Bias and Affective Polarization

Daniel F. Stone
Bowdoin College

September 2016*

Abstract

I propose a model of affective polarization ("that both Republicans and Democrats increasingly dislike, even loathe, their opponents," Iyengar et al, 2012). In the model, two agents repeatedly choose actions based on private interests, the social good, their own "character" (willingness to trade private for social gains) and beliefs about the other agent's character. Each agent could represent a political party, or the model could apply to other settings, such as spouses or business partners. Each agent Bayesian updates beliefs about the other's character, and dislikes the other more when its character is perceived as more self-serving. I characterize the dynamic and long-run effects of three biases: a prior bias against the other agent's character, the false consensus bias, and limited strategic thinking. Prior bias against the opponent remains constant or dissipates over time, and actions do not diverge. By contrast, the other two biases, which are not directly related to character, cause actions to become more extreme over time and repeatedly be "worse" than expected, causing affective polarization—even when both players are arbitrarily "good" (unselfish). For some parameter values, long-run affective polarization is unbounded, despite Bayesian updating. The results imply that affective polarization can be caused by cognitive bias, and that subtlety and unawareness of bias are key forces driving greater severity of this type of polarization.

Keywords: affective polarization, partyism, polarization, disagreement, dislike, over-precision, unawareness, media bias

JEL codes: D72; D83

*I thank Steven J. Miller, Dan Wood, Gaurav Sood, Dan Kahan, Roland Bénabou, Andrei Shleifer, and participants at seminars at the University of Western Ontario, Wake Forest University and Southern Methodist University, in particular Saltuk Ozerturk, Tim Salmon, Bo Chen, James Lake, Al Slivinski, Greg Pavlov, and Charles Zheng, for helpful comments and discussion. This paper was written while I visited the University of Virginia's Department of Economics; I am grateful for their hospitality. Email: dstone@bowdoin.edu.

"Most quarrels amplify a misunderstanding."

- André Gide

1 Introduction

Has the US become more politically polarized in the last several decades? Yes, definitely, with respect to legislative voting; see Figure 1. No, not necessarily, in that political scientists have debated ideological polarization of the general public for years and have yet to reach consensus on this topic.[1]

The common intuition that the US general public has indeed become more polarized has recently received stronger empirical support, however. A new literature has emerged documenting relatively strong and unambiguous evidence of mass *affective* polarization—that rank and file partisans have grown to *dislike* members of the out-party more over time (independent of whether their ideologies or true policy preferences have diverged).[2] Affective polarization has likely contributed to further exacerbating party-line voting and political gridlock, as well as other social and political problems.[3]

The political science literature typically attributes affective polarization to growth in "social distance" between the parties (see, e.g., Iyengar and Westwood, 2015). This idea, put briefly, is that it is human nature to automatically dislike others who are different from ourselves, and that there has been a perceived growth in differences between the parties over time. Some papers in this literature stress the importance of strengthened partisan identity and social identity theory (Mason, 2015). Another prominent theory is increasingly vitriolic partisanship in the media environment and political advertising (Lelkes, Sood, and Iyengar, 2015). A related literature from political psychology focuses on the evolutionary adaptiveness

[1] See Abramowitz and Saunders (2008) and Fiorina, Abrams, and Pope (2008) for competing arguments, and Hill and Tausanovitch (forthcoming) for more recent analysis and discussion of the lack of consensus.

[2] Iyengar, Sood, and Lelkes (2012) first coined the term affective polarization. Important later papers include Rogowski and Sutherland (2015) and Mason (2015). The trend seems to continue; when announcing the suspension of his presidential campaign, Marco Rubio said, "[Modern politics is] going to leave us as a nation where people literally hate each other because they have different political opinions." (See http://www.latimes.com/politics/la-pol-prez-marco-rubio-speech-transcript-20160315-story.html.) While the US context is the focus of this paper, partisan polarization is of course not unique to the US; the analysis of this paper may also apply to polarization in other contexts. For example, the recent debate over Brexit became so heated as to cause numerous violent incidents.

[3] See, e.g., Hetherington and Rudolph (2015) for discussion of how affective polarization could exacerbate gridlock. See Mann and Ornstein (2013) and Barber and McCarty (2015) for detailed discussion of potentially harmful welfare effects of partisan voting and gridlock.

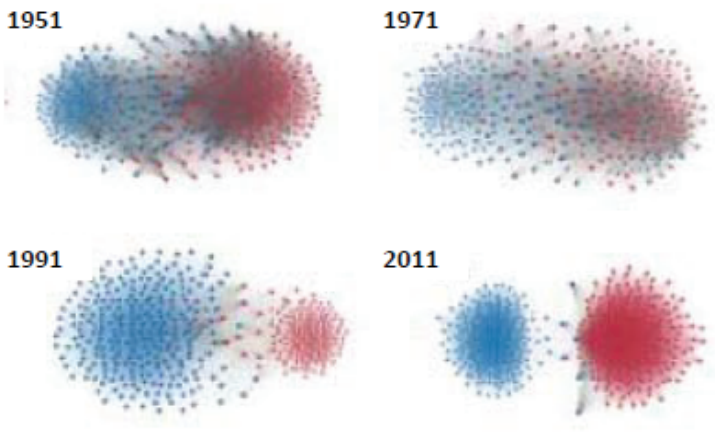

Figure 1: US House of Representatives voting network graphs, adapted from Andris, Lee, Hamilton, Martino, Gunning, and Selden (2015). Red nodes are Republicans, blue nodes Democrats, and the lines connecting nodes and positions of nodes indicate voting similarity.

of tribalism and motivated reasoning (Haidt, 2012).

