. “Ethnic Diversity and Social Trust: A Narrative and Meta-Analytical Review.” *Annual Review of Political Science* 23: 441–65.

DiResta, Renee, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. 2019. “The Tactics & Tropes of the Internet Research Agency.” United States Senate Select Committee on Intelligence. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1003&context=senatedocs.

Dombrowski, Peter, and Simon Reich. 2017. “Does Donald Trump Have a Grand Strategy?” *International Affairs* 93 (5): 1013–37.

Doyle, Joshua. 2021. “The Effect of Cultural Trust on Cooperation in Two Behavioral Experiments.” *Social Psychology Quarterly* 84 (3): 246–66.

Druckman, James N., Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2021a. “How Affective Polarization Shapes Americans’ Political Beliefs: A Study of Response to the COVID-19 Pandemic.” *Journal of Experimental Political Science* 8 (3): 223–34. https://doi.org/10.1017/XPS.2020.28.

———. 2021b. “Affective Polarization, Local Contexts and Public Opinion in America.” *Nature Human Behaviour* 5 (1): 28–38. https://doi.org/10.1038/s41562-020-01012-5.

Druckman, James N., and Mary C. McGrath. 2019. “The Evidence for Motivated Reasoning in Climate Change Preference Formation.” *Nature Climate Change* 9 (2): 111–19.

Dukalskis, Alexander, and Johannes Gerschewski. 2017. “What Autocracies Say (and What Citizens Hear): Proposing Four Mechanisms of Autocratic Legitimation.” *Contemporary Politics* 23 (3): 251–68.

Elswah, Mona, and Philip N Howard. 2020. “‘Anything That Causes Chaos’: The Organizational Behavior of Russia Today (RT).” *Journal of Communication* 70 (5): 623–45. https://doi.org/10.1093/joc/jqaa027.

Entous, Adam, Craig Timberg, and Elizabeth Dwoskin. 2023. “Russian Operatives Used Facebook Ads to Exploit America’s Racial and Religious Divisions.” *Washington Post*, September 25, 2023. http://static.cs.brown.edu/people/jsavage/VotingProject/2017_09_25_WaPo_RussianOperativesUsedFacebookAdsToExploitAmericasRacialAndReligiousDivisions.pdf.

Etudo, Ugo, Victoria Y. Yoon, and Niam Yaraghi. 2019. “From Facebook to the Streets: Russian Troll Ads and Black Lives Matter Protests.” In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/3b1a59b0-f010-4343-b3e2-e677365bf3a1/content.

Facebook. n.d. “Datasets.” Accessed November 29, 2022. https://scontent.fphx1-2.fna.fbcdn.net/v/t39.8562-6/258872344_407777644381740_4909820853571583715_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=ad8a9d&_nc_ohc=H5qMKlTji8cAX-s52ui&_nc_ht=scontent.fphx1-2.fna&oh=00_AfBK9OzhfkFEYqACh-SG1JyyJAyu9wiUHvKEcM1sOgDeSw&oe=638A6808.

Fallis, Don. 2015. “What Is Disinformation?” *Library Trends* 63 (3): 401–26.

Fazio, Lisa K., David G. Rand, and Gordon Pennycook. 2019. “Repetition Increases Perceived Truth Equally for Plausible and Implausible Statements.” *Psychonomic Bulletin & Review* 26 (5): 1705–10.

Fearon, James D. 1994. “Domestic Political Audiences and the Escalation of International Disputes.” *American Political Science Review* 88 (3): 577–92.

———. 1995. “Rationalist Explanations for War.” *International Organization* 49 (3): 379–414.

———. 1997. “Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs.” *Journal of Conflict Resolution* 41 (1): 68–90. https://doi.org/10.1177/0022002797041001004.

Fenno, Richard F. 1973. *Congressmen in Committees*. Little, Brown, & Company. https://www.google.com/books/edition/Congressmen_in_Committees/Ap9CAAAAIAAJ?hl=en&gbpv=0&bsq=congressmen%20in%20committees%20richard%20fenno.

Fiorina, Morris P., Samuel J. Abrams, and Jeremy Pope. 2006. *Culture War?: The Myth of a Polarized America*. Pearson Education.

Flynn, D.j., Brendan Nyhan, and Jason Reifler. 2017. “The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics.” *Political Psychology* 38 (S1): 127–50. https://doi.org/10.1111/pops.12394.

Folkman, Susan, Richard Lazarus, Christine Schetter, Anita DeLongis, and Rand Gruen. 1986. “Dynamics of a Stressful Encounter: Cognitive Appraisal, Coping, and Encounter Outcomes.” *Journal of Personality and Social Psychology* 50 (May): 992–1003. https://doi.org/10.1037/0022-3514.50.5.992.

Frankel, Max. 2002. “Learning from the Missile Crisis.” Smithsonian Magazine. October 2002. https://www.smithsonianmag.com/history/learning-from-the-missile-crisis-68901679/.

Frantz, Erica, Andrea Kendall-Taylor, and Joseph Wright. 2020. “Digital Repression in Autocracies.” *Varieties of Democracy Institute Users Working Paper (27)*.

Freitag, Markus, and Richard Traunmüller. 2009. “Spheres of Trust: An Empirical Analysis of the Foundations of Particularised and Generalised Trust.” *European Journal of Political Research* 48 (6): 782–803.

Friedrichs, Gordon M. 2022a. “Conceptualizing the Effects of Polarization for US Foreign Policy Behavior in International Negotiations: Revisiting the Two-Level Game1.” *International Studies Review* 24 (1): viac010. https://doi.org/10.1093/isr/viac010.

Friedrichs, Gordon M. 2022b. “Polarized We Trade? Intraparty Polarization and US Trade Policy.” *International Politics* 59 (5): 956–80. https://doi.org/10.1057/s41311-021-00344-x.

Friedrichs, Gordon M., and Jordan Tama. 2022. “Polarization and US Foreign Policy: Key Debates and New Findings.” *International Politics* 59 (5): 767–85. https://doi.org/10.1057/s41311-022-00381-0.

Gächter, Simon, Benedikt Herrmann, and Christian Thöni. 2004. “Trust, Voluntary Cooperation, and Socio-Economic Background: Survey and Experimental Evidence.” *Journal of Economic Behavior & Organization* 55 (4): 505–31. https://doi.org/10.1016/j.jebo.2003.11.006.

Gadarian, Shana Kushner, and Bethany Albertson. 2014. “Anxiety, Immigration, and the Search for Information.” *Political Psychology* 35 (2): 133–64.

Gale, David, and Frank M. Stewart. 1953. “Infinite Games with Perfect Information.” *Contributions to the Theory of Games* 2 (245–266): 2–16.

Galeotti, Mark. 2015. “Hybrid War” and ‘Little Green Men’: How It Works, and How It Doesn’t.” *Ukraine and Russia: People, Politics, Propaganda and Perspectives* 156.

Gallarotti, Giulio M. 2011. “Soft Power: What It Is, Why It’s Important, and the Conditions for Its Effective Use.” *Journal of Political Power* 4 (1): 25–47.

Gallina, Marta. 2023. “The Concept of Political Sophistication: Labeling the Unlabeled.” *Political Studies Review* 21 (January): 147892992211460. https://doi.org/10.1177/14789299221146058.

Gallop, Max, and Zachary Greene. 2021. “Polarisation, Accountability, and Interstate Conflict.” *The British Journal of Politics and International Relations* 23 (1): 121–38. https://doi.org/10.1177/1369148120944349.

Garrett, R. Kelly, Shira Dvir Gvirsman, Benjamin K. Johnson, Yariv Tsfati, Rachel Neo, and Aysenur Dal. 2014. “Implications of Pro- and Counterattitudinal Information Exposure for Affective Polarization.” *Human Communication Research* 40 (3): 309–32. https://doi.org/10.1111/hcre.12028.

Gartzke, Erik. 1999. “War Is in the Error Term.” *International Organization* 53 (3): 567–87. https://doi.org/10.1162/002081899550995.

Geissler, Erhard, and Robert Hunt Sprinkle. 2013. “Disinformation Squared: Was the HIV-from-Fort-Detrick Myth a Stasi Success?” *Politics and the Life Sciences* 32 (2): 2–99.

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

Gill, Hyungjin. 2022. "Testing the Effect of Cross-Cutting Exposure to Cable TV News on Affective Polarization: Evidence from the 2020 U.S. Presidential Election." *Journal of Broadcasting & Electronic Media* 66 (2): 320–39. https://doi.org/10.1080/08838151.2022.2087653.

Giusti, Serena, and Elisa Piras. 2020. *Democracy and Fake News: Information Manipulation and Post-Truth Politics*. Routledge.

Glaeser, Edward L., David I. Laibson, José A. Scheinkman, and Christine L. Soutter. 2000. "Measuring Trust*." *The Quarterly Journal of Economics* 115 (3): 811–46. https://doi.org/10.1162/003355300554926.

Goldsmith, Benjamin E., and Yusaku Horiuchi. 2012. "In Search of Soft Power: Does Foreign Public Opinion Matter for US Foreign Policy?" *World Politics* 64 (3): 555–85. https://doi.org/10.1017/S0043887112000123.

Google Scholar. 2022. "Google Scholar." *Advanced Search Results by Keyword in Title* (blog). 2022. https://scholar.google.com/.

Gordon, Susan M, Michael G Mullen, and David Sacks. 2023. "U.S.-Taiwan Relations in a New Era." 81. Independent Task Force. New York, NY: Council on Foreign Relations.

Greene, Steven. 2004. "Social Identity Theory and Party Identification*." *Social Science Quarterly* 85 (1): 136–53. https://doi.org/10.1111/j.0038-4941.2004.08501010.x.

Guess, Andrew, and Alexander Coppock. 2020. "Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments." *British Journal of Political Science* 50 (4): 1497–1515. https://doi.org/10.1017/S0007123418000327.

Guess, Andrew, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. "A Digital Media Literacy Intervention Increases Discernment between Mainstream and False News in the United States and India." *Proceedings of the National Academy of Sciences* 117 (27): 15536–45. https://doi.org/10.1073/pnas.1920498117.

Guess, Andrew, Dominique Lockett, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler. 2020. "'Fake News' May Have Limited Effects beyond Increasing Beliefs in False Claims."

Guess, Andrew, and Benjamin A. Lyons. 2020. "Misinformation, Disinformation, and Online Propaganda." *Social Media and Democracy: The State of the Field, Prospects for Reform*, 10–33.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. “Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook.” *Science Advances* 5 (1): eaau4586. https://doi.org/10.1126/sciadv.aau4586.

Guess, Andrew, Brendan Nyhan, and Jason Reifler. 2020. “Exposure to Untrustworthy Websites in the 2016 US Election.” *Nature Human Behaviour* 4 (5): 472–80.

Gunitsky, Seva. 2015. “Corrupting the Cyber-Commons: Social Media as a Tool of Autocratic Stability.” *Perspectives on Politics* 13 (1): 42–54.

Guo, Bin, Yasan Ding, Yueheng Sun, Shuai Ma, Ke Li, and Zhiwen Yu. 2021. “The Mass, Fake News, and Cognition Security.” *Frontiers of Computer Science* 15: 1–13.

Haase, Amy. 2024. “Brain Cancer Patient Prepares to Run London Marathon.” South West Londoner. January 25, 2024. https://www.swlondoner.co.uk/life/25012024-brain-cancer-patient-prepares-to-run-london-marathon.

Hallin, Daniel C. 1989. *The Uncensored War: The Media and Vietnam*. University of California Press.

Hameleers, Michael. 2020. “Separating Truth from Lies: Comparing the Effects of News Media Literacy Interventions and Fact-Checkers in Response to Political Misinformation in the US and Netherlands.” *Information, Communication & Society*, 1–17.

Hameleers, Michael, and Toni G. L. A. van der Meer. 2020. “Misinformation and Polarization in a High-Choice Media Environment: How Effective Are Political Fact-Checkers?” *Communication Research* 47 (2): 227–50. https://doi.org/10.1177/0093650218819671.

Han, Rongbin. 2015. “Defending the Authoritarian Regime Online: China’s ‘Voluntary Fifty-Cent Army.’” *The China Quarterly* 224: 1006–25.

Hangartner, Dominik, Elias Dinas, Moritz Marbach, Konstantinos Matakos, and Dimitrios Xefteris. 2019. “Does Exposure to the Refugee Crisis Make Natives More Hostile?” *American Political Science Review* 113 (2): 442–55. https://doi.org/10.1017/S0003055418000813.

Hanyok, Robert J. 1998. “Skunks, Bogies, Silent Hounds, and the Flying Fish: The Gulf of Tonkin Mystery, 2-4 August 1964.” *Cryptologic Quarterly*, 1–55.

Harold, Scott W., Nathan Beauchamp-Mustafaga, and Jeffrey W. Hornung. 2021. “Chinese Disinformation Efforts.” Combating Foreign Disinformation on Social Media. Santa Monica, California: RAND Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RR4300/RR4373z3/RAND_RR4373z3.pdf.

Haselswerdt, Jake, and John Sides. 2019. “Campaigns and Elections.” In *New Directions in Public Opinion*, 3rd ed. Routledge.

Hasher, Lynn, David Goldstein, and Thomas Toppino. 1977. “Frequency and the Conference of Referential Validity.” *Journal of Verbal Learning and Verbal Behavior* 16 (1): 107–12.

Haukkala, Hiski, Carina Van de Wetering, and Johanna Vuorelma. 2018. *Trust in International Relations: Rationalist, Constructivist, and Psychological Approaches*. Routledge.

Hawkins, Scott A., and Stephen J. Hoch. 1992. “Low-Involvement Learning: Memory without Evaluation.” *Journal of Consumer Research* 19 (2): 212–25.

Hayward, James. 2016. *Burn the Sea: Flame Warfare, Black Propaganda and the Nazi Plan to Invade England*. The History Press. https://books.google.com/books?hl=en&lr=&id=12erCwAAQBAJ&oi=fnd&pg=PT7&dq=shingle+street+hayward&ots=wY455tMP-a&sig=aoeFyKvBVo7cqvZyfngoel2XNBA.

Heatherly, Kyle A, Yanqin Lu, and Jae Kook Lee. 2017. “Filtering out the Other Side? Cross-Cutting and like-Minded Discussions on Social Networking Sites.” *New Media & Society* 19 (8): 1271–89. https://doi.org/10.1177/1461444816634677.

Heider, Fritz. 1946. “Attitudes and Cognitive Organization.” *The Journal of Psychology* 21 (1): 107–12.

———. 1982. *The Psychology of Interpersonal Relations*. Psychology Press.

Hemming, Henry. 2019. *Agents of Influence: A British Campaign, a Canadian Spy, and the Secret Plot to Bring America into World War II*. E-Book. PublicAffairs.

Hetherington, Marc J. 2015. “Partisanship and Polarization in Contemporary Politics.” In *New Directions in Public Opinion*, 168–86. Routledge.

Hetherington, Marc J., and Jason A. Husser. 2012. “How Trust Matters: The Changing Political Relevance of Political Trust.” *American Journal of Political Science* 56 (2): 312–25. https://doi.org/10.1111/j.1540-5907.2011.00548.x.

Hetherington, Marc J., and Thomas J. Rudolph. 2015. *Why Washington Won’t Work: Polarization, Political Trust, and the Governing Crisis*. University of Chicago Press.

Highton, Benjamin. 2009. “Revisiting the Relationship between Educational Attainment and Political Sophistication.” *The Journal of Politics* 71 (4): 1564–76. https://doi.org/10.1017/S0022381609990077.

Hill, Kim Quaile, Soren Jordan, and Patricia A. Hurley. 2015. *Representation in Congress: A Unified Theory*. Cambridge University Press.

Hoffman, Aaron M. 2002. "A Conceptualization of Trust in International Relations." *European Journal of International Relations* 8 (3): 375–401.

Hoffman, Tod. 2002. "Seeking Intrepid." *Queen's Quarterly* 109 (2): 255–65.

Hoffner, Cynthia, and Raiza A. Rehkoff. 2011. "Young Voters' Responses to the 2004 U.S. Presidential Election: Social Identity, Perceived Media Influence, and Behavioral Outcomes." *Journal of Communication* 61 (4): 732–57. https://doi.org/10.1111/j.1460-2466.2011.01565.x.

Horner, Christy Galletta, Dennis Galletta, Jennifer Crawford, and Abhijeet Shirsat. 2021. "Emotions: The Unexplored Fuel of Fake News on Social Media." *Journal of Management Information Systems* 38 (4): 1039–66. https://doi.org/10.1080/07421222.2021.1990610.

Huddy, Leonie, and Alexa Bankert. 2017. "Political Partisanship as a Social Identity." Oxford Research Encyclopedia of Politics. May 24, 2017. https://doi.org/10.1093/acrefore/9780190228637.013.250.

Huddy, Leonie, and Omer Yair. 2021. "Reducing Affective Polarization: Warm Group Relations or Policy Compromise?" *Political Psychology* 42 (2): 291–309. https://doi.org/10.1111/pops.12699.

Hung, Tzu-Chieh, and Tzu-Wei Hung. 2022. "How China's Cognitive Warfare Works: A Frontline Perspective of Taiwan's Anti-Disinformation Wars." *Journal of Global Security Studies* 7 (4): ogac016. https://doi.org/10.1093/jogss/ogac016.

Hyde, Susan D., and Elizabeth N. Saunders. 2020. "Recapturing Regime Type in International Relations: Leaders, Institutions, and Agency Space." *International Organization* 74 (2): 363–95.

Hyun, Ki Deuk, and Mihye Seo. 2021. "The Effects of HMP and TPP on Political Participation in the Partisan Media Context." *Communication Research* 48 (5): 665–86. https://doi.org/10.1177/0093650218820229.

IAEA. 2018. "Statement on Iran by the IAEA Spokesperson." Text. International Atomic Energy Agency. IAEA. May 1, 2018. https://www.iaea.org/newscenter/pressreleases/statement-on-iran-by-the-iaea-spokesperson.

Ignatius, David. 1989. "How Churchill's Agents Secretly Manipulated the US before Pearl Harbor." *Washington Post*.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. “Identification, Inference and Sensitivity Analysis for Causal Mediation Effects.” *Statistical Science*, 51–71.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto. 2013. “Experimental Designs for Identifying Causal Mechanisms.” *Journal of the Royal Statistical Society Series A: Statistics in Society* 176 (1): 5–51.

Ivshina, Olga. 2015. “Flight MH17: Russia and Its Changing Story.” *BBC News*, October 16, 2015, sec. Europe. https://www.bbc.com/news/world-europe-34538142.

Iyengar, Shanto, Tobias Konitzer, and Kent Tedin. 2018. “The Home as a Political Fortress: Family Agreement in an Era of Polarization.” *The Journal of Politics* 80 (4): 1326–38. https://doi.org/10.1086/698929.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. “The Origins and Consequences of Affective Polarization in the United States.” *Annual Review of Political Science* 22 (1): 129–46. https://doi.org/10.1146/annurev-polisci-051117-073034.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. “Affect, Not Ideology: A Social Identity Perspective on Polarization.” *Public Opinion Quarterly* 76 (3): 405–31. https://doi.org/10.1093/poq/nfs038.

Janiszewski, Chris. 1993. “Preattentive Mere Exposure Effects.” *Journal of Consumer Research* 20 (3): 376–92.

Jefferson, Hakeem, Fabian G. Neuner, and Josh Pasek. 2021. “Seeing Blue in Black and White: Race and Perceptions of Officer-Involved Shootings.” *Perspectives on Politics* 19 (4): 1165–83. https://doi.org/10.1017/S1537592720003618.

Jessee, Stephen A. 2017. “‘Don’t Know’ Responses, Personality, and the Measurement of Political Knowledge.” *Political Science Research and Methods* 5 (4): 711–31. https://doi.org/10.1017/psrm.2015.23.

John, Nicholas A., and Shira Dvir-Gvirsman. 2015. “‘I Don’t like You Any More’: Facebook Unfriending by Israelis during the Israel–Gaza Conflict of 2014.” *Journal of Communication* 65 (6): 953–74. https://doi.org/10.1111/jcom.12188.

Johns, Robert, and Graeme AM Davies. 2012. “Democratic Peace or Clash of Civilizations? Target States and Support for War in Britain and the United States.” *The Journal of Politics* 74 (4): 1038–52.

Jones, David R. 2010. “Partisan Polarization and Congressional Accountability in House Elections.” *American Journal of Political Science* 54 (2): 323–37. https://doi.org/10.1111/j.1540-5907.2010.00433.x.

Jowett, Garth S., and Victoria O'Donnell. 2018. *Propaganda & Persuasion*. SAGE Publications.

Karlin, Samuel. 1953. "The Theory of Infinite Games." *Annals of Mathematics*, 371–401.

Karlsen, Geir Hågen. 2019. "Divide and Rule: Ten Lessons about Russian Political Influence Activities in Europe." *Palgrave Communications* 5 (1): 1–14. https://doi.org/10.1057/s41599-019-0227-8.

Kaufmann, Chaim D. 1994. "Out of the Lab and into the Archives: A Method for Testing Psychological Explanations of Political Decision Making." *International Studies Quarterly* 38 (4): 557–86. https://doi.org/10.2307/2600865.

Kensinger, Elizabeth A. 2007. "Negative Emotion Enhances Memory Accuracy: Behavioral and Neuroimaging Evidence." *Current Directions in Psychological Science* 16 (4): 213–18. https://doi.org/10.1111/j.1467-8721.2007.00506.x.

Keremoğlu, Eda, and Nils B. Weidmann. 2020. "How Dictators Control the Internet: A Review Essay." *Comparative Political Studies* 53 (10–11): 1690–1703.

Kertzer, Joshua D, Jonathan Renshon, and Keren Yarhi-Milo. 2021. "How Do Observers Assess Resolve?" *British Journal of Political Science* 51 (1): 308–30. https://doi.org/10.1017/S0007123418000595.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2014. "Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation." *Science* 345 (6199).

Kragh, Martin, and Sebastian Åsberg. 2017. "Russia's Strategy for Influence through Public Diplomacy and Active Measures: The Swedish Case." *Journal of Strategic Studies* 40 (6): 773–816.

Krolik, Jeff, Raymond Zhong, Paul Mozur, and Aaron Krolik. 2021. "How China Spreads Its Propaganda Version of Life for Uyghurs." ProPublica. June 23, 2021. https://www.propublica.org/article/how-china-uses-youtube-and-twitter-to-spread-its-propaganda-version-of-life-for-uyghurs-in-xinjiang.

Kubin, Emily, and Christian von Sikorski. 2021. "The Role of (Social) Media in Political Polarization: A Systematic Review." *Annals of the International Communication Association* 45 (3): 188–206. https://doi.org/10.1080/23808985.2021.1976070.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (3): 480–98. https://doi.org/10.1037/0033-2909.108.3.480.

Kupchan, Charles A., and Peter L. Trubowitz. 2007. "Dead Center: The Demise of Liberal Internationalism in the United States." *International Security* 32 (2): 7–44. https://doi.org/10.1162/isec.2007.32.2.7.

Kurlantzick, Joshua. 2020. “How China Ramped Up Disinformation Efforts During the Pandemic.” *Council on Foreign Relations In Brief. September* 10.

La Cour, Christina. 2020. “Theorising Digital Disinformation in International Relations.” *International Politics* 57 (4): 704–23.

Lanoszka, Alexander. 2019. “Disinformation in International Politics.” *European Journal of International Security* 4 (2): 227–48.

Larson, Deborah Welch. 1997. “Trust and Missed Opportunities in International Relations.” *Political Psychology* 18 (3): 701–34.

Lee, Amber Hye-Yon, Yphtach Lelkes, Carlee B. Hawkins, and Alexander G. Theodoridis. 2022. “Negative Partisanship Is Not More Prevalent than Positive Partisanship.” *Nature Human Behaviour* 6 (7): 951–63. https://doi.org/10.1038/s41562-022-01348-0.

Lee, Seungsu, and Jaeho Cho. 2023. “Communication Mediation in an Era of Partisan Selectivity: Modeling Effects of Information and Discussion on Participation.” *International Journal of Public Opinion Research* 35 (3): edad020.

Lee, Seungsu, and Kyungmo Kim. 2022. “Perceived Influence of Partisan News and Online News Participation: Third-Person Effect, Hostile Media Phenomenon, and Cognitive Elaboration.” *Communication Research*, November, 00936502221127494. https://doi.org/10.1177/00936502221127494.

Lelkes, Yphtach. 2021. “Policy over Party: Comparing the Effects of Candidate Ideology and Party on Affective Polarization.” *Political Science Research and Methods* 9 (1): 189–96. https://doi.org/10.1017/psrm.2019.18.

Lerner, Jennifer S., Roxana M. Gonzalez, Deborah A. Small, and Baruch Fischhoff. 2003. “Effects of Fear and Anger on Perceived Risks of Terrorism: A National Field Experiment.” *Psychological Science* 14 (2): 144–50. https://doi.org/10.1111/1467-9280.01433.

Lerner, Jennifer S., and Dacher Keltner. 2000. “Beyond Valence: Toward a Model of Emotion-Specific Influences on Judgement and Choice.” *Cognition & Emotion* 14 (4): 473–93. https://doi.org/10.1080/026999300402763.

Levendusky, Matthew, and Neil Malhotra. 2016. “Does Media Coverage of Partisan Polarization Affect Political Attitudes?” *Political Communication* 33 (2): 283–301. https://doi.org/10.1080/10584609.2015.1038455.

Levendusky, Matthew S. 2013a. “Why Do Partisan Media Polarize Viewers?” *American Journal of Political Science* 57 (3): 611–23. https://doi.org/10.1111/ajps.12008.

———. 2013b. “Partisan Media Exposure and Attitudes Toward the Opposition.” *Political Communication* 30 (4): 565–81. https://doi.org/10.1080/10584609.2012.737435.

———. 2018. “Americans, Not Partisans: Can Priming American National Identity Reduce Affective Polarization?” *The Journal of Politics* 80 (1): 59–70. https://doi.org/10.1086/693987.

Levy, Ro’ee. 2021. “Social Media, News Consumption, and Polarization: Evidence from a Field Experiment.” *American Economic Review* 111 (3): 831–70. https://doi.org/10.1257/aer.20191777.

Lin, Mei-Chen, Paul M. Haridakis, and Yan Bing Zhang. 2020. “Political Party Identification and Intergroup Attitudes: Exploring the Effects of Mediated and Direct Contact With the Opposing Party During a Presidential Campaign.” *International Journal of Communication* 14 (0): 18.

Little, Andrew T., Keith E. Schnakenberg, and Ian R. Turner. 2022. “Motivated Reasoning and Democratic Accountability.” *American Political Science Review* 116 (2): 751–67. https://doi.org/10.1017/S0003055421001209.

Lo, Ven-Hwei, Grace Xiao Zhang, and Miao Lu. 2023. “Consequences of Exposure to Misinformation: Negative Emotions and Biased Risk Perception.” In *Miscommunicating the COVID-19 Pandemic: An Asian Perspective*, 89–111. London: Routledge. https://doi.org/10.4324/9781003355984.

Loomba, Sahil, Alexandre de Figueiredo, Simon J. Piatek, Kristen de Graaf, and Heidi J. Larson. 2021. “Measuring the Impact of COVID-19 Vaccine Misinformation on Vaccination Intent in the UK and USA.” *Nature Human Behaviour* 5 (3): 337–48.

Lueders, Hans. 2021. “Electoral Responsiveness in Closed Autocracies: Evidence from Petitions in the Former German Democratic Republic.” *American Political Science Review*, December, 1–16. https://doi.org/10.1017/S0003055421001386.

Luo, Mufan, Jeffrey T. Hancock, and David M. Markowitz. 2020. “Credibility Perceptions and Detection Accuracy of Fake News Headlines on Social Media: Effects of Truth-Bias and Endorsement Cues.” *Communication Research*, 0093650220921321.

Lupton, Robert N., William M. Myers, and Judd R. Thornton. 2015. “Political Sophistication and the Dimensionality of Elite and Mass Attitudes, 1980−2004.” *The Journal of Politics* 77 (2): 368–80. https://doi.org/10.1086/679493.

Luskin, Robert C. 1987. “Measuring Political Sophistication.” *American Journal of Political Science* 31 (4): 856–99. https://doi.org/10.2307/2111227.

Luskin, Robert C., and John G. Bullock. 2011. "'Don't Know' Means 'Don't Know': DK Responses and the Public's Level of Political Knowledge." *The Journal of Politics* 73 (2): 547–57. https://doi.org/10.1017/S0022381611000132.

Lyons, Benjamin A. 2022. "Why We Should Rethink the Third-Person Effect: Disentangling Bias and Earned Confidence Using Behavioral Data." *Journal of Communication* 72 (5): 565–77. https://doi.org/10.1093/joc/jqac021.

MacKuen, Michael, Jennifer Wolak, Luke Keele, and George E. Marcus. 2010. "Civic Engagements: Resolute Partisanship or Reflective Deliberation." *American Journal of Political Science* 54 (2): 440–58. https://doi.org/10.1111/j.1540-5907.2010.00440.x.

Martin, L. John. 1982. "Disinformation: An Instrumentality in the Propaganda Arsenal." *Political Communication* 2 (1): 47–64.

Mason, Lilliana. 2015. "'I Disrespectfully Agree': The Differential Effects of Partisan Sorting on Social and Issue Polarization." *American Journal of Political Science* 59 (1): 128–45. https://doi.org/10.1111/ajps.12089.

———. 2018. *Uncivil Agreement: How Politics Became Our Identity*. University of Chicago Press.

McLoughlin, Killian L., William J. Brady, and Molly J. Crockett. 2021. "The Role of Moral Outrage in the Spread of Misinformation."

McMahon, Robert. 2022. "Russia Is Censoring News on the War in Ukraine. Foreign Media Are Trying to Get Around That." Council on Foreign Relations. March 18, 2022. https://www.cfr.org/in-brief/russia-censoring-news-war-ukraine-foreign-media-are-trying-get-around.

Mcquail, Denis, and Sven Windahl. 2015. *Communication Models for the Study of Mass Communications*. Routledge.

Mearsheimer, John J. 2011. *Why Leaders Lie: The Truth about Lying in International Politics*. Oxford University Press.

Media Insight Project. 2014. "The Personal News Cycle: How Americans Choose to Get Their News." American Press Institute, Associated Press-NORC Center for Public Affairs Research. https://www.americanpressinstitute.org/publications/reports/survey-research/how-americans-get-news/.

Meirick, Patrick C. 2004. "Topic-Relevant Reference Groups and Dimensions of Distance: Political Advertising and First- and Third-Person Effects." *Communication Research* 31 (2): 234–55. https://doi.org/10.1177/0093650203261514.

Mihalka, Michael. 1980. *German Strategic Deception in the 1930s*. Rand Corporation. https://apps.dtic.mil/sti/citations/tr/ADA095120.

Milesi, Patrizia. 2016. "Moral Foundations and Political Attitudes: The Moderating Role of Political Sophistication." *International Journal of Psychology* 51 (August): 252–60. https://doi.org/10.1002/ijop.12158.

Miller, Alan S., and Tomoko Mitamura. 2003. "Are Surveys on Trust Trustworthy?" *Social Psychology Quarterly* 66 (1): 62–70. https://doi.org/10.2307/3090141.

Miller, Gregory D. 2012. *The Shadow of the Past: Reputation and Military Alliances before the First World War*. Cornell University Press.

Miller, Michael K. 2015. "Elections, Information, and Policy Responsiveness in Autocratic Regimes." *Comparative Political Studies* 48 (6): 691–727. https://doi.org/10.1177/0010414014555443.

Miller, Warren E., and Donald E. Stokes. 1963. "Constituency Influence in Congress." *American Political Science Review* 57 (1): 45–56.

Mize, Trenton D. n.d. "Sobel-Goodman Tests of Mediation in Stata." Academic. Trenton D. Mize. n.d. https://www.trentonmize.com/software/sgmediation2.

Mondak, Jeffery J. 2000. "Reconsidering the Measurement of Political Knowledge." *Political Analysis* 8 (1): 57–82.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62 (3): 760–75. https://doi.org/10.1111/ajps.12357.

Morrow, James D. 1994. "Modeling the Forms of International Cooperation: Distribution Versus Information." *International Organization* 48 (3): 387–423.

Myrick, Rachel. 2021. "Do External Threats Unite or Divide? Security Crises, Rivalries, and Polarization in American Foreign Policy." *International Organization* 75: 38.

———. 2022. "The Reputational Consequences of Polarization for American Foreign Policy: Evidence from the US-UK Bilateral Relationship." *International Politics* 59 (5): 1004–27. https://doi.org/10.1057/s41311-022-00382-z.

Naftali, Timothy J. 2012. "Intrepid's Last Deception: Documenting the Career of Sir William Stephenson." In *Espionage: Past, Present and Future?*, 72–99. Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203043813-6/intrepid-last-deception-documenting-career-sir-william-stephenson-timothy-naftali.

Nanz, Andreas, and Jörg Matthes. 2022. “Democratic Consequences of Incidental Exposure to Political Information: A Meta-Analysis.” *Journal of Communication* 72 (3): 345–73. https://doi.org/10.1093/joc/jqac008.

NATO. 2023. “Member Countries.” North Atlantic Treaty Organization. April 5, 2023. https://www.nato.int/cps/en/natohq/nato_countries.htm.

Nayak, Ashutosh, Mayur Garg, and Rajasekhara Reddy Duvvuru Muni. 2023. “News Popularity Beyond the Click-Through-Rate for Personalized Recommendations.” In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1396–1405. Taipei Taiwan: ACM. https://doi.org/10.1145/3539618.3591741.

Nelson, Jacob L., and Harsh Taneja. 2018. “The Small, Disloyal Fake News Audience: The Role of Audience Availability in Fake News Consumption.” *New Media & Society* 20 (10): 3720–37.

New York Herald Tribune. 1940. “Hitler’s Agent Ensconced in Westchester,” August 1, 1940.

Nisbet, Erik. 2020. “Replication Data for "The Presumed Influence of Election Misinformation on Others Reduces Our Own Satisfaction with Democracy.” Harvard Dataverse. https://doi.org/10.7910/DVN/QMEBYZ.

Nisbet, Erik C., Kathryn E. Cooper, and R. Kelly Garrett. 2015. “The Partisan Brain: How Dissonant Science Messages Lead Conservatives and Liberals to (Dis)Trust Science.” *The ANNALS of the American Academy of Political and Social Science* 658 (1): 36–66. https://doi.org/10.1177/0002716214555474.

Nisbet, Erik C., Chloe Mortenson, and Qin Li. 2021. “The Presumed Influence of Election Misinformation on Others Reduces Our Own Satisfaction with Democracy.” *Harvard Kennedy School Misinformation Review*, March. https://doi.org/10.37016/mr-2020-59.

Nocetti, Julien. 2015. “Contest and Conquest: Russia and Global Internet Governance.” *International Affairs (Royal Institute of International Affairs 1944-)* 91 (1): 111–30.

“Nomination of Hon. Mike Pompeo to Be Secretary of State.” 2018. Washington, DC. https://www.congress.gov/115/chrg/CHRG-115shrg29844/CHRG-115shrg29844.pdf.

Noone, Harry. 2019. “Two-Level Games and the Policy Process: Assessing Domestic–Foreign Policy Linkage Theory.” *World Affairs* 182 (2): 165–86.

Norris, Pippa. 2006. “Did the Media Matter? Agenda-Setting, Persuasion and Mobilization Effects in the British General Election Campaign.” *British Politics* 1 (2): 195–221. https://doi.org/10.1057/palgrave.bp.4200022.

Nye, Joseph S. 2003. *The Paradox of American Power: Why the World’s Only Superpower Can’t Go It Alone*. Cary: Oxford University Press, Incorporated. http://ebookcentral.proquest.com/lib/asulib-ebooks/detail.action?docID=3052325.

———. 2008. “Public Diplomacy and Soft Power.” *The Annals of the American Academy of Political and Social Science* 616 (1): 94–109.

———. 2011. *The Future of Power*. e-book.: PublicAffairs.

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J. Wood. 2020. “Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability.” *Political Behavior* 42 (3): 939–60. https://doi.org/10.1007/s11109-019-09528-x.

Nyhan, Brendan, and Jason Reifler. 2010. “When Corrections Fail: The Persistence of Political Misperceptions.” *Political Behavior* 32 (2): 303–30. https://doi.org/10.1007/s11109-010-9112-2.

Ognyanova, Katherine, David Lazer, Ronald E. Robertson, and Christo Wilson. 2020. “Misinformation in Action: Fake News Exposure Is Linked to Lower Trust in Media, Higher Trust in Government When Your Side Is in Power.” *Harvard Kennedy School Misinformation Review*.

Orhan, Yunus Emre. 2022. “The Relationship between Affective Polarization and Democratic Backsliding: Comparative Evidence.” *Democratization* 29 (4): 714–35. https://doi.org/10.1080/13510347.2021.2008912.

Panagopoulos, Costas. 2016. “All about That Base: Changing Campaign Strategies in U.S. Presidential Elections.” *Party Politics* 22 (2): 179–90. https://doi.org/10.1177/1354068815605676.

Panayotov, and Nikolov. 1985. “KGB, Information Nr. 2955 [to Bulgarian State Security],” September 7, 1985. https://digitalarchive.wilsoncenter.org/document/kgb-information-nr-2955-bulgarian-state-security.

Paolini, Stefania, Jake Harwood, and Mark Rubin. 2010. “Negative Intergroup Contact Makes Group Memberships Salient: Explaining Why Intergroup Conflict Endures.” *Personality and Social Psychology Bulletin* 36 (12): 1723–38.

Pasquetto, Irene, Briony Swire-Thompson, and Michelle A. Amazeen. 2020. “Tackling Misinformation: What Researchers Could Do with Social Media Data.” *Harvard*

Kennedy School Misinformation Review, December. https://doi.org/10.37016/mr-2020-49.

Paterson, Pat. 2008. “The Truth About Tonkin.” *Naval History Magazine* 22 (1). https://www.usni.org/magazines/naval-history-magazine/2008/february/truth-about-tonkin.

Patriot Magazine. 1983. “AIDS May Invade India,” July 17, 1983, selvi edition.

Paul, Christopher, and Miriam Matthews. 2016. “The Russian ‘Firehose of Falsehood’ Propaganda Model.” *Rand Corporation*, 2–7.

Pennycook, Gordon, Tyrone D. Cannon, and David G. Rand. 2018. “Prior Exposure Increases Perceived Accuracy of Fake News.” *Journal of Experimental Psychology: General* 147 (12): 1865.

Pennycook, Gordon, and David G. Rand. 2019. “Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning.” *Cognition* 188: 39–50.

———. 2021a. “The Psychology of Fake News.” *Trends in Cognitive Sciences* 25 (5): 388–402. https://doi.org/10.1016/j.tics.2021.02.007.

———. 2021b. “Lack of Partisan Bias in the Identification of Fake (versus Real) News.” *Trends in Cognitive Sciences* 25 (9): 725–26. https://doi.org/10.1016/j.tics.2021.06.003.

Peterson, Erik, and Shanto Iyengar. 2021. “Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading?” *American Journal of Political Science* 65 (1): 133–47. https://doi.org/10.1111/ajps.12535.

Pew Research Center. 2021a. “Country-Specific Methodology: 2021 Global Attitudes Survey.” *Pew Research Center Methods* (blog). 2021. https://www.pewresearch.org/methods/interactives/international-methodology/.

———. 2021b. “Global Attitudes Spring 2021.” Washington, D.C. https://www.pewresearch.org/global/wp-content/uploads/sites/2/2022/03/Pew-Research-Center-Global-Attitudes-Spring-2021-Survey-Data.zip.

Pitkin, Hanna Fenichel. 1967. *The Concept of Representation*. *The Concept of Representation*. University of California Press. https://doi.org/10.1525/9780520340503.

Powell, Robert. 2002. “Bargaining Theory and International Conflict.” *Annual Review of Political Science* 5 (1): 1–30. https://doi.org/10.1146/annurev.polisci.5.092601.141138.

Preacher, Kristopher J., and Andrew F. Hayes. 2008. *Assessing Mediation in Communication Research*. The Sage sourcebook of advanced data analysis methods for communication ….

Press, Daryl G. 2004. "The Credibility of Power: Assessing Threats during the 'Appeasement' Crises of the 1930s." *International Security* 29 (3): 136–69.

Prins, Brandon C. 2003. "Institutional Instability and the Credibility of Audience Costs: Political Participation and Interstate Crisis Bargaining, 1816-1992." *Journal of Peace Research* 40 (1): 67–84.

Putnam, Robert D. 1988. "Diplomacy and Domestic Politics: The Logic of Two-Level Games." *International Organization* 42 (3): 427–60.

———. 2000. *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster.

Quek, Kai. 2016. "Are Costly Signals More Credible? Evidence of Sender-Receiver Gaps." *Journal of Politics* 78 (3): 925–40. https://doi.org/10.1086/685751.

———. 2021. "Four Costly Signaling Mechanisms." *American Political Science Review* 115 (2): 537–49. https://doi.org/10.1017/S0003055420001094.

Renshon, Jonathan, Allan Dafoe, and Paul Huth. 2018. "Leader Influence and Reputation Formation in World Politics." *American Journal of Political Science* 62 (2): 325–39.

Rød, Espen Geelmuyden, and Nils B. Weidmann. 2015. "Empowering Activists or Autocrats? The Internet in Authoritarian Regimes." *Journal of Peace Research* 52 (3): 338–51.

Rogers, Richard. 2020. "Research Note: The Scale of Facebook's Problem Depends upon How 'Fake News' Is Classified." *The Harvard Kennedy School (HKS) Misinformation Review*.

Roggeveen, Anne L., and Gita Venkataramani Johar. 2002. "Perceived Source Variability versus Familiarity: Testing Competing Explanations for the Truth Effect." *Journal of Consumer Psychology* 12 (2): 81–91.

Rogowski, Jon C. 2018. "Voter Decision-Making with Polarized Choices." *British Journal of Political Science* 48 (1): 1–22. https://doi.org/10.1017/S0007123415000630.

Rogowski, Jon C., and Joseph L. Sutherland. 2016. "How Ideology Fuels Affective Polarization." *Political Behavior* 38 (2): 485–508. https://doi.org/10.1007/s11109-015-9323-7.

Roseman, Ira J. 1996. “Appraisal Determinants of Emotions: Constructing a More Accurate and Comprehensive Theory.” *Cognition & Emotion* 10 (3): 241–78. https://doi.org/10.1080/026999396380240.

Rosenberg, Milton J., and Robert P. Abelson. 1960. “An Analysis of Cognitive Balancing.” *Attitude Organization and Change: An Analysis of Consistency among Attitude Components* 3: 112–63.

Ross, Andrew R. N., Cristian Vaccari, and Andrew Chadwick. 2022. “Russian Meddling in U.S. Elections: How News of Disinformation’s Impact Can Affect Trust in Electoral Outcomes and Satisfaction with Democracy.” *Mass Communication and Society* 25 (6): 786–811. https://doi.org/10.1080/15205436.2022.2119871.

Ruhe, Blaise Misztal, Jonathan. 2021. “Is Iran Bluffing About Its Enriched Uranium Stockpile?” *Foreign Policy* (blog). July 28, 2021. https://foreignpolicy.com/2021/07/28/iran-biden-nuclear-deal-weapons-jcpoa-bluffing-enriched-uranium-stockpile-sanctions/.

Samoilenko, Sergei A., and Margarita Karnysheva. 2020. “Character Assassination as Modus Operandi of Soviet Propaganda.” *The SAGE Handbook of Propaganda*, 189–204.

Sanchez, Carmen, and David Dunning. 2021. “Cognitive and Emotional Correlates of Belief in Political Misinformation: Who Endorses Partisan Misbeliefs?” *Emotion*.

Sartori, Anne E. 2002. “The Might of the Pen: A Reputational Theory of Communication in International Disputes.” *International Organization* 56 (1): 121–49. https://doi.org/10.1162/002081802753485151.

Schaffner, Brian F, and Samantha Luks. 2018. “Misinformation or Expressive Responding?” *Public Opinion Quarterly* 82 (1): 135–47. https://doi.org/10.1093/poq/nfx042.

Schilke, Oliver, Martin Reimann, and Karen S. Cook. 2021. “Trust in Social Relations.” *Annual Review of Sociology* 47 (1): 239–59. https://doi.org/10.1146/annurev-soc-082120-082850.

Schub, Robert. 2020. “When Prospective Leader Turnover Promotes Peace.” *International Studies Quarterly* 64 (3): 510–22. https://doi.org/10.1093/isq/sqaa027.

Schultz, Kenneth A. 2017. “Perils of Polarization for U.S. Foreign Policy.” *The Washington Quarterly* 40 (4): 7–28. https://doi.org/10.1080/0163660X.2017.1406705.

Selvage, Douglas. 2019. “Operation ‘Denver’: The East German Ministry of State Security and the KGB’s AIDS Disinformation Campaign, 1985–1986 (Part 1).”

Journal of Cold War Studies 21 (4): 71–123. https://doi.org/10.1162/jcws_a_00907.

Sheagley, Geoffrey, and Scott Clifford. 2023. "No Evidence That Measuring Moderators Alters Treatment Effects." *American Journal of Political Science,* July. https://doi.org/10.1111/ajps.12814.

Sion, Maurice, and Philip Wolfe. 1957. "On a Game without a Value." *Contributions to the Theory of Games* 3 (1957): 299–306.

Smith, Bradley C., and William Spaniel. 2019. "Militarized Disputes, Uncertainty, and Leader Tenure." *Journal of Conflict Resolution* 63 (5): 1222–52. https://doi.org/10.1177/0022002718789738.

Solvak, Mihkel. 2009. "Events and Reliability of Measures: The Effect of Elections on Measures of Interest in Politics." *International Journal of Public Opinion Research* 21 (3): 316–32. https://doi.org/10.1093/ijpor/edp033.

Soontjens, Karolin, and Stefaan Walgrave. 2021. "Listening to the People: Politicians' Investment in Monitoring Public Opinion and Their Beliefs about Accountability." *The Journal of Legislative Studies* 0 (0): 1–21. https://doi.org/10.1080/13572334.2021.2011649.

Spaniel, William, and Iris Malone. 2019. "The Uncertainty Trade-off: Reexamining Opportunity Costs and War." *International Studies Quarterly* 63 (4): 1025–34. https://doi.org/10.1093/isq/sqz050.

Ștefăniță, Oana, Nicoleta Corbu, and Raluca Buturoiu. 2018. "Fake News and the Third-Person Effect: They Are More Influenced than Me and You." *Journal of Media Research* 11 (3(32)): 5–23. https://doi.org/10.24193/jmr.32.1.

Stephenson, William Samuel. 1999. *British Security Coordination: The Secret History of British Intelligence in the Americas, 1940-1945*. New York: Fromm International.

Stiers, Dieter, and Ruth Dassonneville. 2020. "Retrospective Voting and the Polarization of Available Alternatives." *Canadian Journal of Political Science/Revue Canadienne de Science Politique* 53 (1): 99–115. https://doi.org/10.1017/S0008423919000556.

Stiles, Kendall. 2018. *Trust and Hedging in International Relations*. University of Michigan Press.

Stroud, Natalie Jomini, Joshua M Scacco, and Yujin Kim. 2022. "Passive Learning and Incidental Exposure to News." *Journal of Communication* 72 (4): 451–60. https://doi.org/10.1093/joc/jqac015.

Sturgis, P., and P. Smith. 2010. “Assessing the Validity of Generalized Trust Questions: What Kind of Trust Are We Measuring?” *International Journal of Public Opinion Research* 22 (1): 74–92. https://doi.org/10.1093/ijpor/edq003.

Suhay, Elizabeth, Emily Bello-Pardo, and Brianna Maurer. 2018. “The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments.” *The International Journal of Press/Politics* 23 (1): 95–115. https://doi.org/10.1177/1940161217740697.

Sun, Ye, Zhongdang Pan, and Lijiang Shen. 2008. “Understanding the Third-Person Perception: Evidence From a Meta-Analysis.” *Journal of Communication* 58 (2): 280–300. https://doi.org/10.1111/j.1460-2466.2008.00385.x.

Taber, Charles S., and Milton Lodge. 2006. “Motivated Skepticism in the Evaluation of Political Beliefs.” *American Journal of Political Science* 50 (3): 755–69. https://doi.org/10.1111/j.1540-5907.2006.00214.x.

Tajfel, Henri, John C. Turner, William G. Austin, and Stephen Worchel. 1979. “An Integrative Theory of Intergroup Conflict.” *Organizational Identity: A Reader* 56 (65): 9780203505984–16.

Takano, Masanori, Yuki Ogawa, Fumiaki Taka, and Soichiro Morishita. 2021. “Effects of Incidental Brief Exposure to News on News Knowledge While Scrolling Through Videos.” *IEEE Access* 9: 37772–83. https://doi.org/10.1109/ACCESS.2021.3063484.

Tang, Shuo, Lars Willnat, and Hongzhong Zhang. 2021. “Fake News, Information Overload, and the Third-Person Effect in China.” *Global Media and China* 6 (4): 492–507. https://doi.org/10.1177/20594364211047369.

Tewksbury, David, Andrew J. Weaver, and Brett D. Maddex. 2001. “Accidentally Informed: Incidental News Exposure on the World Wide Web.” *Journalism & Mass Communication Quarterly* 78 (3): 533–54. https://doi.org/10.1177/107769900107800309.

The Economist. 2015. “The Economics of Bluffing,” May 28, 2015. https://www.economist.com/finance-and-economics/2015/05/28/the-economics-of-bluffing.

———. 2022. “Never-Ending Nuclear Talks with Iran Are Bordering on the Absurd,” Seo 2022. https://www.economist.com/middle-east-and-africa/2022/09/08/never-ending-nuclear-talks-with-iran-are-bordering-on-the-absurd.

———. 2023. “China Is Flooding Taiwan with Disinformation,” September 26, 2023. https://www.economist.com/asia/2023/09/26/china-is-flooding-taiwan-with-disinformation.

Toal, Gerard, and John O'Loughlin. 2018. "'Why Did MH17 Crash?': Blame Attribution, Television News and Public Opinion in Southeastern Ukraine, Crimea and the De Facto States of Abkhazia, South Ossetia and Transnistria." *Geopolitics* 23 (4): 882–916. https://doi.org/10.1080/14650045.2017.1364238.

Tomz, Michael. 2012. *Reputation and International Cooperation - Google Books*. Princeton University Press. https://www.google.com/books/edition/Reputation_and_International_Cooperation/BeszAAAAQBAJ?hl=en&gbpv=1&dq=tomz+sovereign+debt&printsec=frontcover.

Tomz, Michael R., and Jessica L. P. Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107 (4): 849–65. https://doi.org/10.1017/S0003055413000488.

Treyger, Elina, Joe Cheravitch, and Raphael S. Cohen. 2022. *Russian Disinformation Efforts on Social Media*. Combating Foreign Disinformation on Social Media Series. Santa Monica, Calif: RAND Project Air Force.

Trilling, Damian, Marijn Van Klingeren, and Yariv Tsfati. 2017. "Selective Exposure, Political Polarization, and Possible Mediators: Evidence from the Netherlands." *International Journal of Public Opinion Research* 29 (2): 189–213.

Tropp, Linda R. 2008. "The Role of Trust in Intergroup Contact: Its Significance and Implications for Improving Relations between Groups." In *Improving Intergroup Relations: Building on the Legacy of Thomas F. Pettigrew*, edited by Ulrich Wagner, Linda R. Tropp, Gillian Finchilescu, and Colin Tredoux, 1st ed., 91–106. Wiley. https://doi.org/10.1002/9781444303117.

Tsfati, Yariv, and Joseph N. Cappella. 2003. "Do People Watch What They Do Not Trust? Exploring the Association between News Media Skepticism and Exposure." *Communication Research* 30 (5): 504–29.

Tucker, Joshua, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3144139.

Turner, John C., Rupert J. Brown, and Henri Tajfel. 1979. "Social Comparison and Group Interest in Ingroup Favouritism." *European Journal of Social Psychology* 9 (2): 187–204.

Tyler, James M., Robert S. Feldman, and Andreas Reichert. 2006. "The Price of Deceptive Behavior: Disliking and Lying to People Who Lie to Us." *Journal of Experimental Social Psychology* 42 (1): 69–77. https://doi.org/10.1016/j.jesp.2005.02.003.

U.S. Department of State. 1987. “A Report on Active Measures and Propaganda, 1986-1987.” 9627. Soviet Influence Activities. Washington, DC: U.S. Department of State. https://www.globalsecurity.org/intell/library/reports/1987/soviet-influence-activities-1987.pdf.

———. 2020. “A List of Treaties and Other International Agreements of the United States in Force on January 1, 2020.” Treaties in Force. Washington, D.C. https://www.state.gov/wp-content/uploads/2020/08/TIF-2020-Full-website-view.pdf.

———. 2022. “Supplemental List of Treaties and Other International Agreements.” Treaties in Force. Washington, D.C. https://www.state.gov/wp-content/uploads/2022/05/TIF-Supplement-2022.pdf.

Uslaner, Eric M. 2015. “Measuring Generalized Trust: In Defense of the ‘Standard’Question.” In *Handbook of Research Methods on Trust*. Edward Elgar Publishing.

Valentino, Nicholas A., Ted Brader, Eric W. Groenendyk, Krysha Gregorowicz, and Vincent L. Hutchings. 2011. “Election Night’s Alright for Fighting: The Role of Emotions in Political Participation.” *Journal of Politics* 73 (1): 156–70. https://doi.org/10.1017/S0022381610000939.

VanderWeele, Tyler. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. “The Spread of True and False News Online.” *Science* 359 (6380): 1146–51.

Vraga, Emily K., and Leticia Bode. 2020. “Defining Misinformation and Understanding Its Bounded Nature: Using Expertise and Evidence for Describing Misinformation.” *Political Communication* 37 (1): 136–44.

Vuving, Alexander. 2019. “The Logic of Attraction: Outline of a Theory of Soft Power.” *Available at SSRN 3637662*.

Walker, Christopher. 2018. “What Is" Sharp Power"?” *Journal of Democracy* 29 (3): 9–23.

Walker, Christopher, and Jessica Ludwig. 2017a. “The Meaning of Sharp Power: How Authoritarian States Project Influence.” *Foreign Affairs*, November. https://www.foreignaffairs.com/articles/china/2017-11-16/meaning-sharp-power?cid=int-fls&pgtype=hpg%3E.

———. 2017b. “Sharp Power: Rising Authoritarian Influence.” Sharp Power. International Forum for Democratic Studies. https://www.ned.org/wp-

content/uploads/2017/12/Introduction-Sharp-Power-Rising-Authoritarian-Influence.pdf.

Wang, Wei-Chun, Nadia M. Brashier, Erik A. Wing, Elizabeth J. Marsh, and Roberto Cabeza. 2016. "On Known Unknowns: Fluency and the Neural Mechanisms of Illusory Truth." *Journal of Cognitive Neuroscience* 28 (5): 739–46.

Wardle, Claire. 2018. "The Need for Smarter Definitions and Practical, Timely Empirical Research on Information Disorder." *Digital Journalism* 6 (8): 951–63.

Webb, Margaret A., and June P. Tangney. 2022. "Too Good to Be True: Bots and Bad Data From Mechanical Turk." *Perspectives on Psychological Science*, November, 17456916221120027. https://doi.org/10.1177/17456916221120027.

Weber, Christopher. 2013. "Emotions, Campaigns, and Political Participation." *Political Research Quarterly* 66 (2): 414–28. https://doi.org/10.1177/1065912912449697.

Webster, Steven W., and Alan I. Abramowitz. 2017. "The Ideological Foundations of Affective Polarization in the U.S. Electorate." *American Politics Research* 45 (4): 621–47. https://doi.org/10.1177/1532673X17703132.

Weedon, Jen, William Nuland, and Alex Stamos. 2017. "Information Operations and Facebook." Facebook Security. https://i2.res.24o.it/pdf2010/Editrice/ILSOLE24ORE/ILSOLE24ORE/Online/_Oggetti_Embedded/Documenti/2017/04/28/facebook-and-information-operations-v1.pdf.

Weeks, Brian E., Daniel S. Lane, and Lauren B. Hahn. 2022. "Online Incidental Exposure to News Can Minimize Interest-Based Political Knowledge Gaps: Evidence from Two U.S. Elections." *The International Journal of Press/Politics* 27 (1): 243–62. https://doi.org/10.1177/1940161221991550.

Weeks, Jessica L. 2008. "Autocratic Audience Costs: Regime Type and Signaling Resolve." *International Organization* 62 (01). https://doi.org/10.1017/S0020818308080028.

———. 2012. "Strongmen and Straw Men: Authoritarian Regimes and the Initiation of International Conflict." *American Political Science Review* 106 (2): 326–47. https://doi.org/10.1017/S0003055412000111.

Wei, R., S. C. Chia, and V.-H. Lo. 2011. "Third-Person Effect and Hostile Media Perception Influences on Voter Attitudes toward Polls in the 2008 U.S. Presidential Election." *International Journal of Public Opinion Research* 23 (2): 169–90. https://doi.org/10.1093/ijpor/edq044.

Wei, Ran, Ven-Hwei Lo, and Yicheng Zhu. 2019. "Need for Orientation and Third-Person Effects of the Televised Debates in the 2016 U.S. Presidential Election."

Mass Communication and Society 22 (5): 565–83. https://doi.org/10.1080/15205436.2019.1601227.

Weisiger, Alex, and Keren Yarhi-Milo. 2015. “Revisiting Reputation: How Past Actions Matter in International Politics.” *International Organization* 69 (2): 473–95. https://doi.org/10.1017/S0020818314000393.

Weiss, Jessica Chen. 2013. “Authoritarian Signaling, Mass Audiences, and Nationalist Protest in China.” *International Organization* 67 (1): 1–35.

Weiss, Leonard. 2013. “The Lavon Affair: How a False-Flag Operation Led to War and the Israeli Bomb.” *Bulletin of the Atomic Scientists* 69 (4): 58–68. https://doi.org/10.1177/0096340213493259.

Weith, Paul T, and Andre Krouwel. 2013. “Wrong Assumptions, Poor Results: An Empirical Assessment of the Dimensionality of Political Knowledge.” In , 1–20. Chicago, Illinois.

White House. 2018. “Remarks by President Trump on the Joint Comprehensive Plan of Action – The White House.” Trump White House Archives. May 8, 2018. https://trumpwhitehouse.archives.gov/briefings-statements/remarks-president-trump-joint-comprehensive-plan-action/.

Whittlesea, Bruce WA. 1993. “Illusions of Familiarity.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19 (6): 1235.

Williams, Richard. 2016. “Understanding and Interpreting Generalized Ordered Logit Models.” *The Journal of Mathematical Sociology* 40 (1): 7–20. https://doi.org/10.1080/0022250X.2015.1112384.

Winter, Nicholas, Tyler Burleigh, Ryan Kennedy, and Scott Clifford. 2019. “A Simplified Protocol to Screen Out VPS and International Respondents Using Qualtrics.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3327274.

Wolfe, Philip. 1955. “The Strict Determinateness of Certain Infinite Games.” *Pacific J. Math* 5: 841–47.

Wolford, Scott. 2007. “The Turnover Trap: New Leaders, Reputation, and International Conflict.” *American Journal of Political Science* 51 (4): 772–88. https://doi.org/10.1111/j.1540-5907.2007.00280.x.

———. 2014. “Showing Restraint, Signaling Resolve: Coalitions, Cooperation, and Crisis Bargaining.” *American Journal of Political Science* 58 (1): 144–56. https://doi.org/10.1111/ajps.12049.

Wu, Cathy Xuanxuan, Amanda A Licht, and Scott Wolford. 2021. “Same as the Old Boss? Domestic Politics and the Turnover Trap.” *International Studies Quarterly* 65 (1): 173–83. https://doi.org/10.1093/isq/sqaa074.

Wu, Shenghong. 2023. “專家：疑美論創造符合中國利益世界觀 | 政治.” *中央社 CNA*, September 21, 2023. https://www.cna.com.tw/news/aipl/202309210234.aspx.

Wu, Yenna. 2019. “Recognizing and Resisting China’s Evolving Sharp Power.” *American Journal of Chinese Studies*, 129–53.

Xiao, Xizhu, Yan Su, and Danielle Ka Lai Lee. 2021. “Who Consumes New Media Content More Wisely? Examining Personality Factors, SNS Use, and New Media Literacy in the Era of Misinformation.” *Social Media+ Society* 7 (1): 2056305121990635.

Yarhi-Milo, Keren. 2018. “After Credibility: American Foreign Policy in the Trump Era.” *Foreign Affairs* 97 (1): 68–77.

Yong, Nicholas. 2023. “Afghanistan: Girls’ Education Activist Arrested by Taliban.” BBC. March 28, 2023. https://www.bbc.com/news/world-asia-65095663.

Yoo, Chan Yun. 2009. “Effects beyond Click-through: Incidental Exposure to Web Advertising.” *Journal of Marketing Communications* 15 (4): 227–46. https://doi.org/10.1080/13527260802176419.

Yu, Ansha, and Sami Tas. 2015. “Taking Into Account Time Spent on Stories.” *Meta* (blog). June 12, 2015. https://about.fb.com/news/2015/06/news-feed-fyi-taking-into-account-time-spent-on-stories/.

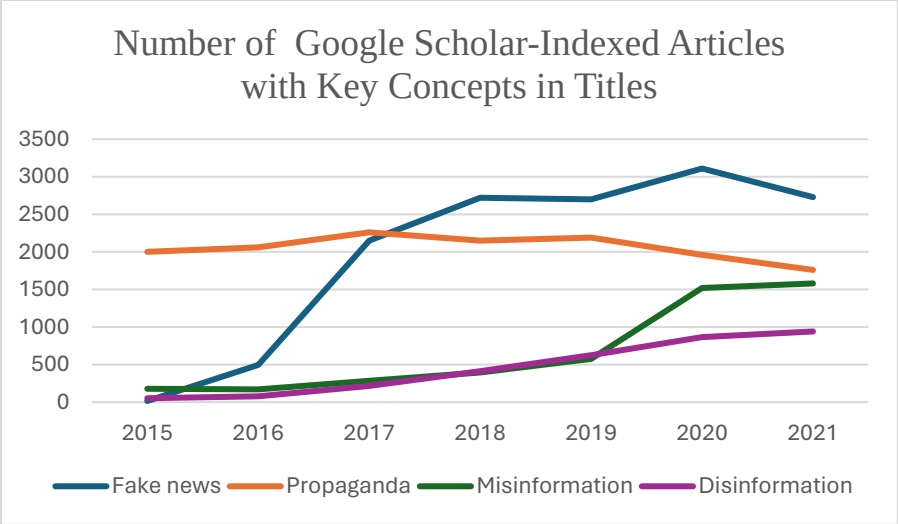
Zagare, Frank C. 2019. “The Moroccan Crisis of 1905–6.” In *Game Theory, Diplomatic History and Security Studies*, edited by Frank C. Zagare, 0. Oxford University Press. https://doi.org/10.1093/oso/9780198831587.003.0004.

Zelenkauskaite, Asta. 2022. *Creating Chaos Online: Disinformation and Subverted Post-Publics*. University of Michigan Press.