But the phenomenon of escalation of extremism of actions and hostility building on one another is not limited to partisan settings. This pattern occurs all too often in a variety of contexts with repeated bilateral interactions, such as spouses, friends, and business partners. The fact that these settings do not involve opposing social groups, media exposure, or motivation to believe in the opposition's inferiority, implies that these are not crucial factors underlying this behavior.[4]

In this paper, I study a novel (but intuitive) explanation for affective polarization across contexts: cognitive bias.[5] Recent research in psychology, neuroscience and even philosophy, argues that interpersonal feelings, and emotions more broadly, are not in fact "non-cognitive"—rather, they reflect information, judgments, and beliefs, conscious or otherwise.[6] In the context

[4]Moreover, the within-party conflicts that have occurred in the current (2016) presidential election campaign, in both parties, also point against the partisan identity theory.

[5]In addition to the literatures referred to above, the social psychology literature, perhaps due to its focus on social identity, appears to largely neglect the subject of within-group conflict and the effects of cognitive bias on interpersonal relationship problems and (unjustified) dislike. I am not aware of work showing the link between the biases studied in psychology most directly related to inference about personal characteristics, the fundamental attribution error and correspondence bias, and inter-party hostility or escalation of conflict in relationships in general. See, for example, Epley (2014) for an informal overview of related psychology research on misperceptions of thoughts of other people, including some discussion of how these misperceptions can lead to conflict. The specialized literature on non-group related hate seems to also largely neglect the role of bias (Rempel and Sutherland, 2016).

[6]Recent work in psychology and neuroscience supports the view that emotional and cognitive processes are not as distinct as previously believed. Haidt (2012) refers to "a prevalent but useless dichotomy between cognition and emotion;" Pessoa (2008) says "parcelling the brain into cognitive and affective regions is inher-

of partisan politics, hostility toward the out-party is caused in large part by beliefs about why the out-party *should* be disliked. Moreover, both common sense and research imply that such beliefs in this context are likely biased—that, more generally, when large groups of people perceive other large groups to be inferior and contemptible, these perceptions are likely off-base (Graham, Nosek, and Haidt, 2012). Even in settings involving just two individuals, the hostility is often based on misunderstanding and skewed beliefs.[7]

Connecting the dots—unjustified dislike is based on skewed beliefs, cognitive bias causes skewed beliefs, thus cognitive bias likely causes dislike—is fairly straightforward. Still, it is unclear which bias or biases have such an effect, to what extent, and why. As alluded to above, there is little existing work studying this issue, either empirically or theoretically. In this paper, I propose and study a model to obtain a more precise and deeper theoretical understanding of these relationships. In particular, I examine the effects of three distinct biases to see which yield outcomes consistent with the two key empirical facts mentioned above for the case of the major parties in the US: 1) increasing extremism of actions over time; 2) increasing dislike of the opposition. Other key questions addressed by the analysis are: how often does affective polarization occur—can it occur even for "good" players—and what is the magnitude of this polarization.

In the model, there are two agents, L and R (left and right). I interpret the model as representing the political context, but the model could apply to other ongoing bilateral relationships. The agents have different interests: L directly benefits from higher levels of a variable, x, and R benefits from lower x. A broad example of x would be the "size" of government; while not all leftists directly benefit from a larger government, on average they are likely to do so, and the same is true of rightists for smaller government. Both agents realize that the level of x that is best for society is not the same as the level that is best for them personally, and agents are heterogeneous in the extent to which they are willing to trade off private for social gains.[8] The agents also may have different "tastes" for socially optimal

ently problematic, and ultimately untenable." From philosophy, see Nussbaum (2003), who writes, "Emotions include in their content judgments that can be true or false."

[7]This is again intuitive, and evidenced by the fact that advice from outsiders, such as marriage counselors or mediators, often helps resolve the issues.

[8]Bernheim and Kartik (2014) model political competition involving a very similar character trait, which they call public spirit. Altruism is a more standard term for a closely related idea, but is not quite the same thing since altruism usually refers to willingness to trade off private gains for gains for another individual(s), and not for society overall.

x: in this case, some truly think society is better with larger x (independent of private gains from this), and some truly believe in smaller x.

Each agent i repeatedly chooses a level of x, partly based on her own "selfishness" parameter s_i, and partly on beliefs about s_{-i}, due to reciprocity or strategic motives. These choices are publicly observed with noise and thus act as signals for updating beliefs about the other agent's s. Dislike of the opposition increases (i.e., affective polarization occurs) when each player's expectation of the other's selfishness increases over time. Both agents are Bayesian updaters, so their beliefs would typically converge to truth in the absence of behavioral factors.

The three biases that I analyze, one-by-one so as to isolate their individual effects, are: prior out-group bias (over-estimation of the opposition's selfishness before the start of play), prior false consensus bias (under-estimation of differences in tastes for socially optimal x before the start of play), and limited strategic thinking (under-estimation of the opposition's strategic sophistication).[9] The out-group bias is the most directly related to the outcome of interest, out-group dislike, and so is the most natural bias to consider. The other two biases are directly related to key elements of bilateral interactions in general—differences in tastes and strategic thinking.[10] An important additional assumption is that both players are unaware of any biases.