Zhu, Qinfeng, Brian E. Weeks, and Nojin Kwak. 2021. “Implications of Online Incidental and Selective Exposure for Political Emotions: Affective Polarization during Elections.” *New Media & Society* 0 (23): 1–23. https://doi.org/10.1177/14614448211061336.

APPENDIX A

GOOGLE SCHOLAR TRENDS: DISINFORMATION-RELATED PUBLICATIONS



.

(Google Scholar 2022)

APPENDIX B

DISINFORMATIONAL METHODS

Objective	Bargaining Spectrum		Method *italics indicate unique methods* *[1], [2], [3], [4a], [5b] indicate method in common with those objectives*	Points of Manip.				
	Element Modified	Modification		Pref. Chng.,	Popularity	Politicization	Chief Negoti. \Autonomy	Chief Negoti. \Credibility
I. GET MORE *Increase chance of opponent concessions*	Win-set (*opponent*)	Widen/ shift towards own	False side payment assurances [2]	X				
			Pro-agreement/anti-status quo disinfo [2,3]	X				
			Promote own soft power [2,3]	X	X			
			Promote sympathetic/undermine antagonist Lvl I/II candidates [2,3,4a-b]	X				
	Win-set (*own*)	Narrow	Misrepresent own bargaining range [2]	X				
	Uncertainty (*own*)	Increase	*False concern of opponent defection*	X				
II. GET SOMETHING *Increase chance of opponent ratification*	Win-set (*opponent*)	Widen/ shift towards own	False side payment assurances [1]	X				
			Pro-agreement/anti-status quo disinfo [1,3]	X				
			Promote own soft power [1,3]	X	X			
			Promote sympathetic/undermine antagonist Lvl I/II candidates [1,3,4a-b]	X				
	Win-set (*own*)	Widen	Misrepresent own bargaining range [1]	X				
III. TAKE SOMETHING *Increase chance of unilateral status quo revision*	Uncertainty (*opponent*)	Increase	*Conflicting disinfo. about status quo*	X				
	Win-set (*third-party*)	Widen/ shift towards own	Pro-revision/anti-status quo disinfo [1,2]	X				
			Promote own soft power [1,2]	X	X			
			Promote sympathetic/undermine antagonist Lvl I/II candidates [1,2,4a-b]	X				
IVa. KEEP WHAT YOU HAVE *Decrease chances of status quo revision attempts*	Win-set (*opponent*)	Narrow	Anti-agreement/pro-status quo disinfo [4b]	X				
			Promote sympathetic/undermine antagonist Lvl I/II candidates [1,2,3,4b]	X				
IVb. KEEP THE OTHER GUY FROM GETTING SOMETHING *Decrease chances of*	Win-set (*opponent*)	Narrow	Anti-agreement/pro-status quo disinfo [4a]	X				
			Promote sympathetic/undermine antagonist Lvl I/II candidates [1,2,3,4a]	X				
			Undercut opponent soft power	X	X			
	Uncertainty (*opponent*)	Increase	*Disinfo. regarding ratification challenges (increase invol. defection concerns)*	X	X	X		

opponent cooperation			*Encourage polarization to increase opponent chief negotiator autonomy (vol./invol. defection concerns)*	X	X	X	X	X	
			Disinfo. regarding chief negotiator reliability (increase vol. defection concerns)	X				X	

APPENDIX C

RESULTS: NOVEL MEDIATION ANALYSIS (NISBET ET AL. DATA)

Nisbet et al. (E. C. Nisbet, Mortenson, and Li 2021; E. Nisbet 2020) asked respondents how influential they thought misinformation from various sources were, to include liberal and conservative "political groups and organizations" (2021, 12). These unique source questions served as conditions (each respondent was only asked about one source), and Nisbet et al. used the conditions to establish that the statistically significant, negative relationship between perceived influence of misinformation (PIM) and satisfaction in the state of U.S. democracy was not an artifact of their question wording. They also gathered information on their respondents' partisanship, using a seven-point scale from strong Republican to strong Democrat, with independent and independent leaners as the middle three categories.

I used those question conditions and the respondents' partisanship information to generate a new variable: *counter*. *Counter* was whether a Republican or Democrat was asked about the influence of misinformation from outgroups (e.g., whether Republicans were asked about the influence of misinformation from liberal sources, and vice versa). I then generated a third-person effect variable (*TPE*), which was the difference between the respondents' reported PIM on themselves and PIM on other voters. Then, using the control variables that Nisbet et al. had employed (as well as for a dichotomous measure of partisanship), I conducted a mediation analysis following the method of Preacher and Hayes (2004). My independent variable (IV) was *counter*, dependent variable was *satisfaction in democracy*, and mediator was *TPE*. The sample size was 815.

The analysis did not meet all their assumptions, but as the data meet no mediation model's assumptions, I proceeded to check face validity of my expectation that respondents would have higher perception of TPE with regards to partisan outgroup

misinformation, and in turn negative political assessments based on the PIM. Particularly, it failed the assumption that all key variables be continuous because *TPE* is an ordinal measure. The assumptions it met were: no multicollinearity, approximately normal distributions for *TPE* and *satisfaction in* democracy, and a linear relationship between the same.

The analysis was conducted using the *sgmediation2* package in Stata (Mize n.d.). It runs an OLS regression, and then conducts Sobel-Goodmen tests of the main IV, mediator, and DV to test mediation. It also facilitates follow-on bootstrapping to produce percentile-based confidence intervals, in accordance with Preacher and Hayes (2004). 5000 bootstrapped samples produced the following values:

Table 6. Mediation Analysis Results

	Coefficient	**S.E.**	**p>z**
X->M (*counter* to *TPE*)	0.592***	0.081	0.000
M->Y (*TPE* to *satis. w/dem.*)	-0.543***	0.062	0.000
Indirect Effect	-0.322***	0.056	0.000
Direct Effect (*c'*)	0.049	0.147	0.739
Total Effect (*c*)	-0.273*	0.149	0.067

The proportion of the total effect mediated was 1.18, which suggests an interaction effect between the IV and mediator (*counter* and *TPE*) in addition to mediation.

What these results show is that the partisan congeniality of misinformation increases TPE, which in turn increases a negative assessment regarding what that influence means for politics, as captured in *satisfaction with democracy*. These results illustrate my expectation that exposure to counter-attitudinal misinformation will tend toward increased negative perceptions of partisan out-groups because one assesses they are more strongly influenced by it than one's in-group would be.

The results also illustrate how disaggregating not only source congeniality, but content congeniality, could account for the conflicting extant results regarding the backlash potential of passive exposure to counter-attitudinal information. When not disaggregated, Nisbet et al. found that the PIM of their various source (conservative and liberal, but also domestic, foreign, and general sources) were statistically indistinguishable from each other. They also found that PIM on other voters' relationship with *satisfaction with democracy* was not conditioned on partisanship at all. This suggests partisan in-group/out-group dynamics play little role at all. However, when disaggregated, the importance of in-/out-group dynamics is evident. Partisans asked about the influence of outgroup misinformation had higher perceptions of TPE. Even discarding the mediation findings, this statistically significant correlation offers support for the inference that scholars can find more variation based on content congeniality (versus identity coincidence congeniality), as partisan outgroup misinformation is an example of out-group informational content whose influence one would almost certainly not find desirable.

APPENDIX D

DISCUSSION: MTURK TREATMENT VALIDATION SURVEYS

I fielded the treatment validation survey on Mturk twice (n = 100 both times). The results suggested the respondents were often inattentive at best and bots at worst. In the first sample, the primary warning sign was the sample's unusually unvarying assessments of misinformation plausibility and slant (below). The average plausibility means ranged only 56.27-62.91 on a scale of 0 to 100: that is: roughly equally likely to be true/untrue. This relative constancy held even when mean plausibility was disaggregated by partisanship. Both Republicans and Democrats largely agreed that all the misinformation stories were neither plausible nor implausible, regardless of their intended partisan slant.

Plausibility Difference-in-Means (0/100 = more likely to be untrue/true)

	All	Republicans	Democrats	(R)-(D)
UKR Sanctions = Oil Exec. Profits	61.281	61.676	61.684	-0.01
	(2.255)	(3.318)	(2.981)	[1.00]
Uvalde false flag assertion	57.315	56.235	59.123	-2.89
	(2.502)	(4.142)	(3.154)	[0.58]
Evangelical sanctions Iraq	59.236	57.667^	60.737	-3.07
	(2.386)	(4.038)	(2.978)	[0.54]
GOP wants to criminalize ectopic abortions	62.236	56.294	66.614	-10.32
	(2.322)	(4.222)	(2.574)	[0.03]***
AL Senator arm homeless	57.629	58.500	58.228	0.27
	(2.532)	(4.301)	(3.142)	[0.96]
Trump aids burn class. docs	62.910	62.559	63.912	-1.35
	(2.255)	(3.656)	(2.847)	[0.77]
Fox host sympathetic to Nazis	58.270	56.559	60.596	-4.04
	(2.499)	(4.182)	(3.151)	[0.44]
TX textbooks on slavery	60.371	58.176	62.737	-4.56
	(2.492)	(3.911)	(3.206)	[0.38]
MSNBC host fired for beliefs	61.360	63.382	60.772	2.61
	(2.378)	(3.394)	(3.170)	[0.59]
Charity fentanyl front	58.820	59.941	58.807	1.13
	(2.450)	(3.629)	(3.225)	[0.82]
Election surveillance coverup by media	58.899	60.364^	58.596	1.77
	(2.276)	(3.667)	(2.923)	[0.71]
Fear of Trump delayed UKR invasion	61.315	58.618	63.877	-5.26
	(2.421)	(4.018)	(2.997)	[0.29]
State Sen. pregnancy solidarity	61.157	59.853	62.684	-2.83
	(2.421)	(3.631)	(3.168)	[0.57]
LGBTQ+ affirmation test in ed.	60.146	58.324	62.035	-3.71
	(2.306)	(3.758)	(2.888)	[0.44]
Prisoner release for votes	60.281	59.882	61.404	-1.52
	(2.358)	(3.615)	(3.070)	[0.75]
CA climate corruption	56.270	54.618	58.386	-3.77
	(2.577)	(4.209)	(3.258)	[0.48]
N	89	33^-34	57	

Notes: R=Republicans; D=Democrats. Standard errors in parentheses. P values in brackets for t tests of difference in means.

The partisan slant results, too, were unusual. The misinformation means again were tightly-ranged from 56.76-64.489, and again showed exceptionally little difference within partisan subgroups (below).

Partisan Favorability Difference-in-Means (0/100 = Pro-Repub./Pro-Dem.)

	All	Republicans	Democrats	(R)-(D)
UKR Sanctions = Oil Exec. Profits (R-D)	62.182	62.324	62.093	0.23
	(2.254)	(4.326)	(2.500)	[0.96]
Uvalde false flag assertion (R-D)	58.932	55.088	61.352	-6.26
	(2.681)	(4.632)	(3.246)	[0.26]
Evangelical sanctions Iraq (R-D)	65.580	60.147	69.000	-8.85*
	(2.401)	(4.168)	(2.835)	[0.07]
GOP wants to criminalize ectopic abortions (R-D)	62.375	60.353	63.648	-3.30
	(2.346)	(4.077)	(2.851)	[0.50]
AL Senator arm homeless (R-D)	61.807	59.735	63.111	-3.38
	(2.537)	(4.394)	(3.093)	[0.52]
Trump aids burn class. docs (R-D)	63.170	61.265	64.370	-3.11
	(2.237)	(4.014)	(2.646)	[0.50]
Fox host sympathetic to Nazis(R-D)	63.750	59.765	66.259	-6.49
	(2.394)	(4.535)	(2.638)	[0.19]
TX textbooks on slavery (R-D)	60.693	58.147	62.296	-4.15
	(2.446)	(4.236)	(2.975)	[0.41]
MSNBC host fired for beliefs (R-D)	61.443	59.118	62.907	-3.79
	(2.628)	(4.634)	(3.155)	[0.49]
Charity fentanyl front (R-D)	61.227	63.618	59.722	3.90
	(2.415)	(3.920)	(3.076)	[0.44]
Election surveillance coverup by media (R-D)	64.489	67.559	62.556	5.00
	(2.415)	(3.846)	(3.101)	[0.32]
Fear of Trump delayed UKR invasion (R-D)	56.761	58.059	55.944	2.11
	(2.494)	(4.159)	(3.135)	[0.68]
State Sen. pregnancy solidarity (R-D)	58.455	57.941	58.778	-0.84
	(2.458)	(3.847)	(3.218)	[0.87]
LGBTQ+ affirmation test in ed. (R-D)	61.614	57.441	64.241	-6.80
	(2.351)	(4.183)	(2.755)	[0.16]
prisoner release for votes	63.205	62.000	63.963	-1.96
	(2.473)	(4.139)	(3.101)	[0.70]
CA climate corruption (R-D)	61.818	62.441	61.426	1.02
	(2.470)	(3.884)	(3.225)	[0.84]
N	88	34	54	

Notes: R=Republicans; D=Democrats. Standard errors in parentheses. P values in brackets for t tests of difference in means.

I assessed the unusualness of these results was due to ambiguities in question wording, inattentive respondents, and the influence of bots. So, after (1) conducting cognitive interviewing to refine the question and headline wording, and (2) adding two "trap" question to identify bots, and (3) adding attention check questions to identify inattentive respondents. This second sample appeared even lower quality than the first

(below). Of the 106 respondents not filtered out by IPHub, only 24 remained after I filtered out probable low quality respondents (duplicates, speeders, likely bots per reCAPTCHA and RelevantID), failed both attention checks, failed trap question). These results encouraged me to switch from Mturk to CloudConnect Research, as discussed in the dissertation.

Sample Quality for Second Mturk Survey

Factor	**Resp. #**	**Notes**
Duplicates (RelevantID, me)	3	• RelevantID: 2 • Me: 1
Speeders (completed ≤ 90 sec)	10	• Clicking at random w/o reading: ~50sec • Reading all and thinking about Qs: ~480sec
Likely bot per **Captcha3**	35	Captcha range: 0-1, with 0.5 indicating equal likelihood of being human or bot.
Likely bot per **RelevantID**	60	Fraud score range: 0-135, with scores greater than 30 indicating equal likelihood of being human or bot
Likely bot per **bot filter Qs**	11	• Bizarre response to "what word is underlined in this question?": 10 • Failed both attention checks: 1

APPENDIX E

IRB APROVAL: PAPER 2

ASU Knowledge Enterprise Development

EXEMPTION GRANTED

Timothy Peterson
CLAS-SS: Politics and Global Studies, School of (SPGS)
-
Timothy.M.Peterson@asu.edu

Dear Timothy Peterson:

On 2/13/2024 the ASU IRB reviewed the following protocol:

Type of Review:	Modification / Update
Title:	Potential Underestimations of Political Misinformation's Impacts: Mistrust, Polarization, Anger, and a Selection Bias
Investigator:	Timothy Peterson
IRB ID:	STUDY00017922
Funding:	Name: Global Studies, School for; CLAS; Name: Educational Outreach and Student Services (EOSS)
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• consent forms 02-13-2024, Category: Consent Form; • recruitment_methods_advertisement_02-12-2024, Category: Recruitment Materials; • social behavioral protocol 02_13_2024, Category: IRB Protocol; • supporting documents 02_12_2024, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions);

The IRB determined that the protocol is considered exempt pursuant to Federal Regulations 45CFR46 (2)(ii) Tests, surveys, interviews, or observation (low risk) on 2/13/2024.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

APPENDIX F

SURVEY: TREATMENT VALIDATION

Note: this is the final version of the treatment validation survey. I can provide copies of the two initial treatment validation survey attempts on Mturk.

Start of Block: 1. Consent, VPN Warning

CONSENT FORM:

Dear respondent,

I am a graduate student under the direction of Professor Timothy Peterson the School of Politics and Global Studies at Arizona State University. I am conducting a research study to examine media dynamics.

I am inviting your participation, which will involve an approximately 8 minute survey. You will be asked to evaluate the plausibility and partisan appeal of 16 headlines, and 7-8 other questions regarding common knowledge, your demographics, and data privacy.

You have the right not to answer any question, and to stop participation at any time. Your participation in this study is voluntary. You must be 18 or older and reside in the United States to participate in the study.

Grounds for disqualification from the study are as follows: (1) use of a Virtual Private Network/Server to mask one's location; (2) having a non-US IP address; (3) failure to correctly answer two questions that are exceptionally easy for a human but exceptionally difficult for a bot; (4) failure to correctly answer a question that all human respondents will be able to answer correctly; and/or (5) exceptionally poor performance on Qualtrics' fraud identification metrics [e.g., reCAPTCHA, RelevantID, duplicate entries, etc.].

You will receive $1.42 for completing the study, and a possible indirect benefit of your participation to society is better understanding of societal cohesion in the United States. There are no foreseeable risks or discomforts to your participation.

We will not ask your name or any other identifying information in this survey. For research purposes, an anonymous numeric code will be assigned to your responses. However, your CloudResearch ID will be temporarily stored with your responses in order to pay you for your time; this data will be deleted as soon as it is reasonably possible.

To ensure we only survey US residents, the survey will automatically check the IP address of whatever computer you are using through a service called IPHub. This service does not have and will not give us any identifying information associated with your IP address; it only allows us to filter out non-US IPs and location-masking software like

Virtual Private Networks. We will delete all IP addresses upon completion of the study.

The results of this study may be used in reports, presentations, or publications but your name will not be used. De-identified data collected as a part of current study may be shared with others (e.g., investigators or industry partners) for future research purposes or other uses.

If you have any questions concerning the research study, please contact the research team at: mjcantre@asu.edu or tmpete15@asu.edu. If you have any questions about your rights as a subject/participant in this research, or if you feel you have been placed at risk, you can contact the Chair of the Human Subjects Institutional Review Board, through the ASU Office of Research Integrity and Assurance, at (480) 965-6788.

By checking the "I agree" box, you are electronically signing this consent form to participate in this study. You affirm you are 18 years or older and live in the United States. To agree: Check the "I agree" box and click NEXT to participate in the study. If you do not wish to participate in this study, simply close out of this browser window.

☐ I agree

WARNING!

This survey uses a protocol to check that you are responding from inside the U.S. and not using a Virtual Private Server (VPS), Virtual Private Network (VPN), or proxy to hide your country.

In order to take this survey, please turn off your VPS/VPN/proxy if you are using one, and also any ad blocking application. Then, refresh the page. Failure to do this will prevent you from completing the study.

JS

Approximately how long will it take for you to complete this survey? [this question is invisible to human respondents]

- ○ 5 minutes or fewer
- ○ 6-15 minutes
- ○ 16-30 minutes
- ○ 31-60 minutes
- ○ Greater than 60 minutes

End of Block: 1. Consent, VPN Warning

Start of Block: 2b. VPN Use Warning

Our system has detected that you are using a Virtual Private Server (VPS), Virtual Private Network (VPN), or proxy to mask your country location.

Because of this, we cannot let you participate in this study. If you are located in the U.S., please turn off your VPN/VPS the next time you participate in a survey, as we requested in the warning message at the beginning. If you are outside the U.S., we apologize, but this study is directed towards U.S. participants only.

Thank you for your interest in our study.

If you have received this message in error, please contact the point of contact for this survey and enter your CloudResearch ID AND the answer to "eighteen minus ten" into the box below.

_____

End of Block: 2b. VPN Use Warning

Start of Block: 2c. Outside of US Warning

Our system has detected that you are attempting to take this survey from a location outside of the U.S. Unfortunately, this study is directed only towards participants in the U.S. and we cannot accept responses from those in other countries (as per our IRB protocol).

Thank you for your interest in our study.

If you have received this message in error, please contact the point of contact for this

survey and enter your CloudResearch ID AND the answer to "eighteen minus ten" into the box below.

End of Block: 2c. Outside of US Warning

Start of Block: 2e. Unresolved Location IPs

For some reason we were still unable to verify your country location. We ask you to please assist us in getting this protocol correct. Please enter your CloudResearch ID below and contact the point of contact for this survey to report the problem

Once you click "Next", you will be taken to the survey (and are certifying that you are taking the survey from the U.S. and not using a VPS). We will check locations manually for those who reach this point and we will contact you if this check identifies you as violating those requirements.

End of Block: 2e. Unresolved Location IPs

Start of Block: 2d. Inauthentic Response Warning

You have been identified this as a probable inauthentic response due to duplicate entries, poor reCAPTCHA performance, other factors identified by RelevantID, and/or displaying inhuman abilities.

Because of this, we cannot let you participate in this study.

If you have received this message in error, please contact the point of contact for this survey and enter your CloudResearch ID AND the answer to "eighteen minus ten" into the box below.

End of Block: 2d. Inauthentic Response Warning

Start of Block: 3. Instructions (Truthfulness)

INSTRUCTIONS PART 1

This survey will ask you to evaluate news headlines.

The first section asks: what is each headline's likelihood of being true?

(Note: we do NOT ask whether a headline likely contains a kernel of truth, but whether it IS true.)

End of Block: 3. Instructions (Truthfulness)

Start of Block: 4. Truthfulness

What is this headline's likelihood of being true? (Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)

White politician names own child after racial slur to protest NFL team name change

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

NRA claims liberal activists conduct school shootings to undermine gun rights

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Alongside abortion bans, most Evangelicals want to force women to marry their rapists

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Fmr. Repub. candidate DeSantis signaled support for law to execute women that receive abortions

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Senior Alabama senator: arm the homeless to protect against active shooters, provide dignity of work

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Trump directed Mar-e-Largo aides to burn classified documents to prevent FBI discovery

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Fox News journalist says swastika-carrying Nazis are 'misunderstood'

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Texas’ proposed history textbooks say slaves received sufficient compensation for their labor

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

MSNBC's secret internal policy to fire commentators critical of welfare fraud

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

Brief instruction check: In this section, are we asking you whether a headline contains a kernel of truth, or whether it is true?

- ○ Contains a kernel of truth
- ○ Is true

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Large minority of migrant 'families' crossing border are actually child sex traffickers and their victims

- Likely
- Somewhat likely
- Neither likely nor unlikely
- Somewhat unlikely
- Unlikely

Page Break

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

New York Times covered-up illegal FBI surveillance of conservative candidates

- Likely
- Somewhat likely
- Neither likely nor unlikely
- Somewhat unlikely
- Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Russia decided to invade Ukraine after learning 'weak' Biden would succeed Trump as president

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Education Secretary: fire teachers that refuse to encourage all students to question their gender

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Dem. governors manufacture prison 'overpopulation' to justify early release of Dem. voters

- Likely
- Somewhat likely
- Neither likely nor unlikely
- Somewhat unlikely
- Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Saving the planet or lining their pockets? California politicians profit from climate change regulations

- Likely
- Somewhat likely
- Neither likely nor unlikely
- Somewhat unlikely
- Unlikely

What is this headline's likelihood of being true? *(Note: we do not ask whether a headline likely contains a kernel of truth, but whether it is true.)*

Revealed: Black Lives Matter plans race riots as cover for bloody, Marxist redistribution of wealth

- ○ Likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Unlikely

End of Block: 4. Truthfulness

Start of Block: 5. Sample Quality Qs

That finishes the first half of the survey.

Below are four brief questions before moving on to the second half of the survey.

Important! The questions are very easy, and the answers will seem obvious, because they are.

What are these things? List them by name in the box below

_____

Waht wrd is undrelned in tihs qestuion?

What is your profession?

- Forest ranger
- Commercial fisherman
- Lumberjack
- Tailor
- None of the above / not applicable

How many mornings per week do you bicycle to the moon on a hot air balloon?

- 0 mornings per week
- 1-2 mornings per week
- 3-4 mornings per week
- 5-6 mornings per week
- 7 mornings per week

End of Block: 5. Sample Quality Qs

Start of Block: 6. Instructions (Favor)

INSTRUCTIONS PART 2

On the next page are the same headlines as before, but with a different question:

Assuming the headlines are entirely accurate, how favorable do you find them to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

Important! We're NOT asking: How favorable people in general might find them
How favorable a ${e://Field/party1_singular} or ${e://Field/party2_singular}

might find them

But how favorable do YOU find the headlines, assuming they are entirely accurate?

(note: a headline is more favorable to one party if it makes that party look good and/or their opponent party look bad)

End of Block: 6. Instructions (Favor)

Start of Block: 7. Partisan Favor

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

White politician names own child after racial slur to protest NFL team name change

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

NRA claims liberal activists conduct school shootings to undermine gun rights

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Poll: alongside abortion bans, most Evangelicals want to force women to marry their rapists

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Fmr. Repub. candidate DeSantis signals support for law to execute women that receive abortions

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Senior Alabama senator: arm the homeless to protect against active shooters, provide dignity of work

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Trump directed Mar-e-Largo aides to burn classified documents to prevent FBI discovery

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Fox News journalist says swastika-carrying Nazis are 'misunderstood'

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Texas’ proposed history textbooks say slaves received sufficient compensation for their labor

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

MSNBC's secret internal policy to fire commentators critical of welfare fraud

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Almost there!

To mentally refresh the instructions for this section, below is a recall task.

Which of these three things did the instructions ask you to do?

- ○ Assess how favorable I find the headlines to one party vs. another
- ○ Treat headlines as more favorable to one party if they make that party look good and/or its opponent party look bad
- ○ Assume the headlines are entirely accurate
- ○ The instructions asked me to do all three of these things

---

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Large minority of migrant 'families' crossing border are actually child sex traffickers and their victims

- ○ More favorable to ${e://Field/party1_abrv}
- ○ Somewhat more favorable to ${e://Field/party1_abrv}
- ○ Equally favors both parties
- ○ Somewhat more favorable to ${e://Field/party2_abrv}
- ○ More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

New York Times covered-up illegal FBI surveillance of conservative candidates

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Russia decided to invade Ukraine after learning 'weak' Biden would succeed Trump as president

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Education Secretary: fire teachers that refuse to encourage all students to question their gender

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Dem. governors manufacture prison 'overpopulation' to justify early release of Dem. voters

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Saving the planet or lining their pockets? California politicians profit from climate change regulations

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

Assuming this headline is entirely accurate, how favorable do you find it to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

(hover here for definition reminder)

Revealed: Black Lives Matter plans race riots as cover for bloody, Marxist redistribution of wealth

- More favorable to ${e://Field/party1_abrv}
- Somewhat more favorable to ${e://Field/party1_abrv}
- Equally favors both parties
- Somewhat more favorable to ${e://Field/party2_abrv}
- More favorable to ${e://Field/party2_abrv}

End of Block: 7. Partisan Favor

Start of Block: 8a. Partisanship Information

You're nearly done. Three final questions.

Would you be willing to grant researchers access to your anonymized web browser history and/or social media activity for a limited period of time?

- ○ Yes
- ○ No

Display This Question:

If orderRepubHigh = 0

Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent, or what?

- ○ Democrat
- ○ Republican
- ○ Independent
- ○ Other

Display This Question:

If orderRepubHigh = 1

Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what?

- ○ Republican
- ○ Democrat
- ○ Independent
- ○ Other

End of Block: 8a. Partisanship Information

Start of Block: 8b. Partisanship Information

Display This Question:

If Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent,... = Republican

Or Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Republican

Final question: would you call yourself a strong Republican, or a not very strong Republican?

- Strong Republican
- Not very strong Republican

Display This Question:

If Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent,... = Democrat

Or Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Democrat

Final question: would you call yourself a strong Democrat, or a not very strong Democrat?

- Strong Democrat
- Not very strong Democrat

Display This Question:

If Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent,... = Independent

Or Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent,... = Other

Or Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Independent

Or Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Other

Final question: do you think of yourself as closer to the Republican Party or to the Democratic Party?

- ○ Closer to Republican Party
- ○ Closer to Democratic Party
- ○ Neither

That concludes the study. Thank you for participating!

!!!--- DEBRIEFING STATEMENT ---!!!

ALL the headlines in this study were made up by the researcher. NONE WERE REAL.

Also, NO MATTER your answer to the question about willingness to grant access to internet/social media history, you did NOT grant us access to that data. We were solely interested in your willingness to grant access.

---> Please click the "next" arrow to be redirected to CloudResearch Thank you again!

End of Block: 8b. Partisanship Information

APPENDIX G

SURVEY: MISINFORMATION TREATMENT

Start of Block: 1. Consent, VPN Warning

CONSENT FORM:

Dear respondent,

I am a graduate student under the direction of Professor Timothy Peterson in the School of Politics and Global Studies at Arizona State University. I am conducting a research study to examine social media dynamics.

I am inviting your participation, which will involve an approximately 8 minute survey. You will be asked questions regarding demographics, social media habits/assessments, emotions, society, and data privacy.

You have the right not to answer any question, and to stop participation at any time. Your participation in this study is voluntary. You must be 18 or older and reside in the United States to participate in the study.

Grounds for disqualification from the study are as follows: (1) use of a Virtual Private Network/Server to mask one's location; (2) having a non-US IP address; (3) failure to correctly answer two questions that are exceptionally easy for a human but exceptionally difficult for a bot; (4) failure to correctly answer a question that all human respondents will be able to answer correctly; and/or (5) exceptionally poor performance on Qualtrics' fraud identification metrics [e.g., reCAPTCHA, RelevantID, duplicate entries, etc.].

You will receive $1.42 for completing the study, and a possible indirect benefit of your participation to society is better understanding of societal cohesion in the United States. There are no foreseeable risks or discomforts to your participation.

We will not ask your name or any other identifying information in this survey. For research purposes, an anonymous numeric code will be assigned to your responses. However, your CloudResearch ID will be temporarily stored in order to pay you for your time; this data will be deleted as soon as it is reasonably possible.

To ensure we only survey US residents, the survey will automatically check the IP address of whatever computer you are using through a service called IPHub. This service does not have and will not give us any identifying information associated with your IP address; it only allows us to filter out non-US IPs and location-masking software like Virtual Private Networks. We will delete all IP addresses upon completion of the study.

If you have any questions concerning the research study, please contact the research team at: mjcantre@asu.edu or tmpete15@asu.edu. If you have any questions about your rights as a subject/participant in this research, or if you feel you have been placed at risk, you can contact the Chair of the Human Subjects Institutional Review Board, through the

ASU Office of Research Integrity and Assurance, at (480) 965-6788.

By checking the “I agree” box, you are electronically signing this consent form to participate in this study. You affirm you are 18 years or older and live in the United States. To agree: Check the “I agree” box and click NEXT to participate in the study. If you do not wish to participate in this study, simply close out of this browser window.

☐ I agree

WARNING!

This survey uses a protocol to check that you are responding from inside the U.S. and not using a Virtual Private Server (VPS), Virtual Private Network (VPN), or proxy to hide your country.

In order to take this survey, please turn off your VPS/VPN/proxy if you are using one, and also any ad blocking application. Then, refresh the page. Failure to do this will prevent you from completing the study.

JS

\Approximately how long will it take for you to complete this survey? [this question is invisible to human respondents]

- ○ 5 minutes or fewer
- ○ 6-15 minutes
- ○ 16-30 minutes
- ○ 31-60 minutes
- ○ Greater than 60 minutes

End of Block: 1. Consent, VPN Warning

Start of Block: 2b. VPN Use Warning

Our system has detected that you are using a Virtual Private Server (VPS), Virtual Private Network (VPN), or proxy to mask your country location.

Because of this, we cannot let you participate in this study. If you are located in the U.S.,

please turn off your VPN/VPS the next time you participate in a survey, as we requested in the warning message at the beginning. If you are outside the U.S., we apologize, but this study is directed towards U.S. participants only.

Thank you for your interest in our study.

If you have received this message in error, please contact the point of contact for this survey and enter your CloudResearch ID AND the answer to "eighteen minus ten" into the box below.

End of Block: 2b. VPN Use Warning

Start of Block: 2c. Outside of US Warning

Our system has detected that you are attempting to take this survey from a location outside of the U.S. Unfortunately, this study is directed only towards participants in the U.S. and we cannot accept responses from those in other countries (as per our IRB protocol).

Thank you for your interest in our study.

If you have received this message in error, please contact the point of contact for this survey and enter your CloudResearch ID AND the answer to "eighteen minus ten" into the box below.

End of Block: 2c. Outside of US Warning

Start of Block: 2e. Unresolved Location IPs

For some reason we were still unable to verify your country location. We ask you to please assist us in getting this protocol correct. Please enter your CloudResearch ID below and contact the point of contact for this survey to report the problem

Once you click "Next", you will be taken to the survey (and are certifying that you are taking the survey from the U.S. and not using a VPS). We will check locations manually for those who reach this point and we will contact you if this check identifies you as violating those requirements.

End of Block: 2e. Unresolved Location IPs

Start of Block: 2d. Inauthentic Response Warning

You have been identified this as a probable inauthentic response due to duplicate entries, poor reCAPTCHA performance, other factors identified by RelevantID, and/or displaying inhuman abilities.

Because of this, we cannot let you participate in this study.

If you have received this message in error, please contact the point of contact for this survey and enter your CloudResearch ID AND the answer to "eighteen minus ten" into the box below.

_____

End of Block: 2d. Inauthentic Response Warning

Start of Block: 3. Pre-Survey

JS

Welcome to the study!

First, some common knowledge questions.

Important! The questions are very easy and the answers will seem obvious, because they are

What are these things? List them by name in the box below

_____

Waht wrd is undrelned in tihs qestuion?

_____

Next: brief demographics questions.

We have most of your demographics from CloudResearch, but we need to ask you about your political and religious involvement.

Do you go to religious services every week, almost every week, once or twice a month, a few times a year, or never?

- Every week
- Almost every week
- Once or twice a month
- A few times a year
- Never

Some people don’t pay much attention to political campaigns. How about you? Would you say that you have been very much interested, somewhat interested or not much interested in the political campaigns so far this year?

- Very much interested
- Somewhat interested
- Not much interested

How often do you pay attention to what’s going on in government and politics?

- Always
- Most of the time
- About half of the time
- Some of the time
- Never

Display This Question:

If orderRepubHigh = 1

Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent, or what?

- ○ Democrat
- ○ Republican
- ○ Independent
- ○ Other

Display This Question:

If orderRepubHigh = 0

Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what?

- ○ Republican
- ○ Democrat
- ○ Independent
- ○ Other

Display This Question:

If Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent,... = Independent

Or Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent,... = Other

Or Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Independent

Or Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Other

Do you think of yourself as closer to the Republican Party or to the Democratic Party?

- ○ Closer to Republican Party
- ○ Neither
- ○ Closer to Democratic Party

Display This Question:

If Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent,... = Democrat

Or Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Democrat

Would you call yourself a strong Democrat or a not very strong Democrat?

- ○ Strong Democrat
- ○ Not very strong Democrat

Display This Question:

If Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent,... = Republican

Or Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Republican

Would you call yourself a strong Republican or a not very strong Republican?

- ○ Strong Republican
- ○ Not very strong Republican

End of Block: 3. Pre-Survey

Start of Block: 4. PolKnow Qs

JS

Thank you. Next is a short section to wrap up the introductory questions.

Instructions: we are interested in the guesses people make when they do not know the answer to a question.

We will ask you four questions. Some may be easy, but others are meant to be so difficult that you will have to guess.

- ○ I promise to try my best without looking up any answers
- ○ I do not want to make that promise

Here is the first one. It is an example of a difficult question:

In what year did the Supreme Court of the United States decide Geer v. Connecticut?

Type the year.

_____

End of Block: 4. PolKnow Qs

Start of Block: 5. PolKnow Qs

Display This Question:

If Here is the first one. It is an example of a difficult question:In what year did the Supreme Court of the United States decide Geer v. Connecticut?Type the year. Text Response Is Equal to 1896

You are right!

Did you look up the answer to that question, or did you already know it yourself?

- ○ I looked it up
- ○ I already knew it

Do you happen to know which party currently has the most members in the U.S. House of Representatives in Washington, D.C.?

- ○ Republicans
- ○ Democrats

Do you happen to know which party is more conservative?

- ○ Republicans
- ○ Democrats

How much of a majority is required for the U.S. Senate and House of Representatives to override a presidential veto?

- ○ 1/2
- ○ 3/5
- ○ 2/3
- ○ 3/4
- ○ The U.S. Senate and House cannot override a presidential veto

End of Block: 5. PolKnow Qs

Start of Block: 5. Social Media Habits

Thank you. Now on to basic social media habits and opinions:

How often do you use social media on average?

- ○ 5+ times a day
- ○ 2-4 times a day
- ○ About once a day
- ○ A few times each week
- ○ Once or twice a month
- ○ Less than once a month
- ○ Never

Have you ever unfollowed, unfriended, or muted a person/page on social media due to the content they were posting?

- ○ Yes
- ○ No

Would you be willing to grant researchers access to your anonymized web browser history and/or social media activity for a limited period of time?

- ○ Yes
- ○ No

How many mornings per week do you bicycle to the moon on a hot air balloon?

- 0 mornings per week
- 1-2 mornings per week
- 3-4 mornings per week
- 5-6 mornings per week
- 7 mornings per week

End of Block: 5. Social Media Habits

Start of Block: 6. Treatment Exposure

Below are four articles shared by randomly-selected social media users${e://Field/period1} ${e://Field/whoAre} ${e://Field/repubDem2}

They endorsed the articles as being interesting, important, or useful to know.

${e://Field/finalListItem1}

${e://Field/finalListItem2}

${e://Field/finalListItem3}

${e://Field/finalListItem4}

Imagine you are scrolling through your social media feed and you come across any of the articles the users shared.

Which of the articles would you read, if any?

Please select as many as apply.

- ☐ ${e://Field/finalListItem1}
- ☐ ${e://Field/finalListItem2}
- ☐ ${e://Field/finalListItem3}
- ☐ ${e://Field/finalListItem4}
- ☐ ⊗I would not read any of these

End of Block: 6. Treatment Exposure

Start of Block: 7. Emotions, Trust I

We're now in the second half of the survey. It will first ask you about emotions.

Generally speaking, how do you feel about the way things are going in the country these days?

Please tell us how much you feel each of the following emotions:

	Not at all	A little	Somewhat	Very	Extremely
Hopeful	○	○	○	○	○
Angry	○	○	○	○	○
Proud	○	○	○	○	○
Afraid	○	○	○	○	○
Disgusted	○	○	○	○	○
Happy	○	○	○	○	○
Nervous	○	○	○	○	○
Outraged	○	○	○	○	○

We'd like to get your feelings toward the two main political parties in the news these days. We'd like you to rate the parties using something we call the feeling thermometer. The thermometer is from 0 to 100 degrees, in which:

- 100 degrees: means you feel favorable and warm toward the party.
- 0 degrees: means you don't feel favorable toward the party and you don't care too much for that party.
- 50 degrees: means you don't feel particularly warm or cold toward the party.

Unfavorable Favorable

0 10 20 30 40 50 60 70 80 90 100

Republicans	
Democrats	

Page Break

Now the survey will ask you about trust.

"You can’t count on strangers anymore."

Do you:

- More or less agree
- More or less disagree

Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?

- Most people can be trusted
- Can't be too careful

Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?

- Try to be helpful
- Just look out for themselves

Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair?

- ○ Would take advantage of you
- ○ Would try to be fair

End of Block: 7. Emotions, Trust I

Start of Block: 8. Trust II

Sidebar question: what is your profession?

- ○ Forest ranger
- ○ Commercial fisherman
- ○ Lumberjack
- ○ Tailor
- ○ None of the above / not applicable

We are going to name some groups/institutions in this country.

As far as the people in these groups/institutions are concerned, how much trust would you say you have in them?

	None	A little	A moderate amount	A lot	A great deal
The White House	○	○	○	○	○
News Media	○	○	○	○	○
U.S. Supreme Court	○	○	○	○	○
Congress	○	○	○	○	○
The Military	○	○	○	○	○
The Justice System	○	○	○	○	○
Democrats	○	○	○	○	○
Republicans	○	○	○	○	○
Fact Checkers for Social Media Companies	○	○	○	○	○

When it comes to public policy decisions, whom do you tend to trust more: ordinary people, experts, or trust both the same?

- ○ Trust ordinary people more
- ○ Trust experts more
- ○ Trust both the same

End of Block: 8. Trust II

Start of Block: 9. Trust III

Display This Question:

If When it comes to public policy decisions, whom do you tend to trust more: ordinary people, expert... = Trust ordinary people more

Do you trust ordinary people much more or somewhat more than experts when it comes to public policy decisions?

- ○ Much more
- ○ Somewhat more

Display This Question:

If When it comes to public policy decisions, whom do you tend to trust more: ordinary people, expert... = Trust experts more

Do you trust experts much more or somewhat more than ordinary people when it comes to public policy decisions?

- ○ Much more
- ○ Somewhat more

End of Block: 9. Trust III

Start of Block: 10. Belief

Below are some article headlines. We showed you some of the headlines earlier in the study, but some of the headlines we have not showed you before. What is each headline's likelihood of being true?

(Note: we do NOT ask whether the headline likely contains a kernel of truth, but whether it IS true.)

	Likely	Somewhat likely	Equally likely and unlikely	Somewhat unlikely	Unlikely
Trump directed Mar-e-Largo aides to burn classified documents to prevent FBI discovery	○	○	○	○	○
Texas' proposed history textbooks say slaves received sufficient compensation for their labor	○	○	○	○	○
New York Times covered-up illegal FBI surveillance of conservative candidates	○	○	○	○	○
Large minority of migrant 'families' crossing border are actually child sex traffickers and their victims	○	○	○	○	○

Facial recognition firm Clearview AI used nearly 1m times by US police	○	○	○	○	○
Afghanistan girls' education activist arrested by Taliban	○	○	○	○	○
Flat-packed pasta could help revolutionize food production	○	○	○	○	○
Brain cancer patient prepares to run London Marathon	○	○	○	○	○

End of Block: 10. Belief

Start of Block: 11. Manipulation Check, Placebo

Final four questions.

Earlier in the study, we mentioned that randomly-selected social media users had shared the original four article headlines we showed you.

Did the survey say which political party the social media users belonged to?

- ○ Yes, they were Republicans
- ○ Yes, they were Democrats
- ○ Yes, they were Independent/Other
- ○ No, the survey did not say anything about their political party

---

Display This Question:

If ctrlGrp = 0

Thank you.

Regardless of what the survey did/did not say, what major political party would you

guess each of the article sharers most likely belongs to?

	Likely to be ${e://Field/party1_abrv}	Somewhat likely to be ${e://Field/party1_abrv}	Equally likely to be either party	Somewhat likely to be ${e://Field/party2_abrv}	Likely to be ${e://Field/party2_abrv}
Sharer of: ${e://Field/finalListItem1}	○	○	○	○	○
Sharer of: ${e://Field/finalListItem2}	○	○	○	○	○
Sharer of: ${e://Field/finalListItem3}	○	○	○	○	○
Sharer of: ${e://Field/finalListItem4}	○	○	○	○	○

Carry Forward Displayed Choices from "Below are some article headlines. We showed you some of the headlines earlier in the study, but some of the headlines we have not showed you before.What is each headline's likelihood of being true?(Note: we do NOT ask whether the headline likely contains a kernel of truth, but whether it IS true.) "

A great many articles are shared on social media and the internet. Do you remember having seen any of the article headlines before you participated in this study?

	Remember having seen	Might have seen, but don't clearly remember	Don't remember having seen
Trump directed Mar-e-Largo aides to burn classified documents to prevent FBI discovery	○	○	○
Texas’ proposed history textbooks say slaves received sufficient compensation for their labor	○	○	○
New York Times covered-up illegal FBI surveillance of conservative candidates	○	○	○
Large minority of migrant 'families' crossing border are actually child sex traffickers and their victims	○	○	○
Facial recognition firm Clearview AI used nearly 1m times by US police	○	○	○
Afghanistan girls' education activist arrested by Taliban	○	○	○
Flat-packed pasta could help revolutionize food production	○	○	○
Brain cancer patient prepares to run London Marathon	○	○	○

Display This Question:

If ctrlGrp = 1

Thank you.

Below, we list the original four article headlines you saw in the study.

Assuming these headlines are entirely accurate, how favorable do you find them to ${e://Field/party1_plural} versus ${e://Field/party2_plural}?

Important! We're NOT asking: How favorable people in general might find them
How favorable a ${e://Field/party1_singular} or ${e://Field/party2_singular} might find them
But how favorable do YOU find the headlines, assuming they are entirely accurate?

(*note: a headline is more favorable to one party if it makes that party look good and/or their opponent party look bad*)

	More favorable to ${e://Field/party1_abrv}	Somewhat more favorable to ${e://Field/party1_abrv}	Equally favors both parties	Somewhat more favorable to ${e://Field/party2_abrv}	More favorable to ${e://Field/party2_abrv}
${e://Field/finalListItem1}	○	○	○	○	○
${e://Field/finalListItem2}	○	○	○	○	○
${e://Field/finalListItem3}	○	○	○	○	○
${e://Field/finalListItem4}	○	○	○	○	○

Page Break

Final Question!

Inaccurate information (misinformation) is sometimes discussed these days. Some people say it is very influential, and some people say it's not very influential.

How much influence would you say inaccurate information has on the political opinions of following people/groups?

	None	A little	A moderate amount	A lot	A great deal
Yourself	○	○	○	○	○
People in general	○	○	○	○	○
Republicans	○	○	○	○	○
Democrats	○	○	○	○	○

End of Block: 11. Manipulation Check, Placebo

Start of Block: End of Survey

That concludes the study. Thank you for participating!

!!!--- DEBRIEFING STATEMENT ---!!!

NO MATTER your answer to the question about willingness to grant access to internet/social media history, you did **NOT** grant us access to that data. We were solely interested in your willingness to grant access.
The random social media users who were the alleged sources of the original four articles did **NOT** actually exist. The researcher was the source of the articles.
The following headlines were **made up by the researcher**. They were **NOT** true.

Trump directed Mar-e-Largo aides to burn classified documents to prevent FBI discovery

Texas' proposed history textbooks say slaves received sufficient compensation for

their labor

New York Times covered-up illegal FBI surveillance of conservative candidates

Large minority of migrant 'families' crossing border are actually child sex traffickers and their victims

The focus of this study was social media dynamics: specifically the effects of misinformation exposure on social media. We are telling you that now rather than at the beginning of the study because knowing the study involved misinformation likely would have influenced respondents' answers.

---> Please click the "next" arrow to be redirected to CloudResearch Thank you again!

End of Block: End of Survey

APPENDIX H

ROBUSTNESS CHECKS: BELIEF

Brant Test

Importantly, the key explanatory variable of interest—congeniality as proxied by partisanship—does not violate the parallel regressions assumption in a way that undermines the validity of the regression results. Brant test results suggest it only violates the assumption once (column 2, $p < 0.01$), and does not do so in a way that changes the direction of the results. Though they vary in magnitude in a statistically significant way, all four constituent coefficients are negative.

Brant Test Results by Regression (Not Odds Ratios)

Coefficients	(1) Burn Docs	(2) TX Textbooks	(3) NYT Coverup	(4) Child Trafficking
Republican	p = 0.718	**p = 0.007**	p = 0.814	p = 0.987
Coefficient 1	-2.012	-0.710	0.707	1.324
2	-1.756	-1.205	0.748	1.274
3	-1.862	-1.572	0.685	1.263
4	-1.997	-1.797	0.940	1.358
Two Exposures	p = 0.846	p = 0.400	p = 0.708	p = 0.928
Coefficient 1	-0.045	0.119	-0.626	-0.409
2	-0.033	-0.055	-0.463	-0.329
3	-0.196	0.058	-0.354	-0.263
4	-0.326	-0.337	-0.107	-0.042
Repub*Two Exp.	p = 0.798	p = 0.972	p = 0.482	p = 0.760
Coefficient 1	0.157	-0.189	1.266	0.017
2	0.050	-0.273	0.605	0.225
3	0.233	-0.190	0.639	0.465
4	-0.236	-0.009	0.343	0.195
Illusory Truth	**p = 0.012**	p = 0.770	p = 0.907	**p = 0.041**
Coefficient 1	0.547	0.889	0.621	0.456
2	0.744	0.773	0.713	0.745
3	0.996	0.682	0.680	0.866
4	1.104	0.760	0.593	0.696
Trust: Media	**p = 0.000**	p = 0.214	p = **0.001**	p = 0.872
Coefficient 1	0.687	0.200	-0.282	-0.281
2	0.448	0.101	-0.445	-0.325
3	0.298	0.003	-0.590	-0.373
4	-0.077	-0.120	-1.006	-0.403

Low Attention	p = 0.696	p = 0.731	p = **0.017**	**p = 0.013**
Coefficient 1	0.184	0.149	0.392	0.614
2	-0.068	0.147	0.819	0.306
3	-0.204	0.218	0.160	-0.133
4	-0.188	-0.159	-0.208	-0.913
Observations	811	811	811	811

P values < 0.1 are bolded

Alternate Specification

Save one minor difference, the original analysis' findings are robust to a more simple model specification that omits perceived illusory truth, trust in media, and low attention. Partisanship still shows significant effects conditioned on headline congeniality; and "Police AI Use" remains the one headline without significant effects. Further, Republicans remain marginally more skeptical of the three other real headlines (columns 6-8). Finally, illusory truth decreased belief in the "Patient Runs Marathon" and "NYT Coverup" headline, but also, novelly, the "Child Trafficking" headline as well. Illusory truth maintained its differential effects by partisanship for the former two headlines as well. The new significance of "Child Trafficking" does not have any major theoretical implications.

Estimated Impact of Exposure on Belief

VARIABLES	(1) Burn Docs	(2) TX Textbooks	(3) NYT Coverup	(4) Child Trafficking	(5) Police AI Use	(6) Afg. Activist Arrest	(7) Flat Packed Pasta	(8) Patient Runs Marathon
Republican	0.113***	0.232***	3.240***	4.983***	0.891	0.714**	0.606***	0.759*
	(0.020)	(0.038)	(0.502)	(0.792)	(0.115)	(0.095)	(0.094)	(0.120)
Saw Twice	0.829	0.886	0.700*	0.785	--	--	1.353	1.614**
	(0.153)	(0.165)	(0.135)	(0.155)			(0.250)	(0.311)
Repub.*Twice	1.503	1.024	1.993**	1.343	--	--	1.186	0.605*
	(0.397)	(0.268)	(0.546)	(0.371)			(0.318)	(0.165)
Observations	811	811	811	811	811	811	811	811

seEform in parentheses

*** p<0.01, ** p<0.05, * p<0.1

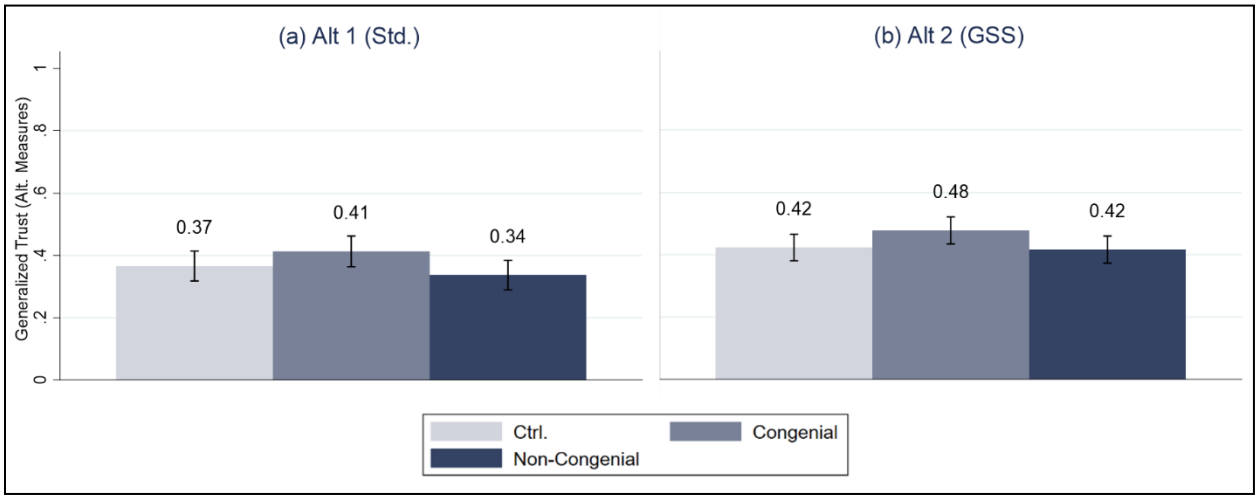
APPENDIX I

ROBUSTNESS CHECKS: TRUST

Alternate GT Operationalizations

As with the primary GT measure, descriptive analyses employing the alternate two GT measures do not show statistically-significant differences-in-means between the treatment and control groups (below). For the first alternate measure, the standard dichotomous question, the means range from 0.34-0.41, anchored on the control group's mean of 0.37. For the second alternate measure, the GSS index), the means again are tightly ranged (0.42-0.48).

Means: Alternate GT Measures



Ordinal logistic regression analyses that take into account (dis)belief do not majorly differ from the primary GT analysis, either (below). Using the dichotomous alternate measure, the results again show the conditional effect of misinformation exposure on GT. Republicans again have lower odds of higher GT (column 3, -67,4%, $p < 0.05$): and impact again counterbalanced by an overall increase in odds of higher GT of 110.6% ($p < 0.05$). The alternate results also again show two other parameters in the analysis approach statistical significance in column 3. NCM exposure among Democrats

increases odds of higher GT by 118.0% and disbelief in said NCM decreases odds of higher GT by 59.7%. These parameters only have respective p-values of 0.216 and 0.202 respectively. The former p-value is roughly the same as in the original analysis, and the latter is slightly less significant, but these differences are minor.

Estimated Impact of Exposure on GT (Alt. Measure 1)

	CM		NCM	
	(1)	(2)	(3)	(4)
VARIABLES	Dem.	Repub.	Dem.	Repub.
CM	1.184	2.106**		
	(0.379)	(0.723)		
Believed both headlines	1.009	1.112		
	(0.363)	(0.423)		
CM*Believed both headlines	0.831	0.326**		
	(0.416)	(0.179)		
Low attent.	1.105	0.974	1.221	3.349**
	(0.436)	(0.391)	(0.577)	(1.742)
Religious Attendance	0.872#	1.082	0.957	1.042
	(0.093)	(0.089)	(0.105)	(0.091)
NCM			2.280	0.616
			(1.518)	(0.399)
Disbelieved 1+ headline			3.096**	1.135
			(1.656)	(0.572)
NCM* Disbelieved 1+ headline			0.403	1.084
			(0.287)	(0.768)
Constant	0.815	0.393***	0.270**	0.395*
	(0.240)	(0.130)	(0.145)	(0.203)
Observations	283	261	264	276

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

A similar conclusion results from ordinal logistic analysis using the second alternate GT measure: the GSS GT index (below). Again, the parameters on CM and the interaction variable in column 2 indicates decreased odds of higher GT at significant or nearly-significant levels. The only difference is that the equivalent parameters in column

3 (NCM, the interaction) are slightly less significant than in either the primary or other alternate analysis (p = 0.261-0.315). The non-significance of NCM disbelief among Democrats would limit the support for Hypothesis 2a, but not remove it entirely.

Estimated Impact of Exposure on GT (Alt. Measure 1)

	CM		NCM	
	(1)	(2)	(3)	(4)
VARIABLES	Dem.	Repub.	Dem.	Repub.
CM	0.066	0.127*		
	(0.068)	(0.071)		
Believed both headlines	0.015	-0.022		
	(0.076)	(0.077)		
CM*Believed both headlines	-0.039	-0.144#		
	(0.106)	(0.110)		
Low attent.	0.005	-0.013	0.028	0.183*
	(0.084)	(0.082)	(0.102)	(0.106)
Religious Attendance	-0.022	0.024#	-0.017	0.017
	(0.022)	(0.017)	(0.023)	(0.017)
NCM			0.144	-0.055
			(0.127)	(0.120)
Disbelieved 1+ headline			0.199**	0.064
			(0.097)	(0.098)
NCM* Disbelieved 1+ headline			-0.140	0.013
			(0.139)	(0.132)
Constant	0.476***	0.345***	0.311***	0.296***
	(0.062)	(0.066)	(0.097)	(0.099)
Observations	283	261	264	276
R-squared	0.008	0.028	0.022	0.020

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Brant Tests for OPT Ordinal Logistic Regressions

Below are the Brant test results for the OPT ordinal logistic regression analysis. Three coefficients in column 4 (NCM, Republicans) vary widely enough in the constituent

logistic regressions that they differ at a statistically significant level. Of them, the parameter that theoretically matters the most is that of NCM treatment*Disbelief. In the original analysis, it was positive but statistically insignificant, but the Brant test shows that one of its estimated parameters is 1.707, more closely approaching significance. Such an impact would be unexpected. NCM disbelief should decrease OPT. But, as the OPT results have been entirely counter-expectation, the positive impact of CM disbelief on OPT among Republicans would only add further support to this section's support for the null hypothesis—that misinformation exposure has no impact on OPT.

Brant Test Results by OPT Regression (Not Odds Ratios)

	CM		NCM	
	(1)	(2)	(3)	(4)
Coefficients	Dem.	Repub.	Dem.	Repub.
Treatment	p = 0.968	p = 0.974	p = 0.172	**p = 0.055**
Coefficient 1	-0.157	0.716	0.223	0.760
2	-0.179	0.731	1.799	-0.658
Believed/Disbel.	p = 0.475	p = 0.470	p = 0.221	**p = 0.065**
Coefficient 1	-0.688	-1.068	-0.713	-0.496
2	-1.445	-1.815	0.610	-1.619
Treatm.*Bel/Disbel.	p = 0.671	p = 0.796	p = 0.107	**p = 0.019**
Coefficient 1	0.401	-0.426	-0.317	-0.342
2	0.935	-0.805	-2.362	1.707
Low Attent.	p = 0.518	p = 0.643	p = 0.870	p = 0.246
Coefficient 1	0.310	-0.346	0.914	0.070
2	0.720	-0.061	1.029	0.724
Relig. Freq.	p = 0.730	p = 0.940	p = 0.843	p = 0.195
Coefficient 1	0.270	0.137	0.237	0.119
2	0.322	0.127	0.202	-0.047
Observations	811	811	811	811

P values < 0.1 are bolded

Regression Results for Alternate Model Specifications

These regressions all omit the three potential confounders I included in the original analyses: partisanship, religious service attendance, and low respondent attention.

Whole-Sample

These results roughly mirror those of the original analysis. CM treatment only had a statistically significant impact on GT, increasing it. NCM treatment, on the other hand, increased only OPT in a statistically significant manner.

Estimated Impact of Exposure on Trust

VARIABLES	(1) GT	(2) OPT
CM	1.414*	1.229
	(0.253)	(0.209)
NCM	1.300	1.396**
	(0.235)	(0.238)
Observations	811	811

seEform in parentheses

*** p<0.01, ** p<0.05, * p<0.1

GT (Disaggregated)

The results from Table 10 are unchanged when one drops *religious attendance* and *low attention* from the specification. CM treatment again has positive and near-significant effects among both parties (respective p = 0.1490 and 0.107. Again, CM belief only decreases GT among Republicans at near-significant levels (p = 0.203). Further, NCM treatment still only has positive, significant impacts among Democrats, and Democrats remain the only group that experienced a decrease in GT if they later reported that they disbelieved the NCM headline they saw.

Estimated Impact of Exposure on GT (Disaggregated)

VARIABLES	(1) Dem.	(2) Repub.	(3) Dem.	(4) Repub.
CM treatment	1.566#	1.690#		
	(0.487)	(0.550)		
Believed both headlines	0.705	0.772		
	(0.267)	(0.301)		
CM treatment*Believed both headlines	0.949	0.493		
	(0.491)	(0.274)		
NCM treatment			4.442**	0.889
			(3.031)	(0.576)
Disbelieved 1+ headline			2.949*	1.316
			(1.708)	(0.683)
NCM treatment*Disbelieved 1+ headline			0.378#	1.010
			(0.278)	(0.715)
Observations	283	261	264	276

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1 # p<0.2

OPT (Disaggregated)

The results from Table 11 are robust to the simpler specification, save for one modification. The key results remain unchanged: neither CM belief nor NCM disbelief impact OPT, regardless of party. CM treatment also only matters among Republicans, increasing their outgroup trust. The one change regards NCM treatment. This specification's results assign near-significance to NCM treatment's effects on OPT among Democrats (column 3). It increased their odds of higher OPT 114.1% (p = 0.164). This does not change the overall lack of support for Hypothesis 2b, though.

Estimated Impact of Exposure on OPT (Disaggregated)

VARIABLES	(1) Dem.	(2) Repub.	(3) Dem.	(4) Repub.
CM treatment	0.888	1.796*		
	(0.266)	(0.547)		
Believed both headlines	0.417**	0.319***		
	(0.156)	(0.125)		
CM treatment*Believed both headlines	1.735	0.685		
	(0.889)	(0.371)		
NCM treatment			2.141#	1.412
			(1.170)	(0.765)
Disbelieved 1+ headline			0.559#	0.489#
			(0.236)	(0.226)
NCM treatment*Disbelieved 1+ headline			0.461	1.136
			(0.284)	(0.684)
Observations	283	261	264	276

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1 # p<0.2
Model: ordinal logistic model

APPENDIX J

ROBUSTNESS CHECKS: AFFECTIVE POLARIZATION

Below are regressions employing an alternate specification to those used in the main analysis. This specification drops potential confounders of partisanship and low respondent attention.

Whole-Sample

As in Table 12, neither the CM nor NCM treatment groups had levels of affective polarization or outgroup dislike that were significantly different from those of the control group.

Estimated Impact of Exposure on Affective Polarization/Outgroup Like

	(1)	(2)
VARIABLES	Affect. Polariz.	likeOut
CM	-0.939	0.347
	(2.680)	(1.900)
NCM	-1.043	-0.117
	(2.690)	(1.907)
Constant	49.744***	19.971***
	(1.892)	(1.341)
Observations	811	811
R-squared	0.000	0.000

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Affective Polarization (Disaggregated)

Table 13's results are also robust to the alternate specification. The interaction of CM and belief and NCM and disbelief remain statistically insignificant. None of them approach statistical significance either, with the lowest p-value of the four being 0.333 (column 4).

Estimated Impact of Exposure on Affective Polarization (Disaggregated)

VARIABLES	(1) Dem.	(2) Repub.	(3) Dem.	(4) Repub.
CM treatment	0.897	-3.816		
	(4.238)	(5.064)		
Believed both headlines	12.094**	19.134***		
	(4.863)	(5.733)		
CM *Believed both headlines	-1.645	1.384		
	(6.828)	(8.132)		
NCM treatment			-4.823	7.211
			(8.183)	(9.355)
Disbelieved 1+ headline			12.795**	27.515***
			(6.323)	(7.657)
NCM *Disbelieved 1+ headline			6.293	-10.022
			(9.070)	(10.325)
Constant	47.035***	40.682***	41.240***	24.682***
	(3.031)	(3.467)	(5.727)	(7.001)
Observations	283	261	264	276
R-squared	0.038	0.086	0.047	0.067

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Outgroup Like (Disaggregated)

I also checked whether the anticipated effects were perhaps evident in the outgroup dislike component of affective polarization. They were not. The four interaction terms did not approach statistical significance, with the lowest p-value being 0.275 (column 3).