I find that out-group bias remains constant over time, or dissipates, partially or completely, depending on parameter values. But out-group bias never builds on itself. By contrast, both of the other biases cause dislike to eventually grow over time for all parameter values, and for some, for dislike to eventually grow *for all realizations of the s parameters*. Moreover, both of these biases can cause dislike to become *arbitrarily large* for all realization of the s's, i.e., even when both agents are minimally self-serving, and despite the variance of beliefs declining to zero.

The intuition for why these biases cause affective polarization "on average" is relatively straightforward. When one's true tastes for a social outcome are consistent with one's private

[9] As discussed below, I use the level-k model of strategic thinking. Both players are (always) level-2 thinkers, which means their belief about the level of thinking of the opponent is biased downward (and never updated). As I discuss, this is a stark assumption.

[10] The false consensus bias can be seen as an extension and formalization of Haidt (2012), an especially prominent work on political hostility, which argues that political-moral values are analogous to tastes, and that under-appreciation of heterogeneity in these tastes is an important cause of partisan hostility. One could also view the analysis of the false consensus and strategic thinking biases as illustrating how biases that seem unrelated to character inferences can potentially contribute to affective polarization.

interests, acting in a way that one believes is best for society appears self-serving. E.g., if Republicans see Democrats taking a seemingly unsavory action in support of big government (say, filibustering a proposed tax cut), and under-estimate how much Democrats "truly believe in" big government, Republicans may interpret this action as self-serving and thus deserving of contempt. Similarly, when strategic sophistication of an actor is underrated, the actor may appear more self-serving than she really is.[11] The reason dislike does not increase due to out-group bias is that in this case, each player expects the other to be a bad actor. So when the other acts accordingly (or less bad), this is, at worst, in line with expectations.

The intuition for the stronger results—that dislike can increase unboundedly, even with arbitrarily "good" players, and despite precision of beliefs converging to certainty—is more subtle. When player i acts "selflessly" in period 1, she over-estimates the extent to which her opponent updates beliefs about s_i downward. Thus, player $-i$ reciprocates unexpectedly "aggressively" in period 2, causing i's beliefs about s_{-i} to be updated upward. Player $-i$ is unaware of this, causing this cycle to continue and compound. This occurs whether the initial bias and true s's are large or small. The race is then on between the increase in precision of beliefs over time, and exacerbation of bad actions due to repeated misunderstanding. The latter dominates when precision of prior beliefs about the opposition's character is sufficiently low. A similar dynamic occurs in the limited strategic thinking case.

Since the results for the false consensus and strategic thinking biases are so similar, each can be seen as a robustness check for the other, and we can look to see what the cases have in common to determine the deeper causes of extreme affective polarization. In both cases, behavior must continue over time to surprise (in a negative way), and the causes of surprise are subtle. In particular, there are three key ingredients that the two cases have in common: 1) the players have a reciprocity motive; 2) the bias is *not* directly related to dislike; 3) the players are unaware of the bias. The results thus imply that subtlety of bias, and unawareness of bias, i.e., overconfidence in one's knowledge and one's unbiasedness (closely related to the concepts of WYSIATI and overprecision), are more important than the specific nature of the bias in driving misunderstanding and thus growth of dislike. I discuss interpretation and

[11]For example, suppose two parties bargain over a pie of size one. Democrats are selfish, but not fully selfish, and wish to take 75%. Both parties make claims for the fractions of the pie they should receive, and the outcome the Democrats actually receive increases in their own claimed share and decreases in the Republicans' claimed share. Taking this strategic interaction into account, the Democrats make a claim more aggressive than their ideal, say 90%. Then, if Republicans fail to sufficiently account for the Democrats' strategic motive, they will infer that the Democrats are more selfish than they truly are.

implications further in the paper's concluding remarks.

2 Additional related literature

This paper is the first from economics, to my knowledge, to address affective polarization. The economics literatures most closely related to this paper are those on political extremism and polarization, on disagreement more generally, and on ethnic conflict. In this first category, Ortoleva and Snowberg (2015) is particularly relevant; they provide a theoretical argument, and empirical evidence, that cognitive bias (overprecision) can cause ideological extremism (but do not study inter-party hostility).[12]

Since Aumann (1976), many economists have considered persistent disagreement to be puzzling. Recent examples of papers on this topic include Sethi and Yildiz (2012), Baliga, Hanany, and Klibanof (2013) and Andreoni and Mylovanov (2012).[13] The existing papers on this topic do not address interpersonal feelings. My paper shows how feelings can contribute to exacerbation of disagreement about policy choices. Andreoni and Mylovanov (2012) is most relevant to mine in that it shows that individuals holding different models of the world can cause opinions to diverge in response to common information (heterogeneous priors on one dimension of the state of the world cause diverging responses to new information on another dimension), and provide experimental evidence supporting this point. An important substantive difference between our models is that they do not address common knowledge of disagreement, while that is actually a key cause of increasing disagreement in my model. But an important similarity is that different beliefs about one dimension (in my case, character) is the cause of increasing disagreement on the other.[14]

Glaeser (2005) models hatred towards an out-group, but the hatred is not based on Bayesian inference. Klumpp and Mialon (2013) study the effects of hate but do not explain the cause of hate. Bordalo, Coffman, Gennaioli, and Shleifer (forthcoming) is more similar to

[12]Related papers include Blomberg and Harrington (2000) and Roux and Sobel (2015). The former show that, with heterogeneous priors, correlation in political extremism and rigidity (precision) of beliefs may arise simply due to Bayesian updating. The latter show how groups may rationally be more polarized than individuals.

[13]The game theory literature on misspecified models is also related; see Esponda and Pouzo (2016) for a very recent paper and references.

[14]Indeed, Andreoni and Mylovanov (2012) also mention unawareness as an important factor causing long-run disagreement. Benoît and Dubra (2014) provide further development of a model related to Andreoni and Mylovanov (2012)'s.

my paper in that they model distorted stereotypes of other social groups as based on biased belief formation; however, their model is quite different overall as it does not study dislike. Acemoglu and Wolitzky (2014) is perhaps the most similar paper to mine. They show how cycles of conflict between groups can arise due to misperceptions ("good" actions are misperceived as "bad"), causing beliefs about the quality of the other side's character to decline. In their model, actions are binary and a spiral of conflict is a sequence of periods in which both parties play bad actions (as opposed to actions actually becoming more extreme over time). They show how cycles can end when groups eventually rationally infer the cycle likely began by mistake; my focus instead is on how small behavioral errors can cause compounding misunderstanding and greater degrees of hostility over time.

3 The model

The goal of the model is to efficiently capture the essence of how actions and beliefs evolve in a repeated bilateral interaction involving both conflicting and shared interests. The model has the structure of an infinitely repeated game (it is convenient to use game theoretic terminology although much of the analysis will be non-game theoretic). In Section 3.1, I describe the stage game set-up. In Section 3.2, I describe how players update beliefs about each other across stages, and how actions depend on beliefs in each stage. In Section 3.3, I define two types of affective polarization that may occur. I include discussion of the model assumptions throughout the section. In each subsection of Section 4, I analyze the model with a single modification of the baseline assumptions presented here, one for each of the three biases studied (the analysis of Section 4.1 subsumes the baseline case with no modifications, i.e., no biases).

3.1 Players, actions and payoffs

There are two players, L (the leftist) and R (the rightist). The stage game payoff for each player i is

$$u_i(x) = u_i^p(x) + \alpha_i u^s(x; \tau_i). \tag{1}$$

Stage subscripts are omitted for now. $x \in \mathbb{R}$ is a choice variable (chosen privately or jointly). Player i receives a direct, private payoff from x of $u_i^p(x)$, and a (subjective) social welfare payoff from x, given i's taste parameter $\tau_i \in \mathbb{R}$, of $u^s(x;\tau_i)$. The pro-sociality of i—the extent to which i is willing to trade off private for social gains—is represented by $\alpha_i > 0$. α_i is a function of other parameters, as will be defined shortly.

I use a quadratic loss function for u^s, and the functions $u_i^p(x) = x$ for $i = L$ and $u_i^p(x) = -x$ for $i = R$ for the private payoffs. The baseline assumptions are that x is chosen unilaterally (that is, $x = x_L$ for L and $x = x_R$ for R) and that $\tau_i = 0$ for both players. The baseline payoffs are thus:

$$u_R(x_R) = -x_R - \alpha_R x_R^2;$$
$$u_L(x_L) = x_L - \alpha_L x_L^2.$$

L privately benefits from x being larger, but thinks society is best off when x is as close as possible to 0. R faces an analogous trade-off, privately preferring smaller x.[15]

The pro-sociality weight, α_i, is defined as:

$$\alpha_i = (1/2)\left(\frac{1}{s_i + rE_i(s_{-i})}\right).$$

s_i is i's "selfishness" and is assumed to have a support of $(0, \infty)$. The term $rE_i(s_{-i})$ represents "reciprocal selfishness": in general, i cares less about social welfare when i expects that s_{-i} is larger, and this effect is strengthened when r (for short, "reciprocity") is larger.[16] It is natural to assume that $r \in [0, 1]$.[17] The $(1/2)$ just simplifies algebra. Thus, $\alpha_i \in (0, \infty)$, and α_i approaches zero as either s_i or s_{-i} approaches ∞, and α_i approaching ∞ requires both s_i and s_{-i} to approach zero (if $r > 0$).

Given the political application, the players can be thought of as either politicians or voters, and the actions as directly affecting policy (this is more relevant for politicians) or statements

[15]The social payoff is not just the sum of the direct payoffs to L and R because there may be other members of society outside of the model, or because each player perceives that the other would be better off acting in accordance with one's own tastes (rather than that player's perception of own preferences).

[16]The payoff for i is a direct function of $E_i(s_{-i})$, rather than s_{-i} with the expected payoff then maximized, simply for the sake of tractability.

[17]See, for example, Levine (1998) for an example of a model that incorporates preferences for reciprocity in a similar way (with regard to others' types and not just their actions). See, e.g., Batson and Powell (2003) for discussion of the (very large) broader literature on the importance of reciprocity for pro-social behavior.

of opinion.[18] x would be a choice of a broad, repeatedly debated policy. As mentioned above, a natural interpretation of x is the size of government: leftists, on average, directly benefit from larger x, since more of them are government employees, beneficiaries of redistributive policies, or if involved in politics, gain politically from such policies; rightists, on average, directly benefit from smaller x for analogous reasons. And across ideologies, people face trade-offs between doing what is best for themselves and what they feel is best for society.[19] Individuals in other bilateral relationships face similar tradeoffs.[20] In some situations agents do not have the power to determine an outcome of interest unilaterally, even in the very short-run; in section 4.3, I modify the model so that x is determined by the players jointly.

I assume that agents are not forward-looking, and in each period, each agent simply maximizes subjective expected utility for the stage game. This myopia assumption is made both to simplify the analysis and because it is likely realistic, at least for most political actions by individual citizens, but also, to a lesser extent, for politicians. Still, the lack of intertemporal payoffs implies the players do not have reputation concerns, which is a questionable aspect of the model that should be kept in mind. I discuss interpretation of payoffs further at the end of this section.