Estimated Impact of Exposure on Outgroup Like (OLS, Disaggregated)

VARIABLES	(1) Dem.	(2) Repub.	(3) Dem.	(4) Repub.
CM treatment	-3.528	2.339		
	(3.136)	(3.478)		
Believed both headlines	-10.130***	-18.133***		
	(3.598)	(3.938)		
CM *Believed both headlines	4.876	2.676		
	(5.052)	(5.585)		
NCM treatment			3.597	-4.156
			(6.022)	(6.325)
Disbelieved 1+ headline			-10.066**	-12.942**
			(4.654)	(5.177)
NCM *Disbelieved 1+ headline			-7.306	6.589
			(6.675)	(6.981)
Constant	23.000***	27.541***	27.320***	31.727***
	(2.243)	(2.381)	(4.214)	(4.733)
Observations	283	261	264	276
R-squared	0.036	0.127	0.067	0.030

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

APPENDIX K

ROBUSTNESS CHECKS: ANGER, FEAR

These specifications dropped the partisanship and low attention variables that were included in the original analysis. The results are robust.

Whole Sample

The whole-sample results are robust to the alternate specification. Neither CM nor NCM treatment have statistically significant or near-significant effects on anger or fear.

Estimated Impact of Exposure on Emotion (OLS)

VARIABLES	(1) Angry	(2) Afraid
CM	-0.077	0.008
	(0.099)	(0.097)
NCM	0.038	0.087
	(0.100)	(0.097)
Constant	3.028***	2.864***
	(0.070)	(0.068)
Observations	811	811
R-squared	0.002	0.001

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Anger (Disaggregated)

Table 15's results regarding anger are also robust to the alternate specification. Anger increases among Democrats that believed CM at nearly-significant levels ($p = 0.11$), but not Republicans. Similarly, anger increases among Democrats but not Republicans that disbelieved NCM at near-significant levels ($p = 0.159$).

Estimated Impact of Exposure on Anger (Disaggregated)

VARIABLES	(1) Dem.	(2) Repub.	(3) Dem.	(4) Repub.
CM	0.014	0.082		
	(0.164)	(0.183)		
Believed both headlines	0.365*	0.960***		
	(0.188)	(0.207)		
CM*Believed both headlines	-0.424#	-0.281		
	(0.264)	(0.293)		
NCM			-0.397	0.084
			(0.322)	(0.331)
Disbelieved 1+ headline			-0.188	0.549**
			(0.249)	(0.271)
NCM* Disbelieved 1+ headline			0.504#	-0.014
			(0.357)	(0.366)
Constant	2.851***	2.714***	3.147***	2.606***
	(0.117)	(0.125)	(0.225)	(0.248)
Observations	283	261	264	276
R-squared	0.018	0.111	0.008	0.032

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1 # p<0.2

Fear (Disaggregated)

Table 16's results regarding fear are also robust to the alternate specification. Fear increases only among Democrats that disbelieved NCM at near-significant levels (p = 0.198).

Estimated Impact of Exposure on Fear (OLS, Disaggregated)

VARIABLES	(1) Dem.	(2) Repub.	(3) Dem.	(4) Repub.
CM	-0.129	0.018		
	(0.162)	(0.180)		
Believed both headlines	0.191	0.177		
	(0.186)	(0.204)		
CM*Believed both headlines	-0.020	0.312		
	(0.261)	(0.289)		
NCM			-0.262	0.161
			(0.328)	(0.316)
Disbelieved 1+ headline			-0.185	0.183
			(0.253)	(0.259)
NCM* Disbelieved 1+ headline			0.469#	-0.081
			(0.363)	(0.349)
Constant	2.994***	2.588***	3.220***	2.500***
	(0.116)	(0.123)	(0.229)	(0.237)
Observations	283	261	264	276
R-squared	0.011	0.029	0.009	0.004

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1 # p<0.2

APPENDIX L

ROBUSTNESS CHECKS: TPE

The original analysis' findings are robust to an alternate specification that drops respondent attentiveness.

Whole Sample

As in the original analysis, neither CM nor NCM treatment had even near-significant impacts on ingroup and outgroup TPE. The lowest p-value between the columns is 0.683.

Estimated Impact of Exposure on TPE

VARIABLES	(1) Ingroup Party	(2) Outgroup Party
CM	1.067	1.047
	(0.168)	(0.159)
NCM	1.037	1.066
	(0.162)	(0.163)
Observations	811	811

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

TPE (Disaggregated)

Table 18's disaggregated results are also robust to the alternate specification, save one exception. Democrats that disbelieved CM and NCM had higher levels of TPE, but Republicans did not. The one difference is that the CM treatment's effects on ingroup TPE approaches near-significance among Democrats (-34.6%, p – 0.143). In the previous analysis, its p-value was 0.275. This difference does not alter my conclusions.

Estimated Impact of Exposure on TPE

	Ingroup TPE (CM)		Outgroup TPE (NCM)	
	(1)	(2)	(3)	(4)
VARIABLES	Dem.	Repub.	Dem.	Repub.
Treatment	0.668#	1.083	0.384*	0.754
	(0.184)	(0.308)	(0.190)	(0.381)
Disbelieved 1+ headline	0.341***	1.157	1.115	1.317
	(0.114)	(0.382)	(0.425)	(0.565)
Treatment*Disbelieved	3.417***	0.969	3.293**	1.843
	(1.550)	(0.452)	(1.825)	(1.029)
Cong. (R) = 1				
Observations	283	261	264	276

seEform in parentheses

*** p<0.01, ** p<0.05, * p<0.1 # p<0.2

APPENDIX M

DIFFERENCE-IN-MEANS: OPT-IN/OUT BY PARTISANSHIP

Statistically significant differences-in-means are bolded.

Difference-in-Means by Opt-Out Status (Democrats)

	Opt-out	Opt-in	Diff	P-Stat
Female	**0.581**	**0.486**	**-0.095**	**0.07***
Education	4.384	4.314	-0.070	0.63
Age	41.075	39.664	-1.410	0.34
Rural	0.123	0.129	0.005	0.88
Suburban	0.530	0.493	-0.037	0.48
Urban	0.347	0.379	0.032	0.53
Religious Attendance	1.627	1.671	0.045	0.72
Pol. Ideology (L-C)	1.786	1.705	-0.081	0.40
Partisan Strength	3.246	3.343	0.097	0.22
SM: Have Unfullowed/Muted	0.858	0.871	0.013	0.71
SM: Freq. of Use	**6.004**	**6.229**	**0.225**	**0.05****
SM: Reddit	0.836	0.850	0.014	0.71
SM: Facebook	0.791	0.829	0.038	0.37
SM: TikTok	**0.440**	**0.629**	**0.188**	**0.00*****
SM: YouTube	**0.963**	**0.921**	**-0.041**	**0.07***
SM: Instagram	0.765	0.779	0.014	0.76
SM: No SM Usage	0.004	0.000	-0.004	0.47
SM News: Sci/Tech	**0.623**	**0.736**	**0.113**	**0.02****
SM News: Polit.	**0.537**	**0.664**	**0.127**	**0.01*****
SM News: Sports	**0.388**	**0.571**	**0.183**	**0.00*****
SM News: Busin.	**0.354**	**0.493**	**0.138**	**0.01*****
SM News: Celeb.	**0.414**	**0.536**	**0.122**	**0.02****
SM News: Other	0.358	0.343	-0.015	0.76
SM News: None	0.175	0.129	-0.047	0.22
Trust: News Media	2.466	2.329	-0.138	0.14
Trust: SM Fact-Checkers	2.757	2.907	0.150	0.16
Trust: Experts	3.840	3.900	0.060	0.59
Trust: Gov't	2.342	2.290	-0.052	0.49
GT	0.392	0.436	0.044	0.39
Alt: GT (Standard)	0.388	0.457	0.069	0.18
Alt: GT (GSS Index)	0.458	0.507	0.049	0.28
Trust: Ingroup Party	2.963	2.893	-0.070	0.48
Trust: Outgroup Party	1.489	1.429	-0.060	0.37
Pol. Knowledge (Index)	0.815	0.821	0.007	0.77
Pol. Interest: Campaign	**2.209**	**2.329**	**0.120**	**0.08***
Pol. Interest: General	3.381	3.500	0.119	0.26

Affect. Polariz.	51.918	51.671	-0.246	0.93
TPE: General	1.463	1.450	-0.013	0.91
TPE: Ingroup	0.940	1.029	0.088	0.42
TPE: Outgroup	2.000	1.950	-0.050	0.73
True: Burn Docs	3.720	3.593	-0.127	0.30
True: TX Textbooks	3.373	3.214	-0.159	0.25
True: NYT Coverup	2.578	2.600	0.022	0.87
True: Child Trafficking	2.075	2.050	-0.025	0.84
True: Police AI Use	3.929	3.771	-0.158	0.14
True: Afg. Activist Arrest	**4.373**	**4.186**	**-0.187**	**0.04****
True: Flat Packed Pasta	3.131	3.164	0.034	0.77
True: Patient Runs Marathon	3.989	3.979	-0.010	0.92
Click: Burn Docs	0.337	0.327	-0.010	0.90
Click: TX Textbooks	0.370	0.404	0.034	0.69
Click: NYT Coverup	0.260	0.167	-0.093	0.23
Click: Child Trafficking	0.286	0.229	-0.057	0.49
Click: Police AI Use	**0.459**	**0.593**	**0.134**	**0.01*****
Click: Afg. Activist Arrest	0.333	0.425	0.092	0.31
Click: Flat Packed Pasta	0.303	0.400	0.097	0.27
Click: Patient Runs Marathon	0.253	0.225	-0.028	0.73
Observations	408			

Difference-in-Means by Opt-Out Status (Republicans)

	Opt-out	Opt-in	Diff	P-Stat
Female	0.478	0.491	0.012	0.83
Education	4.145	4.009	-0.135	0.39
Age	**48.946**	**45.377**	**-3.569**	**0.05**
Rural	0.232	0.274	0.041	0.40
Suburban	0.562	0.575	0.013	0.81
Urban	0.205	0.151	-0.054	0.22
Religious Attendance	2.404	2.613	0.209	0.24
Pol. Ideology (L-C)	4.191	4.066	-0.125	0.18
Republican	2.000	2.000	0.000	.
Partisan Strength	3.259	3.340	0.080	0.30
SM: Have Unfullowed/Muted	0.778	0.774	-0.004	0.93
SM: Freq. of Use	5.721	5.962	0.242	0.10

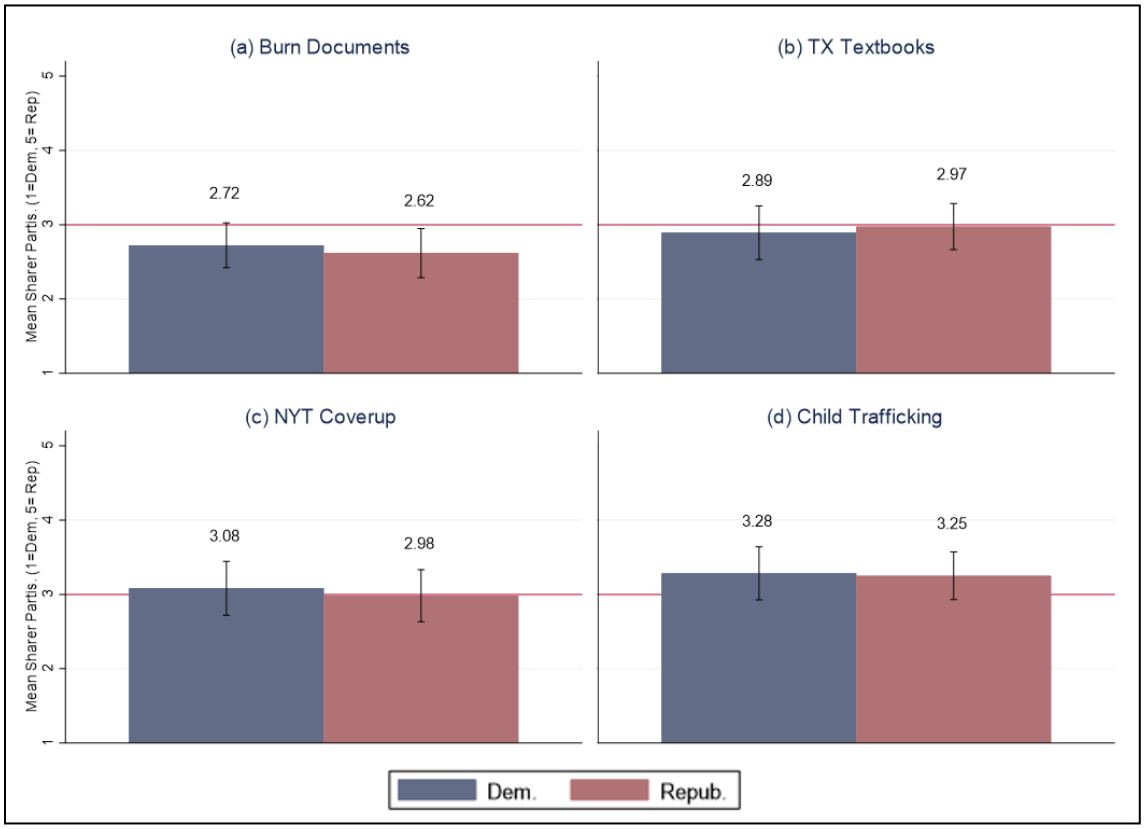
SM: Reddit	0.582	0.623	0.040	0.47
SM: Facebook	0.801	0.858	0.057	0.19
SM: TikTok	**0.337**	**0.462**	**0.126**	**0.02**
SM: YouTube	0.889	0.896	0.007	0.84
SM: Instagram	0.596	0.660	0.064	0.24
SM: No SM Usage	0.010	0.000	-0.010	0.30
SM News: Sci/Tech	**0.498**	**0.651**	**0.153**	**0.01**
SM News: Polit.	0.478	0.547	0.069	0.22
SM News: Sports	0.431	0.472	0.041	0.47
SM News: Busin.	0.387	0.406	0.018	0.74
SM News: Celeb.	0.286	0.283	-0.003	0.95
SM News: Other	0.290	0.245	-0.044	0.38
SM News: None	0.209	0.142	-0.067	0.13
Trust: News Media	1.606	1.736	0.130	0.15
Trust: SM Fact-Checkers	1.892	1.962	0.070	0.53
Trust: Experts	2.808	2.745	-0.063	0.62
Trust: Gov't	2.551	2.660	0.110	0.21
GT	0.313	0.377	0.064	0.23
Alt: GT (Standard)	**0.303**	**0.415**	**0.112**	**0.04**
Alt: GT (GSS Index)	0.389	0.447	0.057	0.24
Trust: Ingroup Party	2.926	3.000	0.074	0.53
Trust: Outgroup Party	1.589	1.698	0.109	0.22
Pol. Knowledge (Index)	0.825	0.818	-0.007	0.79
Pol. Interest: Campaign	2.249	2.302	0.053	0.49
Pol. Interest: General	3.481	3.528	0.047	0.70
Affect. Polariz.	45.694	48.019	2.325	0.54
TPE: General	**1.401**	**1.142**	**-0.259**	**0.04**
TPE: Ingroup	**0.983**	**0.774**	**-0.210**	**0.08**
TPE: Outgroup	**1.603**	**1.236**	**-0.367**	**0.03**
True: Burn Docs	2.229	2.245	0.016	0.91
True: TX Textbooks	2.290	2.264	-0.025	0.85
True: NYT Coverup	3.545	3.642	0.096	0.50
True: Child Trafficking	3.306	3.198	-0.108	0.47
True: Police AI Use	3.865	3.717	-0.148	0.20
True: Afg. Activist Arrest	4.128	4.075	-0.052	0.65
True: Flat Packed Pasta	2.801	2.943	0.142	0.32
True: Patient Runs Marathon	3.724	3.717	-0.007	0.96
Click: Burn Docs	0.262	0.314	0.053	0.55
Click: TX Textbooks	0.280	0.400	0.120	0.19
Click: NYT Coverup	0.473	0.618	0.145	0.15
Click: Child Trafficking	0.344	0.471	0.127	0.20
Click: Police AI Use	0.424	0.491	0.066	0.24

Click: Afg. Activist Arrest	0.227	0.351	0.125	0.14
Click: Flat Packed Pasta	0.227	0.189	-0.038	0.64
Click: Patient Runs Marathon	0.216	0.270	0.054	0.51
Observations	403			

APPENDIX N

INFERRED PARTISANSHIP BY MISINFORMATION HEADLINE

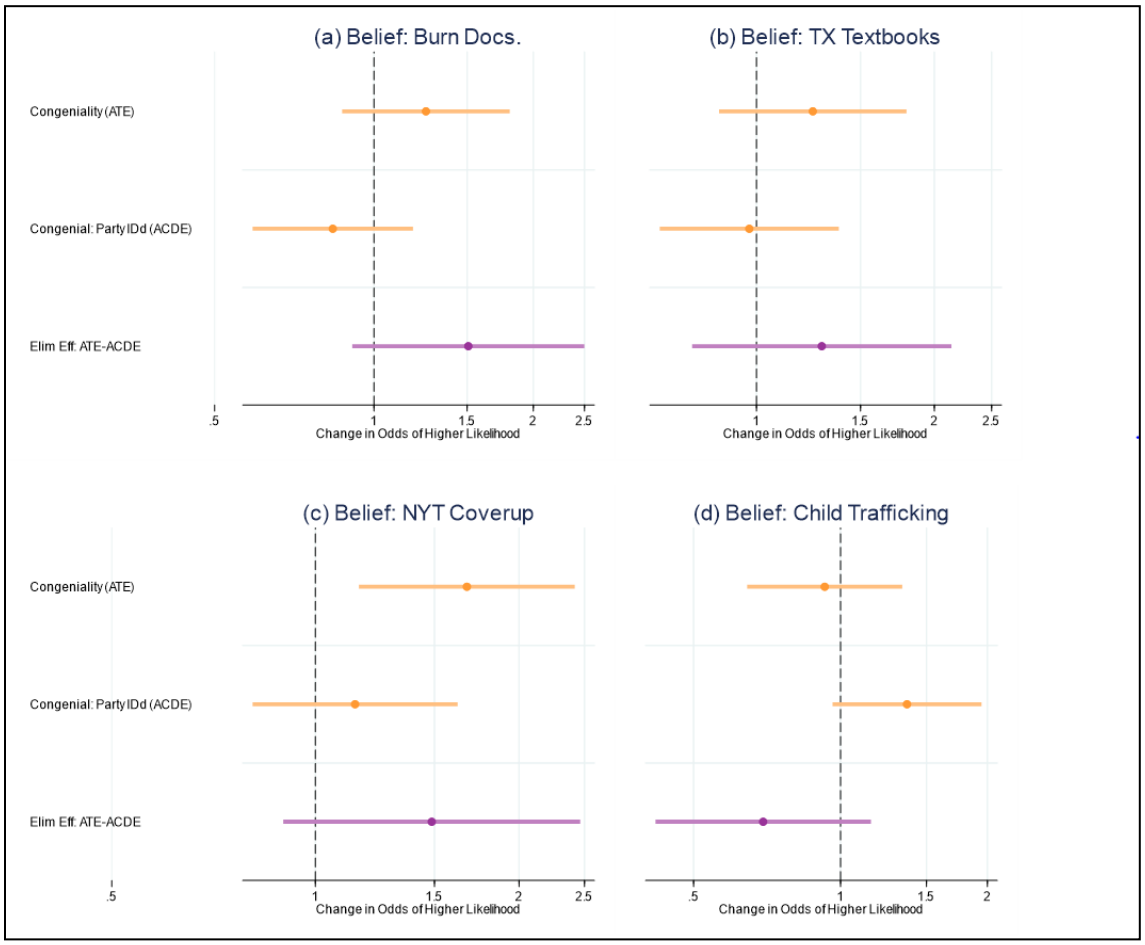
Means: Inferred Sharer Party among Treated Respond. not Told Sharer Party



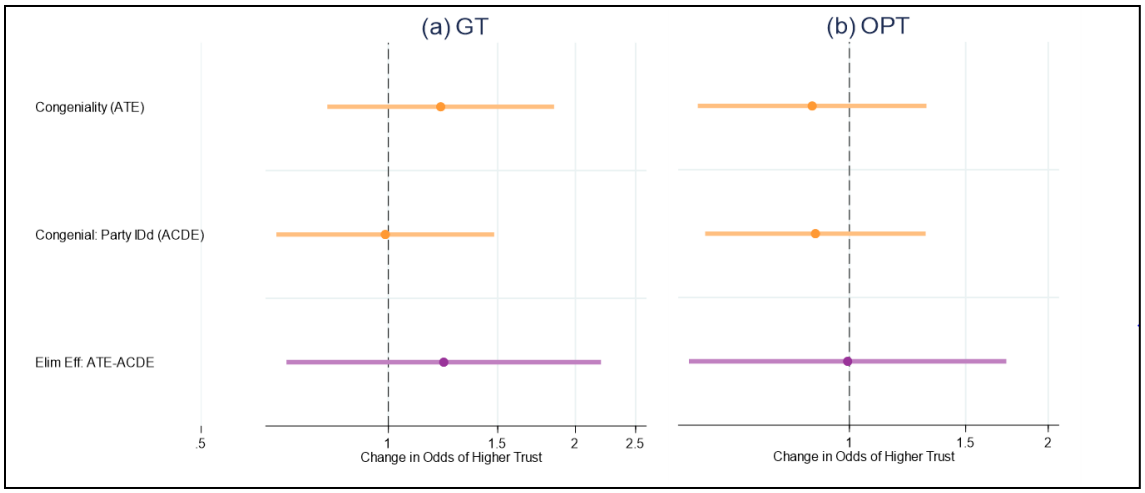
APPENDIX O

NON-EFFECT OF SHARER PARTY INFORMATION

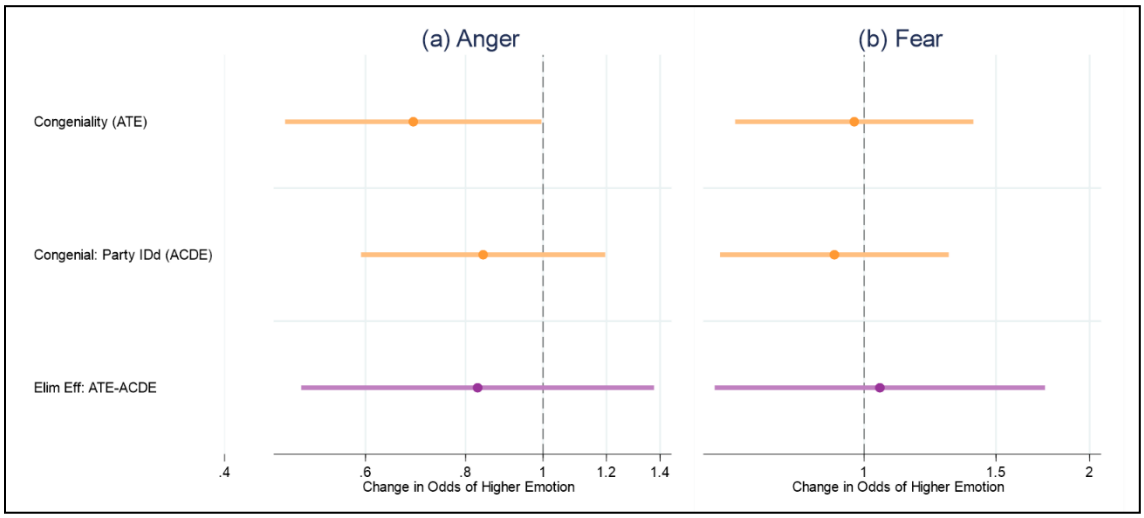
Non-Effect of Sharer Party on Belief



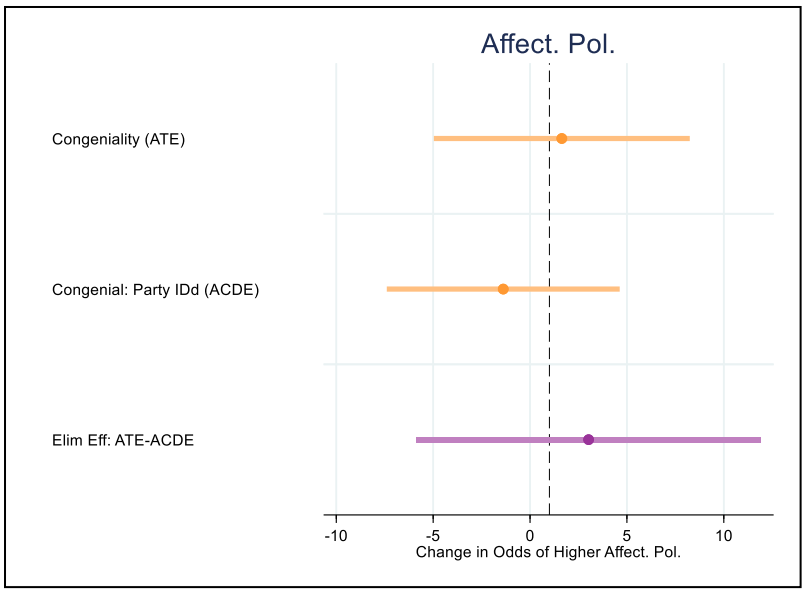
Non-Effect of Sharer Party on Trust



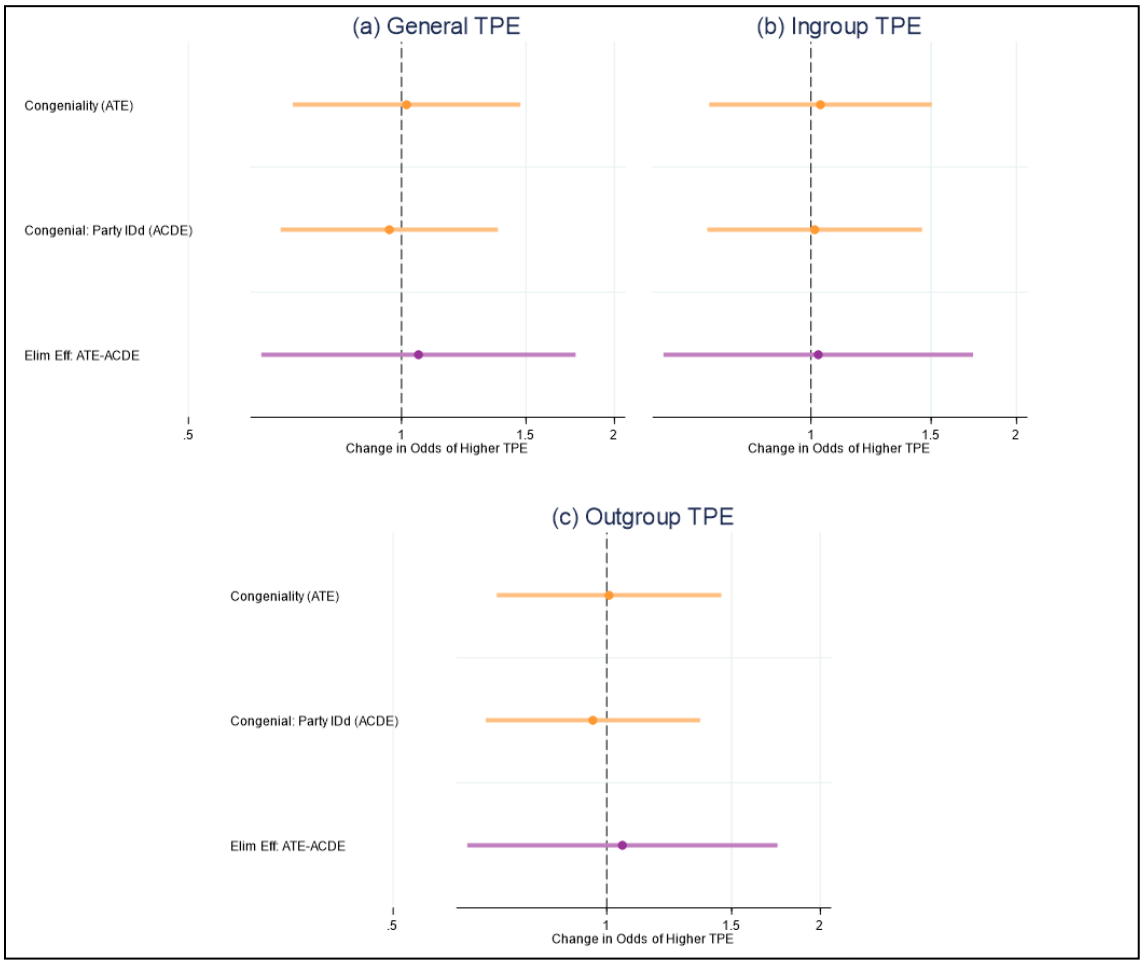
Non-Effect of Sharer Party on Emotion



Non-Effect of Sharer Party on Affect. Pol.



Non-Effect of Sharer Party on TPE



APPENDIX P

RESULTS: CORRELATIONAL ANALYSIS OF PEW DATA

I conducted an exploratory, correlational analysis to test the face validity of Hypotheses 1-2 in a cross-national context using a Pew Global Attitudes Survey dataset (Pew 2021b).[96] The analysis was divided into two parts, testing the link between assessments of domestic political dysfunction and perceptions of reliability, and then testing the link between perceptions of reliability and cooperation preference.[97] The results did not contradict the hypotheses: showing statistically significant and substantially-sized correlations between assessments of a state's domestic political functioning, its reliability in the international arena, and preference for foreign policy cooperation.

Data and Operationalizations

The respondents were from sixteen states in Europe, Asian, and North America, and were surveyed in spring 2021 (Pew 2021a).[98] The final sample was 16,254 respondents, at roughly 1,000 respondents per country. Pew weighted responses based on gender, age, region, and probability of a given respondent of being selected, and some countries' responses were weighted on additional factors (e.g., education, phone use, urbanicity).

The survey provided three variables that captured three key perceptions: how well "the political system in the United States" works, U.S. reliability as a partner to the respondent's country, and whether strong economic ties with the United States or China

[96] In accordance with Pew's terms of use, I include the following disclaimer: Pew Research Center bears no responsibility for the analyses or interpretations of the data presented here. The opinions expressed herein, including any implications for policy, are those of the author and not of Pew Research Center.

[97] Myrick establishes a causal link between domestic political dysfunction and polarization in the UK context, but this data allows me to establish its correlational plausibility in a cross-national context. Myrick does not establish the second link in terms of causation or correlation.

[98] Australia, Belgium, Canada, France, Germany, Greece, Italy, Japan, Netherlands, New Zealand, Singapore, South Korea, Spain, Sweden, Taiwan, United Kingdom

were more important. The first—*U.S. political system dysfunction*—is a proxy for perceived polarization and is the primary independent variable (IV) for the first part of the analysis. It is a four-point variable ranging from "very well" (1) to "not well at all" (4). The second is the dependent variable (DV) for first part of the analysis and the primary IV for the second. For ease of interpretation, I reverse-coded it, resulting in a four-point *U.S. unreliability* variable ranging from "very reliable" (1) to "not at all reliable" (4). The third—*cooperation preference*—is the DV for the second part of the analysis and is a dichotomous variable (0 = U.S. cooperation preference; 1 = China cooperation preference).

In addition to the primary IVs, I variously control for seven additional variables. The first is *disfavor* towards the United States. I created it from the dataset's U.S. favorability variable, which was four-point ranging from "very favorable" (1) to very unfavorable (4). This variable is of particular importance because it increases confidence that the probable relationship between perceptions of U.S. domestic political dysfunction and U.S. unreliability is *not* due to a common denominator of dislike of the United States. For the same reason, I included a four-point variable that captured responses to a question that asked respondents' confidence that U.S. President Joseph Biden would do the "right thing" in "world politics." This variable, *lack of confidence,* controls for the possibility that the DV and main IV are related via an underlying negative assessment of the current U.S. executive.