3.2 Beliefs, learning, and choices

There is common knowledge of payoff functional forms, and each agent i knows her own s_i but does not observe s_{-i}.[21] To create more interesting and realistic learning dynamics, I assume that there is a noise term added to $x_{i,t}$ in each stage game period t, $\epsilon_t^i \sim N(0, \sigma_\epsilon^2)$ (perhaps there are additional exogenous factors affecting the action each period), and that $\hat{x}_{i,t} = x_{i,t} + \epsilon_t^i$

[18] Members of Congress can add riders to bills or make pivotal votes in committees, and certainly executives (mayors, governors etc) often have the opportunity to take unilateral executive actions. Politician public statements or claims certainly can be significant and influential. Individual citizen actions typically do not affect social welfare as directly, but individuals often think their (unilateral) actions, such as campaign efforts or contributions, or statements in conversation or online, are more significant than they really are (Duffy and Tavits, 2008) or may simply wish to express their politics through their actions (Brennan and Hamlin, 1998), and in the aggregate these actions are indeed likely significant.

[19] For readers who prefer a specific interpretation of the players and actions, I would suggest that L and R be interpreted as each being a prominent Congressman/woman (from opposite parties), with x_L and x_R being public statements on the policy issue x. The private benefits of x would be political (politicians gain more support and status within their party, and hence improve career concerns, if they are able to modify x in the direction yielding private benefits to members of the party) or the benefits of x could be purely expressive.

[20] For example, suppose one of two spouses is a "workaholic" and the two spouses jointly decide how many hours she should work. The workaholic might always privately prefer to work more, and the household overall benefits from this to a point, but eventually the household benefits from that spouse being available for other household activities.

[21] I randomize references to her/his.

is observed by both players in each period (though i also knows $x_{i,t}$).

The true distribution of s_i, $i \in \{L, R\}$, is truncated Normal, with lower bound zero, $\mu_s > 0$, and σ_s^2. Due to the truncation, $E(s_i)$ is equal to μ_s plus a normalizing factor; however, it simplifies the analysis, and does not at all affect the nature of the results, to ignore this constant, i.e., to treat this distribution as (non-truncated) $N(\mu_s, \sigma_s^2)$.[22]

Each player always knows the true distribution of his own type; each bias considered only affects a characteristic of the other player. I assume complete unawareness of each bias considered; that is, in addition to being unaware of one's own bias, each player is also unaware of the other player's bias. Thus, each player thinks both players' beliefs about the distributions of both s_L and s_R are objectively correct (equal to the true distributions). As a result of unawareness, players do not have any uncertainty about the other player's beliefs, i.e., the players are certain that their second-order beliefs (beliefs about the other player's beliefs) are correct.

To simplify notation, let $s_{-i}^{i,t}$ denote the first-order expectation $E_{i,t}(s_{-i})$ (i's expectation of s_{-i} given all information observed through—prior to, and including—period t). Similarly, let $s_i^{i,-i,t}$ denote the second-order expectation $E_{i,t}\big(E_{-i,t}(s_i)\big)$ (i's expectation at the end of period t of $-i$'s expectation at the end of t of s_i), and $s_{-i}^{i,-i,i,t}$ denote the third-order expectation.

The unawareness assumption implies that $Var_{i,t}\big(s_i^{-i,t}\big) = 0$ for all i and t. Since i is unaware of any bias and thinks she knows $-i$'s prior beliefs, and i observes all information that $-i$ observes through period t, i is certain that she knows $-i$'s first-order beliefs in period t. Similarly, $Var_{i,t}\big(s_{-i}^{-i,i,t}\big) = 0$. An additional implication of these assumptions is that third-order beliefs are the same as first-order beliefs: i thinks $-i$ holds correct beliefs about i's beliefs, so $s_{-i}^{i,-i,i,t} = s_{-i}^{i,t}$ for all i and t. Complete unawareness is a stark assumption that makes the analysis tractable. Some degree of unawareness is necessary for biases to exist and the prevalence and importance of unawareness in general is supported by a recent literature focusing on this topic (see, e.g., Modica and Rustichini (1999)). The importance of unawareness should be kept in mind when interpreting the results.

Using this notation, given stage-game payoff maximization, the baseline period t choice

[22] To be consistent with this assumption, in the numerical examples I assume μ is at least four standard deviations from zero, implying $Pr(s_i > 0) > 0.9999$, making the truncated and non-truncated distributions essentially equivalent.

functions are:

$$x^*_{L,t} = s_L + rs_R^{L,t-1} + \tau_L = s_L + rs_R^{L,t-1}|_{\tau_L=0},$$
$$x^*_{R,t} = -(s_R + rs_L^{R,t-1}) + \tau_R = -(s_R + rs_L^{R,t-1})|_{\tau_R=0}. \quad (2)$$

3.3 Affective polarization

I assume that agent i's dislike of $-i$ is an increasing function of $s_{-i}^{i,t}$. It is intuitive and supported by research, e.g., Haidt (2012) and Graham, Nosek, and Haidt (2012), that empathy for others and devotion to social welfare are widely considered fundamental moral virtues, and that beliefs about out-party morality are key factors underlying feelings toward the out-party. Based on this assumption regarding dislike, I define two forms of affective polarization as follows, with $E_0(.)$ denoting the objective expectation in period 0 (the expectation using unbiased priors).

Definition 3.1. *"Expected affective polarization" occurs iff: $E_0(s_R^{L,t})$ and $E_0(s_L^{R,t})$ are increasing in t.*

Definition 3.2. *"Strong affective polarization" occurs iff: $E_0(s_R^{L,t}|s_L, s_R)$ and $E_0(s_L^{R,t}|s_L, s_R)$ are increasing in t for all s_L, s_R.*

"Expected affective polarization" occurs when both agents, on average grow to dislike each other more over time. "Strong affective polarization" occurs conditional on any realization of the s parameters—that is, even when s_L and s_R are arbitrarily small. I also characterize the probability limits of expectations of s_{-i} to the extent possible. Clearly as these grow further above the prior mean with higher probability this also indicates affective polarization.

Before proceeding to the analysis, it is worth discussing interpretation of the model a bit further. Since the actions may merely be expressive (e.g., statements of opinion), I do not conduct any welfare analysis and ask the reader to interpret the payoff functions and actions loosely. The payoff functions are merely meant to be reasonable approximations of the drivers of observable actions. The stage game choice functions, (2), show that actions indeed are determined by the parameters in a reasonable way, and allow for tractable analysis of how actions and beliefs evolve over time. Regarding welfare, it is reasonable to assume that welfare would decline as beliefs become more biased (as $E_0(s_{-i}^{i,t})$ and plim $s_{-i}^{i,t}$ diverge from s_{-i}), but "true welfare effects" are not modeled explicitly.

4 Analysis

4.1 Out-group bias (and no bias)

A natural bias to first consider is a prior bias against the out-group. That is, for $i \in \{L, R\}$:

$$E_{i,0}(s_{-i}) = \mu_s + b, \text{ with } b > 0.$$

It is possible that affective polarization results from a small initial bias building on itself. Again, the analysis of this case subsumes the case of no bias, as the results also hold with $b = 0$.

Consider belief updating by R about s_L (L's updating is symmetric). Since $\hat{x}_{i,t} = x^*_{i,t} + \epsilon^i_t$, the new information R observes in period t is $\hat{x}_{L,t} = s_L + rs_R^{L,t-1} + \epsilon^L_t$. R has no uncertainty about $s_R^{L,t-1}$, as discussed in Section 3. Thus, $Var_R(\hat{x}_{L,t}|s_L) = \sigma^2_\epsilon$, and R updates his expectation of s_L in the standard way given a normal prior and normal signal, as a weighted average of the prior mean and the signal, $\hat{x}_{L,t}$, adjusted so that its mean is equal to the parameter value, s_L (by subtracting off $rs_R^{R,L,t-1}$):

$$s_L^{R,t} = \lambda_t s_L^{R,t-1} + (1 - \lambda_t)(\hat{x}_{L,t} - rs_R^{R,L,t-1}), \tag{3}$$

with $\lambda_t = \frac{\sigma^2_\epsilon}{Var_{R,t-1}(s_L)+\sigma^2_\epsilon}$.

L *thinks* that R updates by:

$$\begin{aligned} s_L^{L,R,t} &= \lambda_t s_L^{L,R,t-1} + (1 - \lambda_t)(\hat{x}_{L,t} - rs_R^{L,R,L,t-1}) \\ &= \lambda_t s_L^{L,R,t-1} + (1 - \lambda_t)\left((s_L + rs_R^{L,t-1} + \epsilon^L_t) - rs_R^{L,t-1}\right) \\ &= \lambda_t s_L^{L,R,t-1} + (1 - \lambda_t)(s_L + \epsilon^L_t). \end{aligned} \tag{4}$$

There are two points worth noting here. First, both $s_L^{R,t}$ and $s_L^{L,R,t}$ use the same weight parameter, λ_t. This is because of the assumption that there is common knowledge of all of the variance parameters, and because all updated variances are functions of prior variances (and not means). This correct common knowledge of the updating weighting parameter will occur throughout the paper. Second, the second line uses the fact that third-order beliefs equal first-order beliefs as discussed in Section 2.2: $s_R^{L,R,L,t-1} = s_R^{L,t-1}$. The following result is

immediate.

Lemma 4.1. *(Second-order beliefs converge to truth.)* $s_i^{i,-i,t}$ *converges in probability to* s_i*, for* $i \in \{L, R\}$*, for all values of the parameters.*

(4) shows that the updating procedure for $s_i^{i,-i,t}$ is standard, and the signal is unbiased (its mean is the true value of the parameter, s_L). In this case, it is well known that even if the initial prior mean is biased, the posterior mean converges to truth.[23] Since L thinks the players have common priors and new information, L thinks that R knows how L updates her beliefs about s_R. Thus, L thinks R uses the correct adjusted signal $(\hat{x}_{L,t} - rs_R^{L,t-1})$ when in reality R's adjusted signal is incorrect.

Returning to (3), each side of the first line of (4) can be subtracted from the corresponding side of (3) to obtain:

$$s_L^{R,t} - s_L^{L,R,t} = \lambda_t(s_L^{R,t-1} - s_L^{L,R,t-1}) + (1-\lambda_t)r(s_R^{L,t-1} - s_R^{R,L,t-1}). \tag{5}$$

Thus, the difference between first and second-order beliefs at any point in time is deterministic. The difference is driven only by the initial bias in priors, not random shocks to the $\hat{x}$'s, since these shocks affect both first and second-order beliefs in the same way. The difference in these beliefs at any point in time is completely predictable.