The five other control variables were *age*, *female*, being from a country with a *defense pact* with the U.S., being from a country with a major current *U.S. military presence*, and *2020 UNGA ideal point difference* from the United States. Of them, only

age and *female* are drawn from the Pew dataset, and they are count (18-97) and dichotomous variables, respectively. The other three variables conceptually cover foreign policy coincidence with the United States, and come from various sources.[99] Of note, including *UNGA ideal point difference* excluded the responses of non-UN member Taiwan from the sample.

Modeling Choice

The first part of the analysis (domestic dysfunction→unreliability assessments), I conducted a regression analysis of the entire dataset using generalized ordered logistic model. For robustness, I also conducted an ordinal logistic regression of each country's respondent subsample. For the second part of the analysis (unreliability assessments→cooperation hesitation), I used a logistic model due to the dichotomous DV.

I used a generalized ordinal logistic model rather than an ordinal logistic model for the first part of the analysis because a Wald test indicated that several of the variables violated the parallel regressions assumption. Also known as a partial proportional odds model, the generalized ordinal logistic model allows me to identify whether the non-compliant variables were my main variables of interest (Williams 2016). The model collapses the four-point ordinal DV into all possible dichotomous combinations of its

[99] *Defense pact* is based on NATO's membership list and a U.S. State Department list of bilateral defense agreements current through 2021 (NATO 2023; U.S. Department of State 2020; 2022).[99] *U.S. military presence* is based on the author's knowledge of which surveyed countries had a 30-year or greater presence of a brigade or greater of U.S. ground forces (Army or Marine Corps) or a major multinational ground forces headquarters.[99] Finally, *2020 UNGA ideal point difference* is the difference between each state's UNGA ideal point for 2020 and that of the United States (Bailey, Strezhnev, and Voeten 2017). Like *defense pact* and *U.S. military presence*, this variable conceptually falls under foreign policy coincidence with the United States.

original categories, and then runs simultaneous binary logistic regressions on the three combinations. Comparing the results of the three regressions allows identification of assumption-breaking variables. Those that have constant parameters across the three regressions meet the assumption, while those that have inconstant parameters violate the assumption. Initial review of the results reveal that the few assumption-violating variables did not include my main variables of interest.[100] In this model, I considered and ultimately omitted country fixed effects, as some of the country variables exhibited multicollinearity.

Results

Relationship 1: Political System Dysfunction→Unreliability Perceptions

As expected, greater perception of U.S. *political system dysfunction* is associated with greater perception that the U.S. is an *unreliable partner*. A one-point increase in perception that the political system in the United States "doesn't work" corresponds with a 60.4% increase in probability that the respondent viewed the United States as a more unreliable partner. This result is significant at a 1% level, even though the model controls for respondents' *disfavor* toward the United States, *lack of confidence* in President Biden to "do the right thing in world politics", and recent foreign policy coincidence captured in *2020 UNGA ideal point difference*. I cannot rule out that a latent negative evaluation of the United States is responsible for the relationship between the main IV and DV; but controlling for those three variables reduces that probability.

[100] The assumption-violating variables, as indicated by their varying odds ratios across the three binary logistic regressions, are *lack of confidence*, *2020 UNGA ideal point difference*, and *female*.

Generalized Ordinal Logistic Regression (Odds Ratios, Whole Avail. Sample)

VARIABLES	(1) VR *v*. SR, NTR & NAAR	(2) VR & SR *v*. NTR & NAAR	(3) VR, SR & NTR *v*. NAAR
U.S. Pol. Sys. Dysfunc.	1.604***	1.604***	1.604***
	(0.053)	(0.053)	(0.053)
Disfavor U.S.	1.874***	1.874***	1.874***
	(0.068)	(0.068)	(0.068)
Lack of Confid. in Biden	1.940***	1.868***	2.179***
	(0.126)	(0.067)	(0.131)
Defense Pact w/U.S.	1.071	1.071	1.071
	(0.072)	(0.072)	(0.072)
U.S. Mil. Presence	0.929	0.929	0.929
	(0.053)	(0.053)	(0.053)
UNGA Ideal Point Diff.	1.650***	1.672***	1.065
	(0.136)	(0.096)	(0.112)
Age	0.998	0.998	0.998
	(0.001)	(0.001)	(0.001)
Female	1.727***	1.298***	0.944
	(0.127)	(0.067)	(0.092)
Constant	0.061***	0.003***	0.000***
	(0.013)	(0.001)	(0.000)
Observations	14,250	14,250	14,250

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1
VR = "very reliable", SR = "Somewhat Reliable", NTR = "Not Too Reliable", NAAR = "Not at All Reliable"

These results are largely but not wholly robust within each country's subsample (below). The odds ratios for nearly all the countries across all three constituent logistic regressions show that greater perceived U.S. *political system dysfunction* is associated with greater perception that the U.S. is an *unreliable partner*. The magnitudes vary with the lowest increased odds being 54.8% (Sweden) and highest increased odds being 108.9% (Greece), but they are all significant at a >99% level.

The results were not robust, though, for Australia and the United Kingdom. For both, only two of the logistic regressions showed statistically significant, positive effects (columns 2-3). The first regression, which compared those who thought the U.S. "very reliable" with all the other respondents, showed null effects. These null effects, though, do not call into question the larger conclusions. They may suggest some unique factor involving English-speaking states' relationship with the United States, but the significant results for Canada and New Zealand seem to belie that possibility.

Political System Dysfunction By-Country (Odds Ratios)

VARIABLES	(1) VR *v.* SR, NTR & NAAR	(2) VR & SR *v.* NTR & NAAR	(3) VR, SR & NTR *v.* NAAR
Australia	*1.022*	*1.834***	*1.713**
	(0.157)	(0.266)	(0.473)
Belgium	1.877***	1.877***	1.877***
	(0.283)	(0.283)	(0.283)
Canada	1.645***	1.645***	1.645***
	(0.201)	(0.201)	(0.201)
France	1.918***	1.918***	1.918***
	(0.324)	(0.324)	(0.324)
Germany	1.716***	1.716***	1.716***
	(0.241)	(0.241)	(0.241)
Greece	2.089***	2.089***	2.089***
	(0.328)	(0.328)	(0.328)
Italy	2.049***	2.049***	2.049***
	(0.353)	(0.353)	(0.353)
Japan	1.997***	1.997***	1.997***
	(0.436)	(0.436)	(0.436)
Netherlands	1.724***	1.724***	1.724***
	(0.245)	(0.245)	(0.245)
New Zealand	1.608***	1.608***	1.608***
	(0.173)	(0.173)	(0.173)
Singapore	1.685***	1.685***	1.685***
	(0.242)	(0.242)	(0.242)
South Korea	1.685***	1.685***	1.685***
	(0.242)	(0.242)	(0.242)
Spain	1.724***	1.724***	1.724***

	(0.208)	(0.208)	(0.208)
Sweden	1.548***	1.548***	1.548***
	(0.234)	(0.234)	(0.234)
Taiwan	1.771***	1.771***	1.771***
	(0.330)	(0.330)	(0.330)
United Kingdom	*1.155*	*1.832***	*3.001****
	(0.202)	(0.273)	(0.840)

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Relationship 2: Unreliability Perceptions→Cooperation Hesitation

The second set of results were also as expected. Each 1-unit increase in perceived unreliability of the U.S. as a partner corresponded with a 59% increase in belief that it was more important to have strong economic ties with China than the United States.

Initial Results (Logit, Odds Ratios)

VARIABLES	Prefer China Econ. Partner
U.S. Unreliable Partner	1.590***
	(0.069)
U.S. Pol. Sys. Dysfunc.	1.128***
	(0.042)
Disfavor U.S.	1.419***
	(0.059)
Lack of Confid. in Biden	1.146***
	(0.042)
Defense Pact w/U.S.	0.593***
	(0.050)
U.S. Mil. Presence	1.063
	(0.081)
UNGA Ideal Point Diff.	0.783***
	(0.056)
Age	0.984***
	(0.002)
Female	1.064
	(0.061)
Constant	0.134***
	(0.030)
Observations	12,343

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Discussion

This analysis does not establish causation. Instead, it answers the more modest question of whether the causal chain survives initial empirical efforts to disprove it with samples broader than the United Kingdom. It does. These results show correlations consistent with Hypothesis 1 and 2: a statistically significant, substantially-sized link between assessments of a state's domestic political functioning, its reliability in the international arena, and preference for foreign policy cooperation.

APPENDIX Q

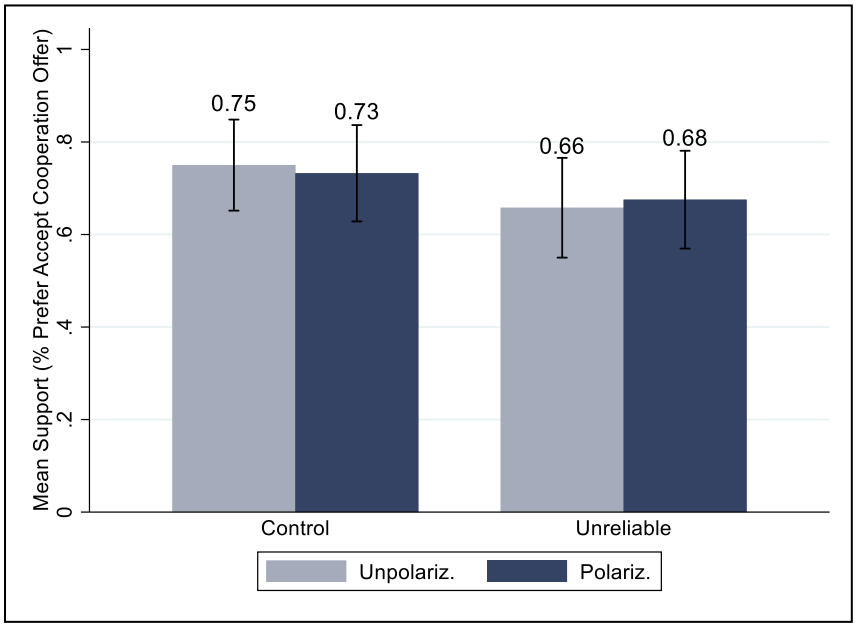
RESULTS: PILOT STUDY

The pilot study found no causal effect of Aurl's polarization on respondents' cooperation preference. The results, though, suggested political sophistication could be a moderator, increasing respondents' abilities to identify the links from polarization to defection risk. The increases still fell short of standard levels of statistical significance (though some approached them), so the results only suggested political sophistication would be a moderator in a larger sample. Further, the results suggest that respondent attentiveness is a moderator, as respondents who encountered the experiment earlier in the survey had results that more conformed to my expectations.

Whole-Sample Results

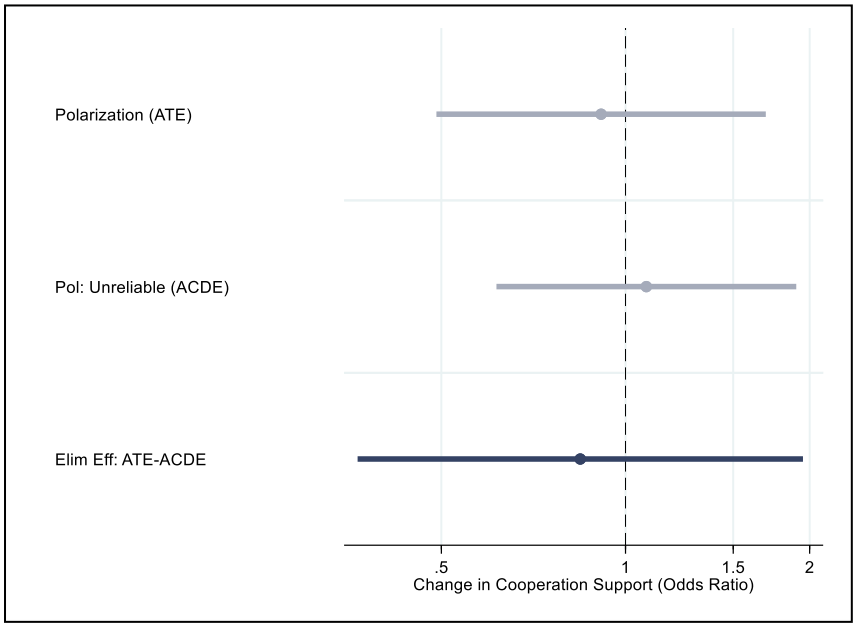
The results, at first glance, seem unsupportive of the hypotheses, both in terms of means and the three estimates of interest. The difference-in-means between the treatment groups' cooperation preference is unremarkable. There is no statistically significant difference between the unpolarized control group (information about unreliability omitted) and the polarized control group, suggesting that polarization status has no impact on cooperation preference.

Means: Offer Acceptance



Regression results for the whole sample are similarly counter-expectation. Below are odds ratios with 90% confidence intervals for the ATE, ACDE, and EE of polarization on cooperation preference. All three are statistically insignificant with the lowest p-value is EE's at 0.738. Polarization seems not to matter at all.

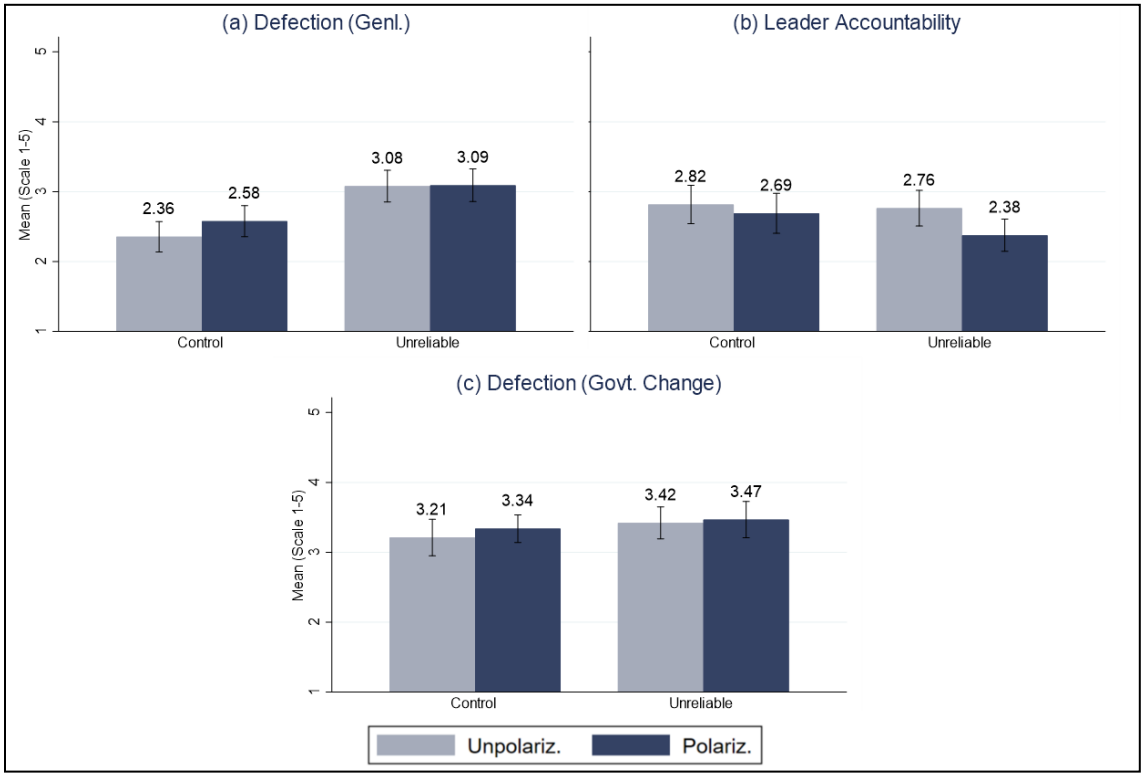
Est. Impact of Polarization on Offer Acceptance (Odds Ratios)



However, examination of the causal logic variables shows that polarization did matter to respondents' perceptions about Aurl defection in general (below). In the control group, those told that Aurl was polarized had a higher mean assessed likelihood of Aurl defecting than those told Aurl was unpolarized (2.58 v. 2.36). The difference approached statistical significance ($p = 0.16$). That said, the level of unreliability inferred from polarization does not appear to have been strong, as the addition of information about Aurl sometimes being unreliable increased the polarized groups' mean (2.58 v. 3.09, $p = 0.00$). But, the near significant impact in the control conditions remains. Similar patterns, though at lower levels of significance are evident for the two causal logic variables. Mean probability of leader accountability under conditions of polarization are lower than under conditions of non-polarization (2.69 v. 2.82, $p = 0.53$), and mean probability of Aurl government changeover meaning defection is higher (3.338 v. 3.211, $p = 0.45$).

The difference in significance could suggest that the measures were noisy; that respondents can infer defection likelihood from polarization, but do not tease out their assessment at lower levels of abstraction; or that the two hypothesized causal pathways do not exist. Further, polarization's possible impact on perceived defection chances suggests that the scenario may have been imbalanced, tilting the scenario risk in favor of cooperation with Aurl. If respondents identified that polarization marginally increased defection chances, and given that majorities of even the unreliable groups preferred offer acceptance, then respondents may have felt that it was better to risk cooperation with a state that might cheat than to place oneself at the mercy of an aggressor state.

Means: Causal Logic Variables

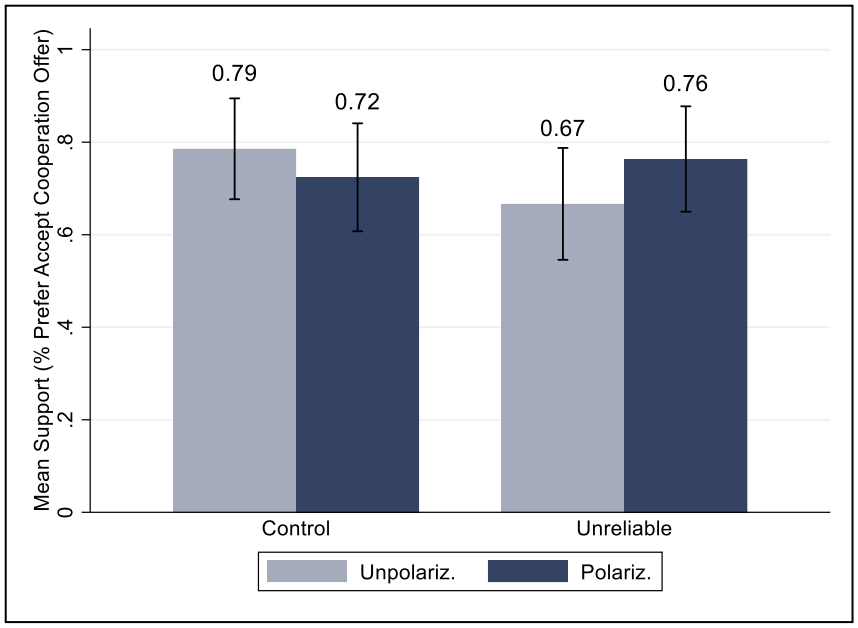


Sub-Sample Results: SPGS Majors

The whole-sample's exploratory results are even more evident within a subsample of respondents. The subsample is respondents who self-identified as majoring or thinking about majoring in disciplines offered by the School of Politics and Global Studies (political science, politics and the economy, global studies). They accounted for 76.33% of the whole sample, with treatment group sizes of 55-60 (n = 229), and for ease of use, I will refer to them as politics-related majors or just majors.

In the "control" condition for unreliability, the politics-related majors told Aurl was polarized had a lower mean acceptance rate than those told Aurl was unpolarized

Means: Offer Acceptance (Subsample)



(0.72 v. 0.79). This difference only has a significance level of 55%, but is much closer than its significance in the whole-sample analysis (19%). Also importantly, the addition of information about Aurl's reliability in the polarized condition hardly changes the mean at all (0.72 v. 0.76), while the addition of information about Aurl's unreliability in the

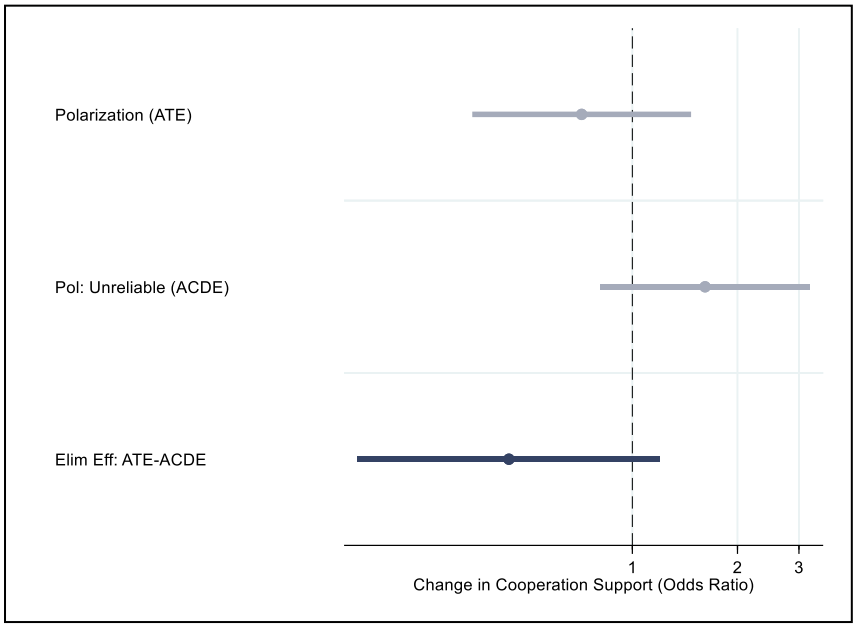
unpolarized condition drops the mean from 0.79 to 0.67 ($p = 0.15$). This suggests that higher levels of unreliability were inferred by respondents told Aurl was polarized than those told it was unpolarized, as the latter were more sensitive to information about Aurl's unreliability.

Finally, the difference between the unpolarized and polarized treatments in the subsample's unreliable condition indicates that there may be an unanticipated, positive interaction between polarization and unreliability. Mean offer acceptance is lower when respondents are told Aurl is unreliable and unpolarized than when they are told it is unreliable and polarized (0.76 v. 0.67, $p = 0.25$). Again, this only approaches standard levels of significance and is in a small-*n* analysis. But, it could suggest that countries that concur about breaking their commitments are more concerning to respondents than countries that perhaps break their commitments because they struggle to reach consensus.

Moving on to regression analysis of the major subsample, Acharya et al.'s indicators of relative direct/indirect impacts of polarization reflect the difference-in-means results (below). Polarization's overall impact (ATE) is in the expected direction, but nowhere near significant (28.41% lower odds, $p = 0.446$). Still insignificant is polarization's direct effect on offer acceptance (ACDE, 61.54% increased odds, $p = 0.253$). Finally, polarization's indirect effect mediated by unreliability (EE) is in the expected direction and almost significant (55.68% lower odds, $p = 0.18$). On the substantial assumption that it would be significant with a larger sample size, this suggests that polarization has no direct effect, but has an indirect causal effect through/by unreliability. Given the possible positive interaction effect between polarization and

unreliability, the negative EE suggests the indirect effect is larger than any positive interaction effect.

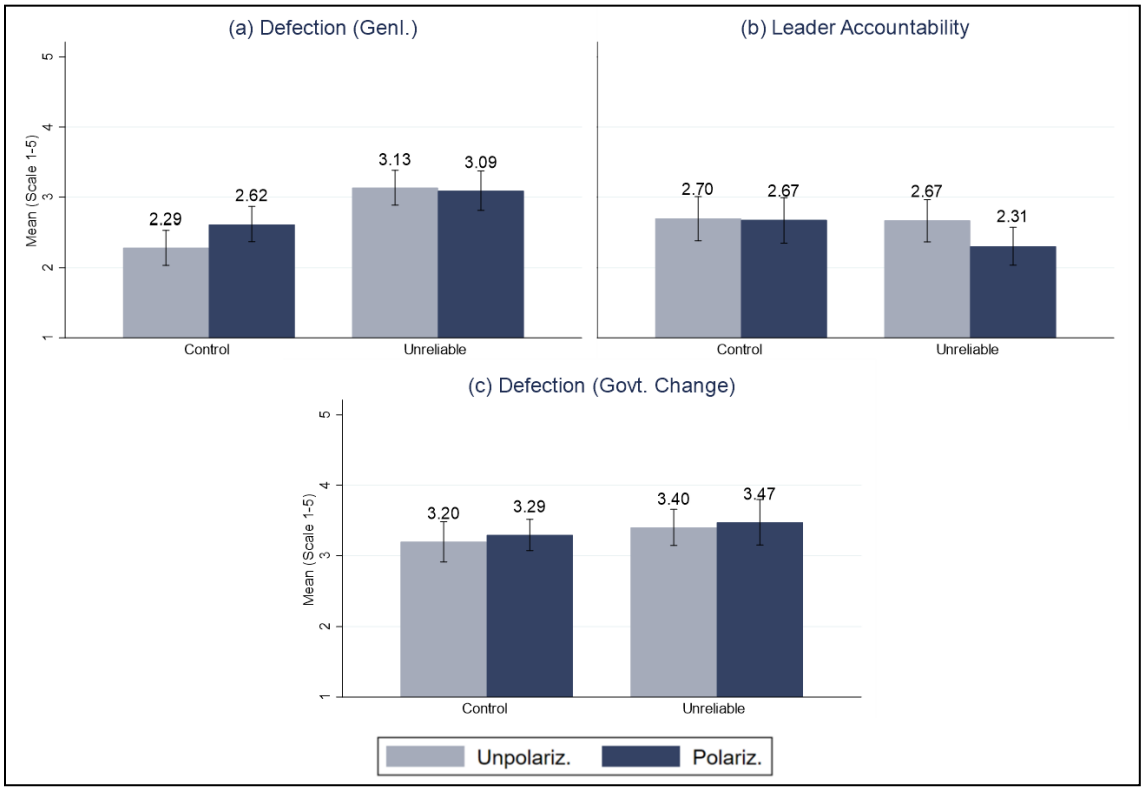
Est. Effect of Polarization on Offer Acceptance (Subsample)



The causal logic outcomes in the major subsample also more strongly evidence the expected polarization impacts on general defection chances (below). In the control condition for unreliability, the positive difference-in-means between those told Aurl was polarized v. non-polarized was greater and closer to statistical significance than in the whole-sample analysis (difference of 0.32, $p = 0.16$). Noteworthily, though, the major subsample results mirror the whole-sample results for leader accountability and defection likelihood in case of government changeover. That is, polarization's impacts on those two variables appears to be null. The difference-in-means between polarized and unpolarized in the control conditions are even more tightly-ranged than in the whole-sample analysis. Given this lack of impact within the politically more-sophisticated subsample, the causal logic results suggest that the null impact of polarization on prospects of leader

accountability or government changeover defection in the whole sample was not due to a weak political reasoning skills.

Means: Causal Logic Variables (Subsample)



Sub-Sample Results: Less Cognitively Taxed SPGS Majors

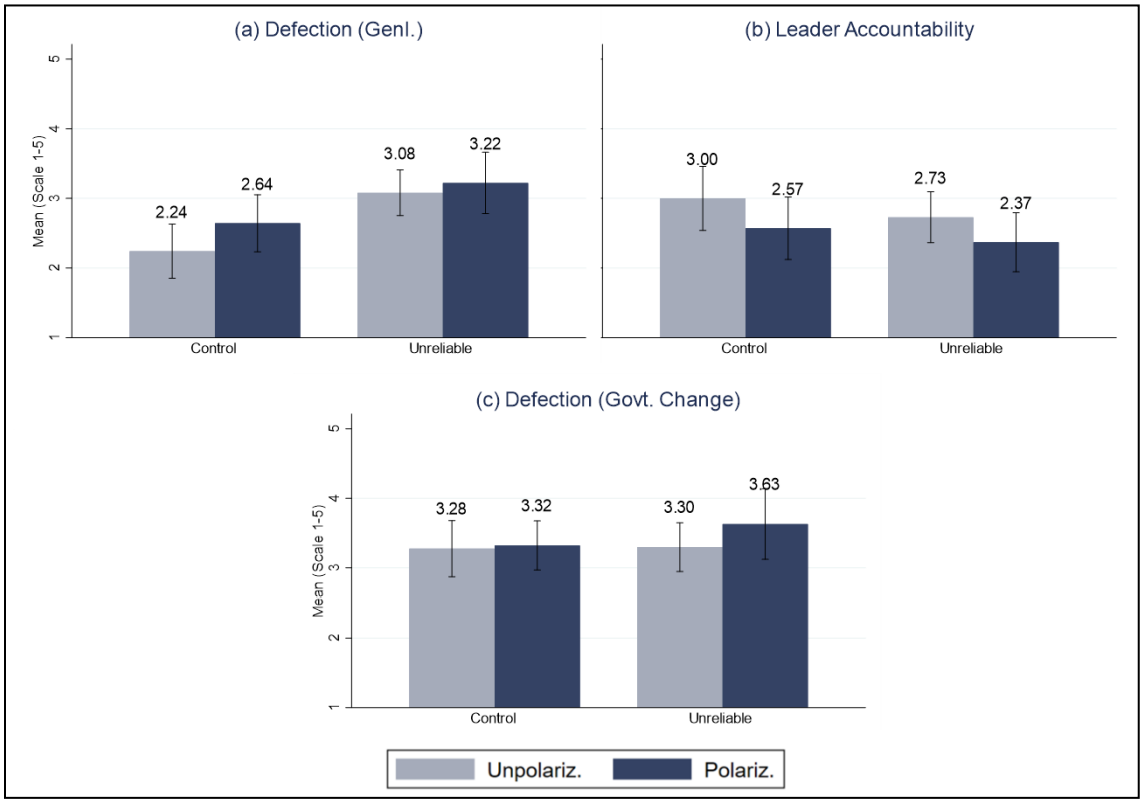
In conducting the analysis, I realized I had failed to include an instructional manipulation check for the pilot study. That meant that I could not see whether the expected polarization impacts were present among attentive respondents. Other than highlighting a refinement to the survey for the study's full fielding, this also led me to search for a proxy variable. I found one in the variable capturing the study's scenario order.

The study contained two sequential scenarios with their order randomized (Country X, Aurl/Zerm). Respondents who encountered the Aurl/Zerm scenario first

likely encountered the scenario less cognitively taxed than those who encountered it second. So, I conducted the causal logic difference-in-means analysis for a third time, this time focusing on the subsample of majors who encountered the Aurl/Zerm scenario first. Due to the small subsample size, I did not re-re-run the difference-in-means and regression analyses regarding polarization's ATE, ACDE, and EE on cooperation preference (treatment group sizes ranged from only 27-37).

The causal logic results suggest that attentiveness is a moderator in addition to political sophistication (below). Despite the smaller sample size, the positive difference-in-means is more significant than in the whole-sample analysis and just as significant as in the larger major analysis ($p = 0.16$). The magnitude is also larger than in both previous analyses (0.4). Further evidence is seen in the differential impact of the unreliability information between the unpolarized and polarized conditions. Addition of the unreliability information resulted in an increase of 0.840 ($p = 0.00$) among respondents told Aurl was unpolarized, but resulted in a smaller increase of 0.579 ($p = 0.06$) among those told Aurl was polarized. As with the previous analyses in this appendix, this differential impact of the unreliability information across the polarization conditions suggests that respondents inferred some level of unreliability from Aurl being polarized.