We can obtain a simple closed-form expression for this difference in beliefs by iterating forward. For $t = 1$:

$$\begin{aligned} s_L^{R,1} - s_L^{L,R,1} &= \lambda_1 b + (1-\lambda_1)rb \\ &= b(r + (1-r)\lambda_1). \end{aligned} \tag{6}$$

By symmetry, the corresponding difference in beliefs about s_R is the same. For $t = 2$:

$$\begin{aligned} s_L^{R,2} - s_L^{L,R,2} &= \lambda_2(s_L^{R,1} - s_L^{L,R,1}) + (1-\lambda_2)r(s_L^{R,1} - s_L^{L,R,1}) \\ &= (r + (1-r)\lambda_2)(s_L^{R,1} - s_L^{L,R,1}) \\ &= b(r + (1-r)\lambda_2)(r + (1-r)\lambda_1). \end{aligned} \tag{7}$$

[23] See, e.g., Bullock (2009).

The pattern continues, so

$$s_{-i}^{i,t} - s_{-i}^{-i,i,t} = b \prod_{i=1:t} (r + (1-r)\lambda_i). \tag{8}$$

This implies the following.

Proposition 4.2. *With out-group bias, common knowledge of tastes and non-strategic payoffs:*

1. *Neither form of polarization occurs for any values of the parameters.*
2. *plim* $s_{-i}^{i,t} = s_{-i} + b$ *if* $r = 1$. *plim* $s_{-i}^{i,t} \in [s_{-i}, s_{-i} + b)$ *if* $r < 1$.
3. $\lim_{t\to\infty} E(x_{L,t}|s_L, s_R) = \lim_{t\to\infty} E(x_{R,t}|s_L, s_R) = s_L + s_R + b$ *if* $r = 1$.

Proofs are in the appendix. The proposition says that out-group bias is not exacerbated over time. The "worst-case scenario" is when $r = 1$; in this case the bias does not decline at all, and stays, on average, equal to the prior bias. Moreover, even in this case the actions not only do not diverge, they converge to be of equal magnitude of $s_L + s_R + b$. This action is stable because it is equal to what is expected. R's belief about s_L is biased upward by b, biasing R's belief about next period's $x_{L,t}$ upward by b. But R's belief about L's belief about s_R is biased downward by b. This bias exactly cancels the upward bias regarding $x_{L,t}$, causing R's belief about next period's $x_{L,t}$ to be unbiased. The same logic holds for L's belief about x_R. If r were less than one, then each player would act less selfishly than the other expects (on average). This causes beliefs about the other's s parameter to move toward truth. The only thing stopping these beliefs from reaching the corresponding true values (as the second sentence in part 2 of the proposition implies is possible) is that the players' belief precisions might become too high too soon.

Table 1 provides an example of how expected beliefs and actions evolve in the first five periods. The parameter values are $s_R = 1$, $s_L = 9$, and $\mu_s = 5$, with two values of r: 0.5 and 1. In both cases, before the start of play first-order beliefs are biased upward by 1 ($b = 1$), and so second-order beliefs are biased downward by 1. The first-order bias causes i to over-estimate $|E_(x_{-i,1})|$ by 1. The second-order bias causes underestimation by $r \times 1$. When $r = 0.5$, the former effect dominates, meaning x_{-i} is less extreme than expected (on average), causing s_{-i}^{i} to decline (on average), and the bias in first-order beliefs (the difference between these and objective first-order beliefs, i.e., second-order beliefs) to always decline (in the table, this decline is from 1 to 0.75 at the end of period 1, then to 0.5 at the end of period 2, etc). Note

that even though $s_L > \mu_s$, R's bias about s_L declines. R's prior is biased upward, and the signal is biased upward as well, but by less than the initial bias, which reduces the bias in the posterior. The signal is less biased to R than the prior because R "filters" x_L by subtracting off $rs_R^{R,L,t-1}$, whereas objectively R should subtract off $rs_R^{L,t-1}$. The bias in the signal is due to the bias in the filtering, which is r times the bias from the previous period, and therefore $r < 1$ ($r = 1$) causes the bias to be reduced (constant) across periods.

Table 1: Example of objective expectations (in $t = 0$) of beliefs and actions in first five periods; $\mu_s = 5$, $s_L = 9$, $s_R = 1$, $b = 1$, $\lambda_t = t/(t+1)$.

	Period	$E_R(x_{L,t})$	$x_{L,t}$	$E_L(x_{R,t})$	$x_{R,t}$	$s_R^{L,t}$	$s_R^{R,L,t}$	$s_L^{R,t}$	$s_L^{L,R,t}$
$r = 0.5$	0	n/a	n/a	n/a	n/a	6.00	5.00	6.00	5.00
	1	8.50	12.00	-8.50	-4.00	3.75	3.00	7.75	7.00
	2	9.25	10.88	-7.25	-4.88	2.17	1.67	8.83	8.33
	3	9.67	10.08	-6.33	-5.42	1.48	1.17	9.15	8.83
	4	9.73	9.74	-5.90	-5.57	1.22	1.03	9.15	8.97
	5	9.67	9.61	-5.70	-5.58	1.11	1.01	9.10	8.99
$r = 1$	0	n/a	n/a	n/a	n/a	6.00	5.00	6.00	5.00
	1	11.00	15.00	-11.00	-7.00	4.00	3.00	8.00	7.00
	2	11.00	13.00	-11.00	-9.00	2.67	1.67	9.33	8.33
	3	11.00	11.67	-11.00	-10.33	2.17	1.17	9.83	8.83
	4	11.00	11.17	-11.00	-10.83	2.03	1.03	9.97	8.97
	5	11.00	11.03	-11.00	-10.97	2.01	1.01	9.99	8.99

Figure 2 presents 100 simulations of the evolution of first-order beliefs for the first 1000 time periods, for $r = 1/3$, $r = 2/3$ and $r = 1$. For all parameter draws (simulations), beliefs quickly converge to close to truth in the first case, and fairly quickly converge in the second, but stay constant after a fairly quick adjustment period in the third. I present this figure largely as a contrast to analogous figures in the subsections below.