Figure 7. Means: Causal Logic Variables (Second Subsample)



The less-abstract causal logic variables demonstrate mixed results in the less cognitively-taxed subsample of majors. Polarization's impact on leader accountability is more pronounced in this subsample than the whole sample and in the larger major subsample, with a difference-in-means between the polarized/unpolarized groups in the control condition being -0.429 ($p = 0.16$). Adding information about Aurl's unreliability also shows differential impacts by polarization condition, as there is a greater difference by unreliability condition among respondents told Aurl was non-polarized than polarized (-0.27 v. -0.20). However, the means suggest no impact of polarization on perceptions that Aurl government changeover could result in defection. Indeed, it shows that the addition of information about unreliability has an impact on respondents told Aurl was

polarized and no impact on respondents told Aurl was non-polarized (0.308, $p = 0.33$). The impact is not significant ($p = 0.33$) but could be at larger sample sizes. However, this differential may communicate more about the subsample than polarization v. non-polarization's relative level of inferred unreliability. More politically sophisticated and encountering the scenario with a fresh mind, the students could have been better able to recognize that Aurl's demonstrated reliability in the time since offer acceptance likely meant it would continue to be reliable through government changeover, despite its reputation for unreliability.

Conclusion: Theoretical Implications and Protocol Refinements

In sum, the pilot study results have three main theoretical takeaways. First, political sophistication and attentiveness may be critical moderators. In none of the analyses did polarization impact offer acceptance; but, in sub-sample analyses that focused on students with politics-related majors and particularly those who encountered the scenario first in the study, it came much closer to. This closeness could suggest impacts of polarization moderated by political sophistication and attentiveness in a larger sample. Second and relatedly, those results suggest Myrick's analysis about the impact of polarization may hold for cases other than U.S. polarization. Thirdly, polarization's impact on defection chances may be broader than my two causal pathways. The sub-sample results show various impacts of polarization on likelihood of leader accountability for defection and government changeover resulting in defection; but, the impacts are smaller in magnitude than polarization's impact on perceived likelihood of defection in general.

These theoretical implications had two practical implications for the full fielding study. First, I needed to add an instructional manipulation check to capture attentiveness. Scenario order was only a rough proxy and will not exist in the full fielding anyway, as I lack funds to repeat the Country X experiment. Second, I needed to relook the scenario to make it less slanted toward cooperation with Aurl. A majority of respondents, even in the major subsample, preferred to accept Aurl's offer regardless of its polarization/unreliability and despite polarization increasing perceptions that Aurl could defect in general. This suggests that respondents may have felt that even though the partner state was unreliable, they had no choice but to gamble and rely on them.

To try to achieve this, I altered the scenario by including a bullet point that clarified Zerm's asserted intentions. Zerm, in the new version, said it would interpret a rejection of Aurl's offer as meaning the country was non-aggressive toward Zerm and acceptance of Zerm's offer as evidence of the respondents' country's aggression. That is, they say they will not invade in case of offer rejection but will invade (given the chance) in case of offer acceptance. This is in contrast to the old version, which said nothing of Zerm's intentions if respondents rejected Aurl's offer. In addition to the alteration, I added a supporting subjective manipulation check. The question queried respondents' perceptions about the likelihood of Zerm invading even if they reject Aurl's offer.

APPENDIX R

IRB APPROVALS: PAPER 3

Pilot Study

EXEMPTION GRANTED

Timothy Peterson
CLAS-SS: Politics and Global Studies, School of (SPGS)
-
Timothy.M.Peterson@asu.edu

Dear Timothy Peterson:

On 11/6/2023 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Foreign Policy: The Link between Polarization and International Cooperation Hesitation
Investigator:	Timothy Peterson
IRB ID:	STUDY00018968
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• consent_form_25-10-2023, Category: Consent Form; • recruitment_methods_email_25-10-2023, Category: Recruitment Materials; • recruitment_methods_syllabus_25-10-2023.pdf, Category: Recruitment Materials; • social_behavioral_protocol_25-10-2023, Category: IRB Protocol; • supporting_documents_25-10-2023, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions);

The IRB determined that the protocol is considered exempt pursuant to Federal Regulations 45CFR46 (2)(ii) Tests, surveys, interviews, or observation (low risk) on 11/6/2023.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Full Fielding Study

ASU Knowledge Enterprise Development

EXEMPTION GRANTED

Timothy Peterson
CLAS-SS: Politics and Global Studies, School of (SPGS)
-
Timothy.M.Peterson@asu.edu

Dear Timothy Peterson:

On 2/29/2024 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Foreign Policy Opinion: Links between Polarization and International Cooperation Hesitation
Investigator:	Timothy Peterson
IRB ID:	STUDY00019702
Funding:	Name: EOSS: Educational Outreach and Student Services
Grant Title:	
Grant ID:	
Documents Reviewed:	• consent_form_02-26-2024, Category: Consent Form; • grant award letter_02_26_2024, Category: Sponsor Attachment; • protocol_02-26-2024, Category: IRB Protocol; • recruitment_methods_advertisement_02-26-2024, Category: Recruitment Materials; • supporting_documents_02-26-2024, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions);

The IRB determined that the protocol is considered exempt pursuant to Federal Regulations 45CFR46 (2)(ii) Tests, surveys, interviews, or observation (low risk) on 2/29/2024.

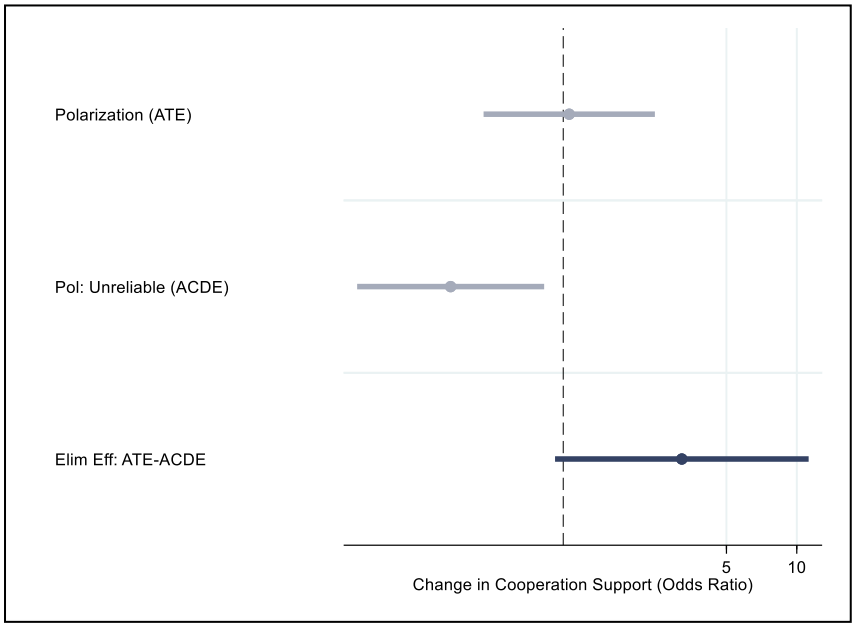
In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

APPENDIX S

RESULTS: EXPERIMENT 1 (SUBSAMPLE)

When focused on attentive respondents in the top quartile of political sophistication, the analysis supports Hypothesis 1 but not Hypothesis 2 (below figure). Polarization's ATE on acceptance of Aurl's offer again is statistically insignificant. But, where the whole-sample analysis found null effects for ACDE, this subsample analysis shows a statistically significant effect (-67.1% odds of offer acceptance, $p = 0.048$) and as a result, a borderline significant EE (222% increased odds of offer acceptance, $p = 0.124$). That is, Aurl being polarized rather than unpolarized directly decreases odds of offer acceptance (ACDE), which supports Hypothesis 1. However, Hypothesis 2's expectation that unreliability would account for a portion of polarization's negative impact is not supported by the results (EE). As EE is the difference between the near-zero ATE and the negative ACDE, Aurl being polarized has an unexpectedly positive indirect/interaction effect on offer acceptance via unreliability perceptions. That is, unreliability plays a role in the causal mechanism from polarization to offer acceptance, but it works against polarization's expected negative impact.

Est. Impact of Polarization on Offer Acceptance (Subsample, Odds Ratios)



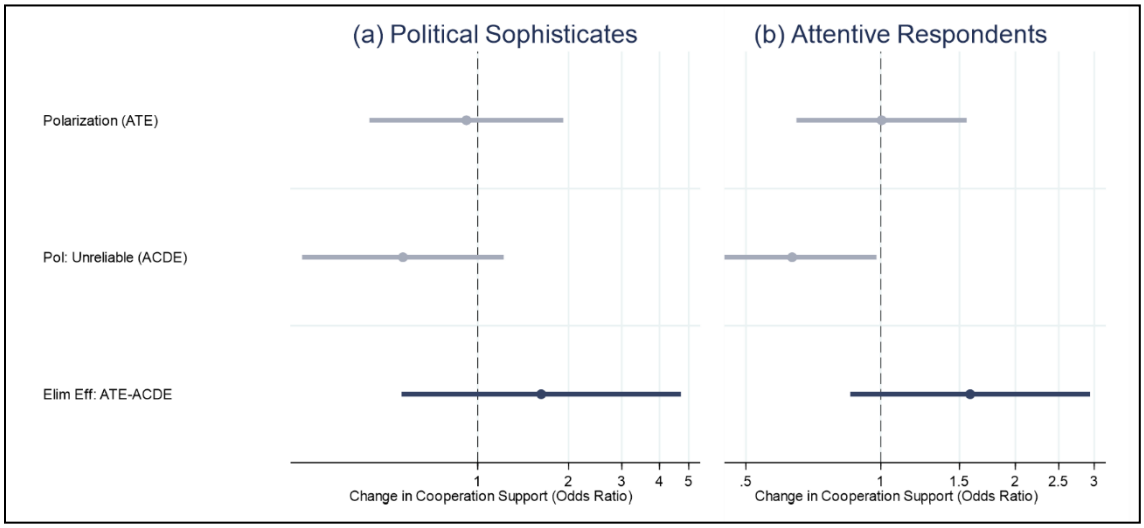
ATE: avg. treatment effect; ACDE: average direct controlled effect; EE: eliminated effect
Coefficients are logarithmically transformed to facilitate ease of visual interpretation
Model: logistic

The unexpectedly positive EE is open to two interpretations (Acharya, Blackwell, and Sen 2018, 375). The first is that the hypothesized negative indirect effect exists but is masked by a countervailing dynamic. EE is composed of polarization's indirect effect via unreliability as well as its interaction effect with unreliability. If the indirect effects were negative but the interaction effects were positive and larger, the result would be a positive EE. Such an interaction is possible. That is, an unreliable, polarized state could seem less likely to defect than an unreliable, unpolarized state. The former is unreliable, but the latter is not only unreliable, but even worse, has domestic consensus about their unreliability. This potential positive interaction also plays a role in the second interpretation of the positive EE, namely: there is no indirect effect via unreliability and only the positive interaction is in play. Theoretically, that seems improbable, but I will test the plausibility of potential alternate mediators Hypothesis 3's analysis.

Other than offering limited support for Hypothesis 1 and no support for Hypothesis 2, the subsample analysis of attentive political sophisticates also raises a new question. Were both political sophistication and attentiveness moderators, or was only one at play? Figure 7 suggests both were moderators, but attentiveness played a larger role than sophistication. Focusing the analysis on just the attentive respondents changes the ACDE from being insignificant in the original whole-sample analysis to significant ($p = 0.084$). That is, attentive respondents showed the direct effect of polarization on unreliability. The same is not true for political sophisticates. Focusing on them made the ACDE more significant than in the original analysis but did not make it significant ($p = 0.222$ v. 0.417).

However, sophistication played a stronger role in moderating the indirect/interaction effects contained in EE. While high attentiveness majorly increased the significance of EE relative to the original whole-sample analysis, The figure shows it did not increase it enough for EE to achieve standard levels of significance ($p = 0.774$ v. 0.220). Political sophistication singly did not achieve that either, with EE being more significant for political sophisticates than the whole sample, but still not significant enough ($p = 0.455$). Thus, the near-significance of EE in the subsample of attentive political sophisticates was due to both factors.

Est. Impact of Polarization by Potential Moderator



ATE: avg treatment effect; ACDE: average direct controlled effect; EE: eliminated effect
Coefficients are logarithmically transformed to facilitate ease of visual interpretation
Model: logistic

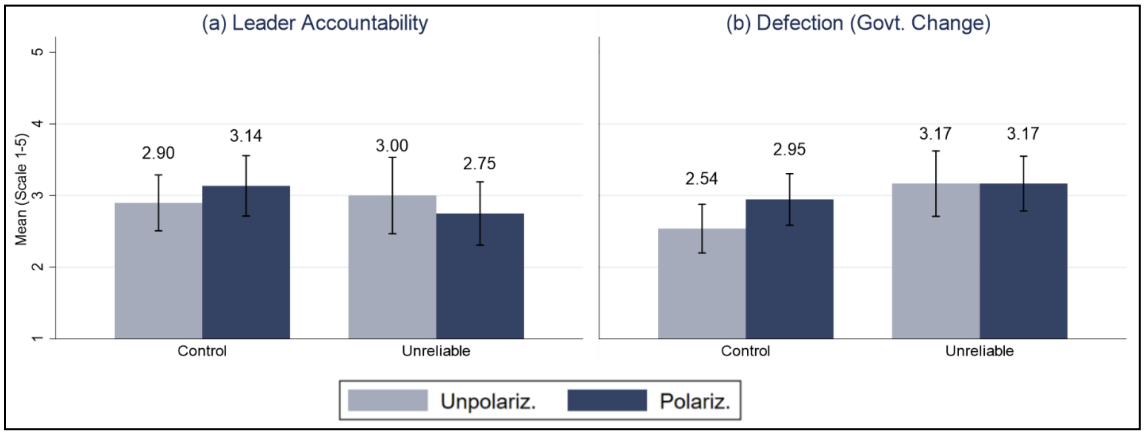
Overall, these analyses offer limited support for Hypothesis 1, no support for Hypothesis 2, and suggest respondent attentiveness and political sophistication as moderators. Polarization decreases support for cooperation, but I unexpectedly only have evidence of it directly decreasing support for cooperation. No indirect effect via unreliability was evidenced, though it may have been washed out by a theoretically plausible positive interaction effect between polarization and unreliability.

Causal Logic: Leader Accountability and Government Changeover

As political sophistication and respondent attentiveness conditioned the results for Hypothesis 1, I conducted a second difference-in-means analysis focusing on the same subgroup of attentive political sophisticates. This subgroup's means offer no more support for Hypothesis 3a than the whole- sample analysis did, but does support

Hypothesis 3b (below figure). First, regarding Hypothesis 3a: respondents' perceived

Means: Gov't Changeover, Accountability by Treatment Group (Subsample)



likelihood that supporters of Aurl's leader will hold the leader accountable if they defect from their agreement is *higher* on average in the polarized/control group than the unpolarized/control group (3.14 v. 2.90). That is counter-expectation, but also is statistically insignificant (p = 0.41). These results do not show the expected decrease in accountability likelihood when a state is polarized. The expected pattern is evidenced, however, with regard to the likelihood of Aurl defection if its government changes hands. Mean perceived likelihood of defection in case of government changeover was higher among respondents told Aurl was polarized than those told the opposite, a difference that approached standard levels of significance (2.95 v. 2.54, p = 0.11). This is consistent with Hypothesis 3b.

A series of ordinal logistic regressions support the difference-in-means findings (below table). For each outcome, I ran an ordinal logistic regression on each of four subsamples: the unreliability "control" group told nothing of Aurl's unreliability, and political sophisticates (PS), attentive respondents (AR), and attentive sophisticates in that group (AS). Consistent with the difference-in-means results, none of the regressions for

leader accountability likelihood were statistically significant, with the lowest p-value being just 0.41 (columns 1-4). Also consistent with the difference-in-means results, polarization has no net effect among those told nothing of Aurl's unreliability (column 1), but does when the analysis is focused on attentive, politically sophisticated respondents (column 8). Among those respondents, being told Aurl was polarized increased their odds of assessing that an Aurl government changeover would result in Aurl defecting from their agreement (107.4%, p = 0.083). Further, the regression results suggest that attentiveness is more responsible for that result than political sophistication. The positive effect of polarization was not found among the control group's political sophisticates (column 5, p = 0.372), but was found among the control group's attentive respondents (column 7, 45.8% increased odds, p = 0.084).

Est. Impact of Polarization on Causal Logic Variables (in "Control" Group, Odds Ratios)

	Leader Accountability if Aurl Defects				Defection (Gov't Changeover)			
VARIABLES	(1) Whole Subsample	(2) PS	(3) AR	(4) AS	(5) Whole Subsample	(6) PS	(7) AR	(8) AS
Polarized	1.002	1.325	0.858	1.403	1.258	1.386	1.458*	2.074*
	(0.187)	(0.484)	(0.188)	(0.577)	(0.234)	(0.507)	(0.319)	(0.874)
Observations	374	98	271	76	374	98	271	76

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In sum, these results provide evidentiary support to only one of the two hypothesized causal pathways from polarization to unreliability. There was no evidence that knowing Aurl was polarized would decrease respondents' assessed likelihood of leader accountability if Aurl defects (Hypothesis 2b). There was evidence, however, that knowing Aurl was polarized would increase respondents' assessed likelihood of Aurl

defecting if their government changed-over (Hypothesis 3b). Particularly, that evidence was found among attentive respondents, and more strongly supported among attentive respondents who were above the 75th quartile in political sophistication.

Beyond testing Hypotheses 3a-b, these results also help interpret this study's unexpected finding for Hypothesis 2: namely, that polarization had a net positive effect on offer acceptance via its relationship with unreliability. One of the possible explanations of the net positive effect was that the expected negative, mediated effect was "washed out" by a large, positive interaction effect. In light of polarization's positive impact on respondents' defection concern in case of government handover, a negative, mediated effect of polarization via unreliability is plausible. This design cannot establish whether respondents considered the government changeover defection risk in their choice to accept or reject Aurl's offer, but it at least rules out there is no empirical connection between Aurl's polarization and respondent perception of its defection chances.

APPENDIX T

ROBUSTNESS CHECKS: EXPERIMENT 2

I conducted three main robustness checks.

Placebo Specification Change

The results in both Table 25 (coethnicity) and Table 26 (Aurl's assumed real-world traits) were robust to an alternate specification. The specification regressed the various outcomes on an interaction of polarization and unreliability rather than on treatment group dummy variables.

For coethnicity, the results still only show significant impacts in columns 1-3 and 6 (below table). Also like in the original analysis—in which groups in the polarization condition had lower means of perceived coethnicity than the unpolarized control group—the polarized treatment is associated with lower odds of perceived coethnicity in the first three columns (-32.3%, $p < 0.05$). Further, as in the original analysis, unreliability's impacts are most seen in columns 2-3 and 5. In columns 2-3, being told Aurl was unreliable is associated with a 26.9%-30.9% decrease in odds of perceived coethnicity ($p < 0.1$). This mirrors the original analysis, which showed that not only polarization mattered, but unreliability did as well, as Group 01 (unpol./unrel.) had statistically significant lower odds of perceived coethnicity relative to the control group (00, unpol./unrel. info omitted). In column 3, a statistically significant interaction effect is evident, perhaps accounting for the high significance of the coefficients by treatment group in the equivalent model in the original analysis. Finally, in column 6, only unreliability had statistically significant effects. This is true int the original analysis as well. In Table 25, only the two groups told of Aurl's unreliability had lower odds of perceived shared goals with Aurl's people (Groups 01, 11).

Est. Impact of Treatment Group Assignment on Aurl Coethnicity Perceptions

VARIABLES	(1) Index	(2) Overall	(3) Values/beliefs	(4) Appearance	(5) Language	(6) Goals
Polarized	0.677**	0.702*	0.532***	1.000	0.748#	0.807
	(0.122)	(0.138)	(0.104)	(0.191)	(0.141)	(0.159)
Unreliable	0.864	0.731*	0.691*	1.058	0.960	0.705*
	(0.153)	(0.139)	(0.132)	(0.196)	(0.177)	(0.134)
Polariz.#Unrel.	1.210	1.261	1.575*	1.018	1.157	1.089
	(0.309)	(0.347)	(0.433)	(0.274)	(0.308)	(0.301)
Observations	750	750	750	750	750	750

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The original analysis' results in Table 26 regarding Aurl's real-world characteristics are similarly though not wholly robust to the alternate specification. As in in the original analysis, statistically significant effects are few: only the polarization factor had a statistically significant increase in odds of Aurl being a former communist state (68.4%, $p < 0.01$). This is akin to the original analysis' result for the *former communist* outcome, which showed Group 10 (pol./om.) as having the only statistically significant difference in odds from Group 00 (unpol./om.) (68.4%, $p < 0.01$).[101] The one point of deviation between the two specifications is in column 10. In the original analysis, one group (Group 11: pol./unrel.) had a statistically significant increased odds of perceiving that Aurl was more likely to be in North America (98.1%, $p < 0.05$). In this robustness check, however, none of the parameters in column 10 are statistically significant. If anything, though, this difference underscores that associations of Aurl with the United States were not asymmetrically distributed across the groups.

[101] This odds ratio's value and significance is not a typo. It is the same as the previous one.

Est. Impact of Treatment Group Assignment on Aurl Coethnicity Perceptions

VARIABLES	(1) Maj. Christ.	(2) Maj. Cauc.	(3) Fmr. comm.	(4) Remind U.S.	(5) Remind Rus./U.S.	--
Polarized	1.041	0.804	1.684***	1.368	1.386	--
	(0.200)	(0.152)	(0.319)	(0.454)	(0.415)	
Unreliable	0.881	0.773	1.187	1.075	1.237	--
	(0.165)	(0.142)	(0.217)	(0.372)	(0.369)	
Polariz.#Unrel.	0.968	1.277	0.648	1.010	0.818	--
	(0.263)	(0.342)	(0.172)	(0.481)	(0.339)	
Observations	750	750	750	317	750	--
CONT'D	(6) Africa	(7) Asia	(8) Australia	(9) Europe	(10) N. Am.	(22) S. Am.
Polarized	1.836	1.329	1.229	0.907	1.036	0.298
	(0.820)	(0.440)	(1.236)	(0.194)	(0.369)	(0.238)
Unreliable	1.251	1.191	2.254	0.986	1.127	0.547
	(0.589)	(0.393)	(1.966)	(0.204)	(0.385)	(0.340)
Polariz.#Unrel.	0.872	0.845	0.194	0.852	1.697	3.246
	(0.533)	(0.388)	(0.292)	(0.258)	(0.802)	(3.476)
Observations	750	750	750	750	750	750

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

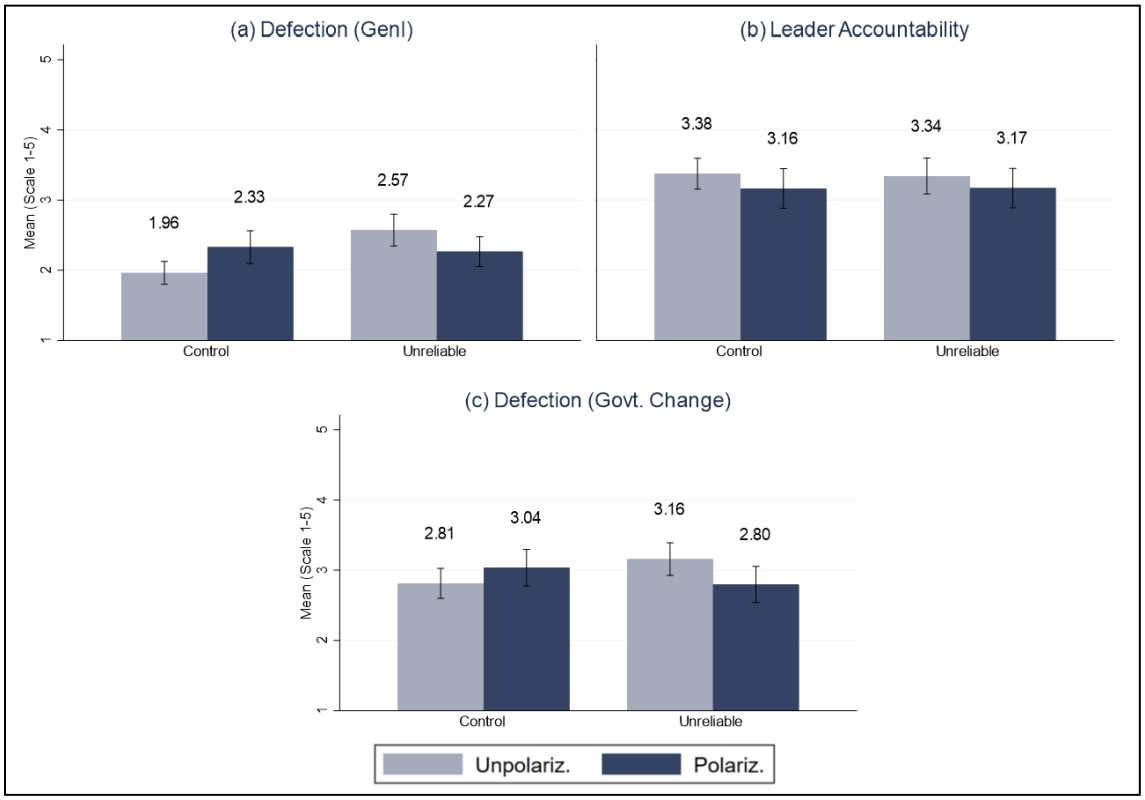
Question Framing

I also examined the impact of question framing. The main analyses' results held among the respondents asked about Aurl "keep[ing] its promise", but not among those asked about Aurl "break[ing] its promise."

Among the "keep promise" group and like in the whole-sample analysis, perceptions of general defection probability increased in the "control" condition among those told Aurl was polarized (2.33 v. 1.96, p = 0.01, below figure). Also like in the whole-sample analysis, no significant impacts to leader accountability probability or defection due to government handover registered in the "control" group (respective p-values: 0.24, 0.19). Thus, respondents exposed to the "keep promise" framing had results

for polarization akin to those of the whole-sample analysis: polarization increased general defection concern, but not strongly via the proposed two causal pathways.

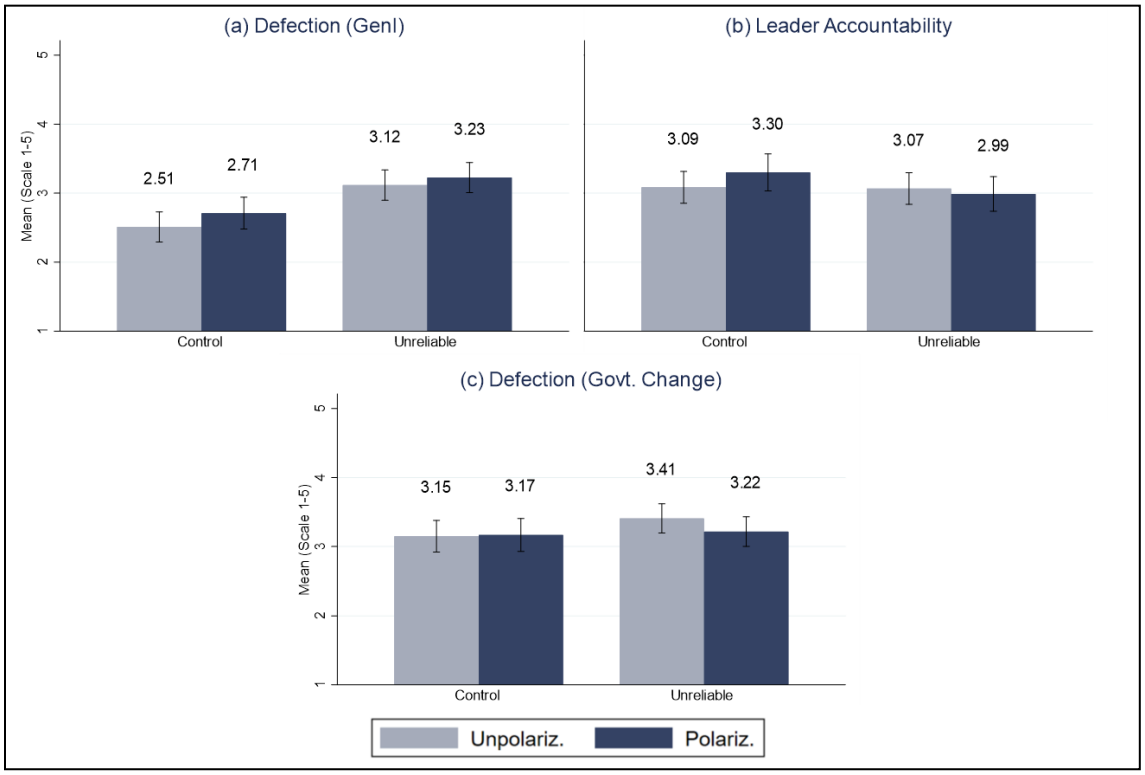
Means: Causal Logic Variables ("Keep Promise")



The “break promise” group results differed from the whole-sample and “keep promise” analysis, though. Polarization had no significant or near-significant impacts on the three causal logic outcomes in the “control” group. Polarization information moved perceived general defection probability in the expected direction, increasing it from 2.51 to 2.71, but its p-value was only 0.22 (below figure). It also increased leader accountability probability from 3.09 to 3.30 ($p = 0.23$), but besides being insignificant, this impact is also in a counter-expectation direction. Finally, polarization had no impact

in the “control” group with respect to defection probability in case of government changeover (3.15 v. 7.17).

Means: Causal Logic Variables ("Break Promise")



However, the two framing groups have statistically significant differences-in-means with regard to variables captured before they were exposed to the framing (below table). The positive framing group were somewhat more likely to be assigned to Group 01 (unpolarized/unreliable) than negative framing group ($p = 0.09$) and had almost-significant differences with regards to Group 00 and being female ($p = 0.12, 0.14$). Further, the positive framing group accepted Aurl’s offer at a rate higher than the negative framing group (7.1% difference, $p = 0.04$). This imbalance, particularly the latter one, suggests some underlying imbalance in an omitted variable between the two groups. The positive/negative framing of questions asked after the Aurl offer choice

could not have possibly impacted it. As a result, the possible framing effects come with a large asterisk.

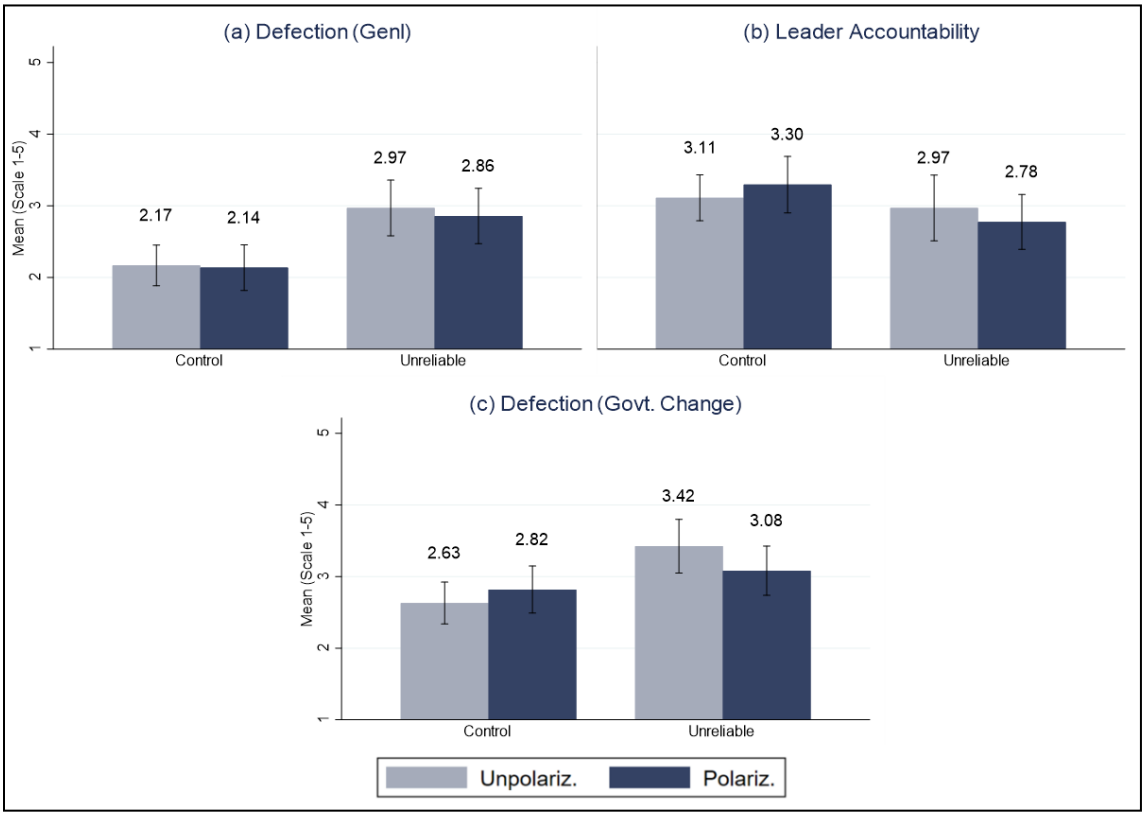
Difference-in-Means for Question Framing by Design Features

	"Keep"	"Break"	Diff	P-Stat
Accept offer	0.694	0.623	-0.071	0.04
Education	4.357	4.263	-0.094	0.37
Pol. sophist.	-0.008	0.008	0.016	0.76
Female	0.531	0.477	-0.053	0.14
00 (unpol, om)	0.300	0.249	-0.051	0.12
01 (unpol, unrel)	0.220	0.273	0.053	0.09
10 (pol, om)	0.228	0.220	-0.008	0.80
11 (pol, unrel)	0.252	0.257	0.005	0.87
Observations	750			

Political Sophistication

The main analysis' findings regarding general defection probability were not robust to a subsample analysis of political sophisticates. In the "control" group, politically sophisticated respondents registered no impact of polarization on perceived general defection chances (2.17 v. 2.14). This is in contrast to the whole-sample results, which evidenced a positive impact of Aurl polarization on perceived probability of defection in general. The other two results from the main analysis: null impacts on the two, more particular causal logic pathways—is robust to an examination of political sophistication, though. These null results suggest sophisticated respondents inferred nothing about Aurl's defection chances from its polarization, although with the significant caveat that the sample size was small (group sizes ranged from 33-54, for a total of 180 respondents).

Means: Causal Logic Outcomes (Sophisticates)



APPENDIX U

DISCUSSION: EXPERIMENT 2 “COUNTRY X” ANONYMIZATION

Adapting Myrick's study to a U.S. sample was necessary, as I had no other sample available, but came with tradeoffs. It necessitated anonymizing the United States as Country X, which in turn meant the scenario no longer provided the amount of background information inherent to the original scenario. Most of the U.K. respondents probably knew the United States is a powerful democracy and historical U.K. ally that is stronger than and a co-ethnic to the United Kingdom in some respects. Given that, an anonymous scenario in one sense simply removes a country name, but in another sense shears off a sizable amount of likely relevant scenario info, with the potential result of a substantially different treatment. I suspect anonymizing it makes it easier to find an effect of affective polarization on cooperation hesitation/unreliability assessments, as it makes the affective polarization treatment a greater proportion of the info available to them.

Further, changing the sample from U.K. to U.S. respondents risked changing respondent calculus due to the respective countries' different geo-strategic positions. Costs of current ally defection or future non-cooperation could be higher for the United Kingdom than for the more powerful United States. The impact of this seems indeterminate: U.S. respondents could be more likely to jettison cooperation with possible unreliable partners because they don't think they need them, but they also could be less dissuaded by possible future unreliability, because opponent defection may seem like it can't hurt them.

Overall, though, I concluded that a result is likely better than no result. With an anonymous scenario and U.S. sample, I judged it would be easier to find the effect I anticipate: a stronger causal effect of affective polarization of cooperation preferences when the affective polarization treatment includes information about the polarization's

policy implications. The anonymized scenario will be removed from Myrick's scenario, but not so far removed so as not to test the hypothesis.

Below is the result of the anonymization. It contains the survey wording from the original study, along with the language modifications necessary to anonymize the partner as Country X. The modifications are in italics and the table rows denote page breaks. The table omits two factual manipulation checks, as they are listed in Appendix W.

The Affective Polarization Operationalizations by Group (Myrick 2022)

Original and Modified Survey Language
We are now interested in understanding your attitudes towards the [United States/*Country X, a historically close military and economic ally of your country that we have anonymized.]* We will first provide you with some information about [American/*its*] politics and then ask you for your opinions. Please read this information carefully.
[The United States/*Country X*] has two major political parties[: the Republican Party and the Democratic Party/**no information provided about the country names**]. Studies show that [the American/*Country X's*] public and its elected officials have become increasingly polarized along party lines. In other words, [Americans/*people in Country X*] increasingly dislike members of the other political party.
Surveys from [the United States/*Country X*] show that, more than ever, [Americans/*people in Country X*]: • Oppose the idea of their child marrying someone from the other political party. • Have 'just a few' or 'no' close friends from the other political party. • ['Strongly dislike' or even 'hate' members of the other political party/*Think their political parties cannot agree on basic facts*].

These differences are reflected in [the US government. More than ever, Republican and Democratic politicians/*Country X's government. More than ever, politicians from Country X's two major political parties*]:

- Use extreme, negative language to taunt politicians of the other party.
- [Post angry or hateful posts on social media about members of the other party/ *Vote the same way as members of their own political party*].

We would like to hear your opinion on [the United States/*Country X*] and its relationship with your country. On the next page, carefully read the statements and indicate whether you agree or disagree.

[The United States/*Country X*] would come to the aid of my country in the event our security is threatened

[The United States/*Country X*] would no longer maintains its commitments to foreign countries

My country should partner with [the United States/*Country X*] would in future international agreements

[The United States/*Country X*] would will not be a reliable future partner for my country

[The United States/*Country X*] has two major political parties[: the Republican Party and the Democratic Party/**no information provided about the country names**].

How often would you say these parties agree?

APPENDIX V

ROBUSTNESS CHECKS: EXPERIMENT 2

Below are four OLS regressions per outcome variable. Each table tests to see whether the treatment's impact may have moderated by respondents' attentiveness and/or political sophistication. Attentiveness is operationalized as having encountered the Country X scenario first in the study as opposed to after the Aurl/Zerm scenario, and political sophistication is proxied by majoring or being interested in majoring in politics-related disciplines.

None of the four outcome variables show any sensitivity to the updated treatment wording overall or in any of the three subsamples. The most significant treatment impact was in Table "Maintains commitments" on attentiveness (p = 0.320).

Estimated Impact of Treatment on "Would Come to Aid"

	Whole Sample	X Scen. First	Major	X First & Major
Treatment	-0.027	-0.004	0.025	0.080
	(0.168)	(0.02)	(0.13)	(0.30)
Constant	3.613	3.640	3.542	3.420
	(0.117)**	(22.78)**	(26.39)**	(17.24)**
R^2	0.00	0.00	0.00	0.00
N	300	160	229	108

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

Estimated Impact of Treatment on "Maintains Commitm."

	Whole Sample	X Scen. First	Major	X First & Major
Treatment	0.045	0.203	0.033	-0.161
	(0.155)	(1.00)	(0.20)	(0.61)
Constant	3.852	3.837	3.814	3.920
	(0.108)**	(27.71)**	(32.18)**	(20.37)**
R^2	0.00	0.01	0.00	0.00
N	300	160	229	108

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

Estimated Impact of Treatment on "Should Partner"

	Whole Sample	X Scen. First	Major	X First & Major
Treatment	-0.089	-0.152	-0.068	-0.058
	(0.168)	(0.63)	(0.36)	(0.22)
Constant	3.406	3.395	3.347	3.420
	(0.117)**	(20.74)**	(25.24)**	(17.71)**
R^2	0.00	0.00	0.00	0.00
N	300	160	229	108

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

Estimated Impact of Treatment on "Will be Reliable"

	Whole Sample	X Scen. First	Major	X First & Major
Treatment	0.064	-0.026	-0.001	0.059
	(0.172)	(0.11)	(0.00)	(0.20)
Constant	3.232	3.256	3.271	3.320
	(0.120)**	(20.24)**	(23.73)**	(15.52)**
R^2	0.00	0.00	0.00	0.00
N	300	160	229	108

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

APPENDIX W

SURVEY: EXPERIMENT 1

Start of Block: Consent

Dear respondent,

I am a Ph.D. candidate under the direction of Professor Timothy Peterson in the School of Politics and Global Studies at Arizona State University. I am conducting a research study to explore the public's policy opinions.

I am inviting your participation, which will consist of an online survey. The survey will take approximately 6 minutes. The survey presents a political scenario and asks a series of questions.

To enable compensation and ensure respondents meet the U.S. location criteria, I will capture your CloudResearch ID and IP address. I will delete both upon completion of survey fielding. You have the right to not answer any question and to stop participation at any time.

If you choose to not participate or to withdraw from the study at any time, there will be no penalty. For completing this study, you will receive $1.06. Y**ou must be 18 years of age or older and be a U.S. resident in the United States to participate in the study**. Your participation in this study has the potential to improve the understanding political scientists have of the public's policy opinions. There are no foreseeable risks to your participation.

Grounds for disqualification from the study are as follows: (1) use of a Virtual Private Network/Server or proxy to mask one's location; (2) having a non-US IP address; (3) failure to correctly answer two questions that are exceptionally easy for a human but exceptionally difficult for a bot; (4) failure to correctly answer a question that all human respondents will be able to answer correctly; and/or (5) exceptionally poor performance on Qualtrics' fraud identification metrics [e.g., reCAPTCHA, RelevantID, duplicate entries, etc.].

If you have any questions concerning the research study, please contact the research team at: mjcantre@asu.edu or tmpete15@asu.edu. If you have any questions about your rights as a subject/participant in this research, or if you feel you have been placed at risk, you can contact the Chair of the Human Subjects Institutional Review Board, through the ASU Office of Research Integrity and Assurance, at (480) 965-6788.

If you agree to participate, please select "I agree" below and continue to the study.

Sincerely,

Michal J. Cantrell

School of Politics and Global Studies
Arizona State University

☐ I agree

WARNING!

This survey uses a protocol to check that you are responding from inside the U.S. and not using a Virtual Private Server (VPS), Virtual Private Network (VPN), or proxy to hide your country.

In order to take this survey, please turn off your VPS/VPN/proxy if you are using one, and also any ad blocking application. Then, refresh the page. Failure to do this will prevent you from completing the study.

JS

NOTE: THIS IS THE "TRAP" QUESTION FOR BOTS.
Approximately how long will it take for you to complete this survey?

- ○ 5 minutes or fewer
- ○ 6-15 minutes
- ○ 16-30 minutes
- ○ 31-60 minutes
- ○ Greater than 60 minutes

End of Block: Consent

Start of Block: 2b. VPN Use Warning

Our system has detected that you are using a Virtual Private Server (VPS), Virtual Private Network (VPN), or proxy to mask your country location.

Because of this, we cannot let you participate in this study. If you are located in the U.S.,

please turn off your VPN/VPS the next time you participate in a survey, as we requested in the warning message at the beginning. If you are outside the U.S., we apologize, but this study is directed towards U.S. participants only.

Thank you for your interest in our study.

If you have received this message in error, please contact the point of contact for this survey and enter your CloudResearch ID AND the answer to "eighteen minus ten" into the box below.

_____

End of Block: 2b. VPN Use Warning

Start of Block: 2c. Outside of US Warning

Our system has detected that you are attempting to take this survey from a location outside of the U.S. Unfortunately, this study is directed only towards participants in the U.S. and we cannot accept responses from those in other countries (as per our IRB protocol).

Thank you for your interest in our study.

If you have received this message in error, please contact the point of contact for this survey and enter your CloudResearch ID AND the answer to "eighteen minus ten" into the box below.

_____

End of Block: 2c. Outside of US Warning

Start of Block: 2e. Unresolved Location IPs

For some reason we were still unable to verify your country location. We ask you to please assist us in getting this protocol correct. Please enter your CloudResearch ID below and contact the point of contact for this survey to report the problem

Once you click "Next", you will be taken to the survey (and are certifying that you are taking the survey from the U.S. and not using a VPS). We will check locations manually for those who reach this point and we will contact you if this check identifies you as violating those requirements.

_____

End of Block: 2e. Unresolved Location IPs

Start of Block: 2d. Inauthentic Response Warning

You have been identified this as a probable inauthentic response due to duplicate entries, poor reCAPTCHA performance, other factors identified by RelevantID, and/or displaying inhuman abilities.

Because of this, we cannot let you participate in this study.

If you have received this message in error, please contact the point of contact for this survey and enter your CloudResearch ID AND the answer to "eighteen minus ten" into the box below.

_____

End of Block: 2d. Inauthentic Response Warning

Start of Block: 1. PolInt

Welcome to the study!

First, some introductory questions:

How often do you pay attention to what's going on in government and politics?

- ○ Always
- ○ Most of the time
- ○ About half the time
- ○ Some of the time
- ○ Never

Some people don't pay much attention to political campaigns. How about you? Would you say that you have been very much interested, somewhat interested or not much interested in the political campaigns so far this year?

- ○ Very much interested
- ○ Somewhat interested
- ○ Not much interested

Some people don't pay much attention to foreign policy. How about you? Would you say that you are very much interested, somewhat interested or not much interested in foreign policy?

- ○ Very much interested
- ○ Somewhat interested
- ○ Not much interested

End of Block: 1. PolInt

Start of Block: 2, PolKno

Thank you. These are the last of the introductory questions.

Instructions: We are interested in the guesses people make when they do not know the answer to a question.

We will ask you several questions.

Some may be easy, but others are meant to be so difficult that you will have to guess.

- ○ I promise to try my best without looking up any answers
- ○ I do not want to make that promise

Here is the first question. It is an example of a difficult one, so you likely will have to guess.

In what year did the Supreme Court of the United States decide Geer v. Connecticut?

Type the year.

_____

Display This Question:

If Here is the first question. It is an example of a difficult one, so you likely will have to guess.In what year did the Supreme Court of the United States decide Geer v. Connecticut?Type the year. Text Response Is Equal to 1896

You are right!

Did you look up the answer to that question, or did you already know it yourself?

- ○ I looked it up
- ○ I already knew it

Do you happen to know which party currently has the most members in the U.S. House of Representatives in Washington?

- ○ Democrats
- ○ Republicans

Do you happen to know who the current Prime Minister of the United Kingdom is?

- ○ Rishi Sunak
- ○ Boris Johnson
- ○ Theresa May
- ○ David Cameron
- ○ Nicola Sturgeon

Are China and Taiwan allies?

- ○ Yes
- ○ No

How much of a majority is required for the U.S. Senate and House of Representatives to override a presidential veto?

- ○ The U.S. Senate and House cannot override a presidential veto
- ○ 1/2
- ○ 3/5
- ○ 2/3
- ○ 3/4

What five countries are permanent members of the United Nations (U.N.) Security Council?

- ☐ United States
- ☐ United Kingdom
- ☐ Australia
- ☐ Canada
- ☐ Japan
- ☐ China
- ☐ Germany
- ☐ France
- ☐ Russia
- ☐ India

End of Block: 2, PolKno

Start of Block: B. Instructions

Good job.

Next is the main part of the study.

You will be asked to read a political scenario and provide your opinions on it.

Please read it carefully. Some parts of it may strike you as important; other parts may seem unimportant.

End of Block: B. Instructions

Start of Block: 1. Scenario (2x2) + DV

Display This Question:

If treatmentGroup = 00

Or treatmentGroup = 10

You are a citizen of a given country. You follow the news about a **major rivalry** between two other nearby countries: **${e://Field/country1}** and **${e://Field/country2}**.

- ${e://Field/country1} and ${e://Field/country2} possess **roughly equal strength**, and **both are stronger** than your country.

- Your country is somewhat **friendlier with ${e://Field/country1}** than ${e://Field/country2}.

- ${e://Field/country1} is **${e://Field/polField1} very politically divided**, with **${e://Field/polField3} disagreements** over policy and **${e://Field/polField2}** between opposing sides.

Display This Question:

If treatmentGroup = 01

Or treatmentGroup = 11

You are a citizen of a given country. You follow the news about a **major rivalry** between two other nearby countries: **${e://Field/country1}** and **${e://Field/country2}**.

- ${e://Field/country1} and ${e://Field/country2} possess **roughly equal strength**, and **both are stronger** than your country.

- Your country is somewhat **friendlier with ${e://Field/country1}** than ${e://Field/country2}.

- ${e://Field/country1} is **${e://Field/polField1} very politically divided**, with **${e://Field/polField3} disagreements** over policy and **${e://Field/polField2}** between opposing sides.

- ${e://Field/country1} **sometimes does not follow through** on its international promises.

${e://Field/country2} invades a country near you that was increasingly friendly with ${e://Field/country1}. It threatens that other countries friendly with ${e://Field/country1} may be next unless they demonstrate that they are not threats to ${e://Field/country2}.

- ${e://Field/country1} offers to immediately station powerful defense systems in your country to protect you from ${e://Field/country2} as long as necessary.

- The systems **would certainly repel** an attack by ${e://Field/country2}.

- **${e://Field/country1} says** it will **not withdraw** the systems until your country is out of danger.

- **${e://Field/country2} says** your country's acceptance of ${e://Field/country1}'s offer would demonstrate that your country is **a threat** to ${e://Field/country2}. **It also says** your country's rejection of ${e://Field/country1}'s offer would demonstrate that your country is **not a threat** to ${e://Field/country2}.

Which policy do you prefer your country follow?

- ○ **Accept** ${e://Field/country1}'s offer to station powerful defense systems in your country to protect you from ${e://Field/country2} as long as necessary
- ○ **Reject** ${e://Field/country1}'s offer to station powerful defense systems in your country to protect you from ${e://Field/country2} as long as necessary

---

Note: the "next" button will appear after thirty seconds, to ensure respondents have adequate time to read the scenario.

End of Block: 1. Scenario (2x2) + DV

Start of Block: 2. Scenario (2x2) CausalLogic

Thank you. Four follow-up questions:

If your country rejects ${e://Field/country1}'s offer, how likely do you think ${e://Field/country2} is to invade your country?

- ○ Unlikely
- ○ Somewhat unlikely
- ○ Neither likely nor unlikely
- ○ Somewhat likely
- ○ Likely

How likely is ${e://Field/country1} to ${e://Field/keepBreak1} its promise to keep its defense systems in your country as long as necessary?

- Unlikely
- Somewhat unlikely
- Neither likely nor unlikely
- Somewhat likely
- Likely

If ${e://Field/country1}'s leader ${e://Field/keepBreak2} their promise to your country, how likely are the leader's supporters to try to hold the leader accountable?

- Unlikely
- Somewhat unlikely
- Neither likely nor unlikely
- Somewhat likely
- Likely

If ${e://Field/country1}'s leader ${e://Field/keepBreak2} their promise to your country, how likely are the leader's opponents to try to hold the leader accountable?

- Unlikely
- Somewhat unlikely
- Neither likely nor unlikely
- Somewhat likely
- Likely

Final section of the scenario:

Your country chooses to accept ${e://Field/country1}'s offer.

- Just one week later, the defense systems are **stationed and operational** in your country

- **${e://Field/country2} says** the defense systems demonstrate that your country **is a threat** to ${e://Field/country2}

After some time passes, **${e://Field/country1}'s government changes hands** from its current administration to the opposing side. **Everything else in the scenario remains the same.**

In this scenario, how likely is ${e://Field/country1} to ${e://Field/keepBreak1} its promise to keep its defense systems in your country as long as necessary?

- Unlikely
- Somewhat unlikely
- Neither likely nor unlikely
- Somewhat likely
- Likely

End of Block: 2. Scenario (2x2) CausalLogic

Start of Block: 3. Manipulation Checks

Only three short sections of questions left.

The first section is about information in the scenario.

What did the scenario say about the level of political division/conflict within ${e://Field/country1}?

${e://Field/country1}:

- ○ Is very politically divided, with major disagreement over policies and mistrust/disrespect between opposing sides
- ○ Is not very politically divided, with only minor disagreement over policies and trust/respect between opposing sides
- ○ It did not say anything about it

What did the scenario say about how well ${e://Field/country1} follows through on its international promises?

${e://Field/country1}:

- ○ Sometimes does not follow through on its international promises
- ○ Almost always follows through on its international promises
- ○ It did not say anything about it

End of Block: 3. Manipulation Checks

Start of Block: 4. Scenario (2x2) Placebo Qs

NOTE: the following section was the same for Country 1 and Country 2, with Country 2 being omitted here for brevity. The only difference is that the Country 2 section was prefaced with "Final section: same questions, but about ${e://Field/country2}."

The last two sections are your opinions about **${e://Field/country1}** and **${e://Field/country2}**.

First, ${e://Field/country1}.

How likely are people from **${e://Field/country1}** to be people like you overall?

- ○ Unlikely
- ○ Somewhat unlikely
- ○ Neither likely nor unlikely
- ○ Somewhat likely
- ○ Likely

How likely are people from **${e://Field/country1}** to be people like you in terms of the following traits?

	Unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Likely
VALUES & BELIEFS (what people think about religion and politics | what people think is right or wrong)	○	○	○	○	○
APPEARANCE (physical characteristics | style of clothing/hair)	○	○	○	○	○
LANGUAGE (use language similar to or the same as one's own)	○	○	○	○	○
GOALS (what people want in life | what things people find meaningful)	○	○	○	○	○

Which, if any, continent does **${e://Field/country1}** seem most likely to be located in?

- ○ North America
- ○ South America
- ○ Asia
- ○ Europe
- ○ Africa
- ○ Australia
- ○ No continent seems a more likely location than the others

How likely is **${e://Field/country1}** to be:

	Unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Likely
Majority Christian	○	○	○	○	○
Majority Caucasian	○	○	○	○	○
A former Communist country	○	○	○	○	○

Does **${e://Field/country1}** remind you of any countries?

- ○ Yes
- ○ No

Display This Question:

If Does ${e://Field/country1} remind you of any countries? = Yes

Which country does **${e://Field/country1}** most remind you of?

_____

Thank you for your participation in this study!

!!! --- DEBRIEFING STATEMENT --- !!!

- **The scenario presented in this study was fictional.**

- **The policy focus of this study was whether people are less likely to prefer cooperation with states that are politically polarized, and if so, why (e.g., concern about lack of accountability, concern of policy change due to government handover, etc.). We did not provide that detailed purpose at the beginning of the study because knowing that was our purpose could have influenced your answers.**

---> Thank you again, and please hit the "next" arrow to be redirected to CloudResearch.

End of Block: 4. Scenario (2x2) Placebo Qs

APPENDIX X

SURVEY: EXPERIMENT 2

Below is the portion of the pilot study that pertained to Country X, as well as the collected demographic data that served as a distraction task. Sometimes the Country X scenario preceded the pilot Aurl/Zerm scenario, and sometimes it came after. The distraction task was always between the two scenarios. After the final scenario was an omitted and unrelated series of questions for a separate study, querying data privacy preferences, partisanship-related variables, and trust in experts.

Start of Block: Consent

Dear respondent,

I am a graduate student under the direction of Professor Timothy Peterson in the School of Politics and Global Studies at Arizona State University. I am conducting a research study to explore foreign policy opinions.

I am inviting your participation, which will consist of an online survey. The maximum duration of the study is one hour, but you can expect to complete it within 10-20 minutes. The survey presents two political scenarios and asks a series of questions. Your responses in the survey will be anonymous, meaning the data I collect cannot be traced back to you. You have the right to not answer any question and to stop participation at any time.

If you choose to not participate or to withdraw from the study at any time, there will be no penalty. By completing this study, you will receive one (1) required research credit for your class. Credit will be granted on Sona after completion of the study and your professor will be notified at the end of your class session of all credits students have earned. If you do not wish to participate in this study, you can complete an alternative assignment to fulfill the credit(s) required for the research component of your class. **You must be 18 years of age or older to participate in the study**. Your participation in this study has the potential to improve the understanding political scientists have of foreign policy opinions. There are no foreseeable risks to your participation.

If you have any questions concerning the research study, please contact the research team at: mjcantre@asu.edu or tmpete15@asu.edu. If you have any questions about your rights as a subject/participant in this research, or if you feel you have been placed at risk, you can contact the Chair of the Human Subjects Institutional Review Board, through the ASU Office of Research Integrity and Assurance, at (480) 965-6788. If you agree to participate, please select "I agree" below and continue to the study.

Sincerely,

Michal J. Cantrell
School of Politics and Global Studies
Arizona State University

☐ I agree

End of Block: Consent

Start of Block: B. Instructions

Welcome to the study!

In this study, you will be asked to read two scenarios.

On the next page is the first scenario. Please read it carefully. Some parts of it may strike you as important; other parts may seem unimportant. After describing the scenario, we will ask your opinions.

End of Block: B. Instructions

Start of Block: DISTRACTION: Demographics

Thank you!

We will now ask some quick demographic questions before presenting the final scenario.

What is your student status: international or U.S. student?

- ○ International student
- ○ U.S. student

What is your gender?

- ○ Male
- ○ Female
- ○ Non-binary / third gender
- ○ Prefer not to say

Below are a list of eight race/ethnicity categories. Please check as many as apply to you.

- ☐ White
- ☐ Black or African American
- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Native Hawaiian or Pacific Islander
- ☐ Middle Eastern or North African
- ☐ Hispanic
- ☐ Prefer not to say / not applicable

Are you majoring in (or thinking about majoring in) political science, politics and the economy, or global studies?

- ○ No
- ○ Yes

End of Block: DISTRACTION: Demographics

Start of Block: Transition to final scenario

Thank you! We are now nearing the end of the survey. What follows is the second and final scenario.

Please read it carefully. Some parts of it may strike you as important; other parts may seem unimportant. After describing the scenario, we will ask your opinions.

End of Block: Transition to final scenario

Start of Block: i. Scenario (1x2): Intro

We are interested in understanding **your attitudes toward Country X, a historically close military and economic ally of your country that we have anonymized**. We will first provide you with some information about its politics and then ask you for your opinions.

Please read this information carefully.

Country X has two major political parties.

Studies show that Country X's public and its elected officials have become **increasingly polarized along party lines.**

In other words, **people in Country X increasingly dislike members of the other political party.**

End of Block: i. Scenario (1x2): Intro

Start of Block: iia. Scenario (1x2): No Gridlock

Surveys from Country X show that, more than ever, people in Country X:

• Oppose the idea of their child marrying someone from the other political party.

• Have "just a few" or "no" close friends from the other political party.

• "Strongly dislike" or even "hate" members of the other political party.

Based on what you read, which of the following is correct?

- ○ People in Country X "strongly dislike" or even "hate" members of the other political party.
- ○ People in Country X oppose the idea of their child marrying someone from the other political party.
- ○ Neither of these statements is correct.
- ○ Both of these statements are correct.

These differences are reflected in Country X's government. More than ever, politicians from Country X's two major political parties:

• Use extreme, negative language to taunt politicians of the other party.

• Post angry or hateful posts on social media about members of the other party.

Based on what you read, which of the following is correct?

- ○ Politicians from Country X use extreme, negative language to taunt politicians of the other party.
- ○ Politicians from Country X post angry or hateful posts on social media about members of the other party.
- ○ Neither of these statements is correct
- ○ Both of these statements are correct

End of Block: iia. Scenario (1x2): No Gridlock

Start of Block: iib. Scenario (1x2): Gridlock

Surveys from Country X show that, more than ever, people in Country X:

• Oppose the idea of their child marrying someone from the other political party.

• Have "just a few" or "no" close friends from the other political party.

• Think their political parties cannot agree on basic facts.

Based on what you read, which of the following is correct?

- ○ People from Country X think their political parties cannot agree on basic facts.
- ○ People from Country X oppose the idea of their child marrying someone from the other political party.
- ○ Neither of these statements is correct.
- ○ Both of these statements are correct.

These differences are reflected in Country X's government. More than ever, politicians from Country X's two major political parties:

• Use extreme, negative language to taunt politicians of the other party.

• Vote the same way as members of their own political party.

Based on what you read, which of the following is correct?

- ○ Politicians from Country X use extreme, negative language to taunt politicians of the other party.
- ○ Politicians from Country X vote the same way as members of their own political party.
- ○ Neither of these statements is correct
- ○ Both of these statements are correct

End of Block: iib. Scenario (1x2): Gridlock

Start of Block: iii. Scenario (1x2): Intro to Qs

We would like to hear **your opinion on Country X and its relationship with your country.** On the next page, carefully read the statements and indicate whether you agree or disagree.

End of Block: iii. Scenario (1x2): Intro to Qs

Start of Block: iv. Scenario (1x2): Main Qs

X→

Country X would come to the aid of my country in the event our security is threatened

- ○ Strongly disagree
- ○ Disagree
- ○ Somewhat disagree
- ○ Neither agree nor disagree
- ○ Somewhat agree
- ○ Agree
- ○ Strongly agree

X→

Country X no longer maintains its commitments to foreign countries

- ○ Strongly disagree
- ○ Disagree
- ○ Somewhat disagree
- ○ Neither agree nor disagree
- ○ Somewhat agree
- ○ Agree
- ○ Strongly agree

X→

My country should partner with Country X in future international agreements

- ○ Strongly disagree
- ○ Disagree
- ○ Somewhat disagree
- ○ Neither agree nor disagree
- ○ Somewhat agree
- ○ Agree
- ○ Strongly agree

X→

Country X will <u>not</u> be a reliable future partner for my country

- ○ Strongly disagree
- ○ Disagree
- ○ Somewhat disagree
- ○ Neither agree nor disagree
- ○ Somewhat agree
- ○ Agree
- ○ Strongly agree

End of Block: iv. Scenario (1x2): Main Qs

Start of Block: v. Scenario (1x2): Manipulation Check

Country X has two major political parties. How often would you say these parties <u>agree</u>?

- ○ Almost Always
- ○ Sometimes
- ○ Rarely
- ○ Almost Never
- ○ I don't know

End of Block: v. Scenario (1x2): Manipulation Check

Start of Block: via. Scenario (1x2) Placebo Qs (PosFrame)

NOTE: In this section I randomized the question wording. Some respondents saw the questions framed positively ("people like you"), and others saw the questions framed negatively ("people unlike you").

How likely are people from Country X to be people like you overall?

- ○ Extremely unlikely
- ○ Somewhat unlikely
- ○ Neither likely nor unlikely
- ○ Somewhat likely
- ○ Extremely likely

How likely are people from Country X to be people like you in terms of the following traits?

	Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
VALUES & BELIEFS (what people think about religion and politics | what people think is right or wrong)	○	○	○	○	○
APPEARANCE (physical characteristics | style of clothing/hair)	○	○	○	○	○
LANGUAGE(use language similar to or the same as one's own)	○	○	○	○	○
GOALS(what people want in life | what things people find meaningful)	○	○	○	○	○

How likely is Country X to be:

	Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
Majority Christian	○	○	○	○	○
Majority Caucasian	○	○	○	○	○
Located in Europe	○	○	○	○	○
Located in Asia	○	○	○	○	○
Located in North America	○	○	○	○	○
Located in South America	○	○	○	○	○
Located in Africa	○	○	○	○	○
A former communist state	○	○	○	○	○

Does Country X remind you of any countries?

- ○ No
- ○ Yes

End of Block: via. Scenario (1x2) Placebo Qs (PosFrame)

End of Block: vib. Scenario (1x2) Placebo Qs (NegFrame)

Start of Block: vii. Scenario (1x2) Placebo Follow-Up

Display This Question:
If Does Country X remind you of any countries? = Yes
Or Does Country X remind you of any countries? = Yes

Which country does Country X most remind you of?

_____

End of Block: vii. Scenario (1x2) Placebo Follow-Up