Note that the two aspects of the model that I considered to be most questionable—complete unawareness of prior bias, and lack of reputation concerns—should not be concerns for these results, since both of these assumptions should cause affective polarization to be more likely to occur. These results also make it clear that affective polarization is not built in to the model in any way, due to these assumptions or others.

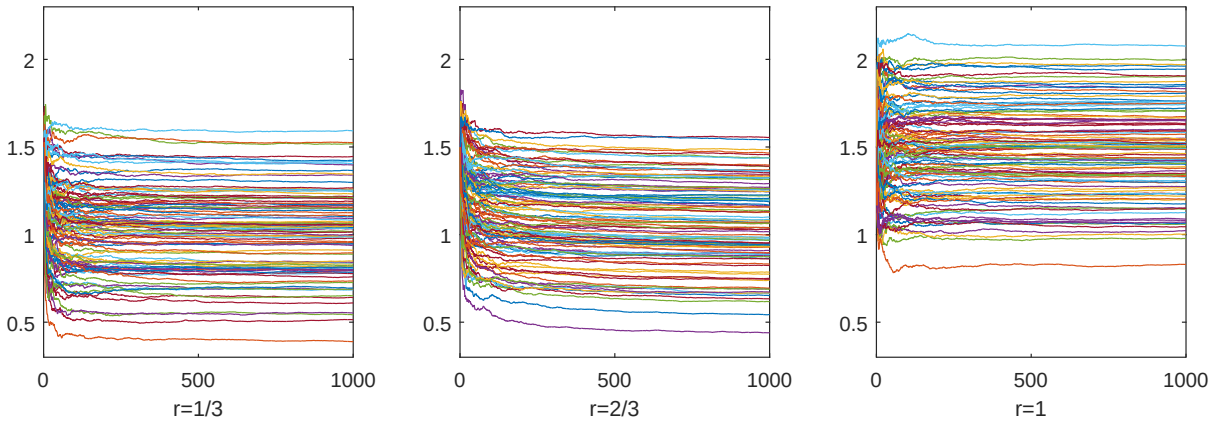

Figure 2: Simulated first-order beliefs $(s_{-i}^{i,t})$ for $t = 0 - 1000$. One hundred simulations per set of variance parameter values; $\mu_s = 1$, $\sigma_s^2 = \sigma_\epsilon^2 = 0.25$, $b = 0.5$ (and thus $(s_{-i}^{i,0} = 1.5)$ for all simulations.

4.2 False consensus bias

The second bias that I consider is the false consensus bias. Informally, this is the tendency to believe others are more similar to us in "horizontally differentiated" characteristics than they really are.[24] Assume now that each player i holds correct priors about s_{-i} but now τ_{-i} is unobserved and for each i, $\tau_i \sim N(\mu_{\tau_i}, \sigma_\tau^2)$. I assume $\mu_{\tau_L} \geq \mu_{\tau_R}$ because it is natural to expect that L is more likely to prefer higher x due to either selection (partisans select into their party due to true political tastes) and/or motivated reasoning (partisans believe what they want to be true is true; e.g., if one privately benefits from big government, one is more likely to then truly believe this is also best for society overall). As mentioned in the introduction, Haidt (2012) is a very well-known work claiming that political discord results from under-estimation of heterogeneity in moral "tastes." I formalize the false consensus bias as follows:

$$E_{L,0}(\tau_R) = \mu_{\tau_R} + b, \text{ and } E_{R,0}(\tau_L) = \mu_{\tau_L} - b, \text{ with } b \in (0, \mu_{\tau_L} - \mu_{\tau_R}].$$

First consider the case of no reciprocity ($r = 0$). In this case, $\hat{x}_{L,t} = s_L + \tau_L + \epsilon_1^L$. R's

[24] See Ross, Greene, and House (1977) for early work from psychology and Butler, Giuliano, and Guiso (2015) for a recent application from economics. By contrast, the over-optimism form of overconfidence makes us over-estimate our advantages over others in vertically differentiated dimensions.

updated expectation of s_L after the first period is:

$$\begin{aligned} s_L^{R,1} &= \lambda_1 s_L^{R,0} + (1-\lambda_1)(\hat{x}_{L,1} - \tau_L^{R,0}) \\ &= \lambda_1 \mu_s + (1-\lambda_1)\Big(s_L + (\tau_L - \tau_L^{R,0}) + \epsilon_1^L\Big). \end{aligned} \tag{9}$$

Taking objective expectations: $E_0(s_L^{R,1}) = \mu_s + (1-\lambda_1)b$ since $E_0(\tau_L - \tau_L^{R,0}) = b$. Simply observing one action by the other player causes dislike of that player to increase, on average. Since the bias is symmetric for L's beliefs about R, we have already obtained at least some affective polarization. Since each player under-estimates how different the other's tastes are from her own, each likely observes the other stating an opinion that appears more self-serving than it really is.

To see what happens asymptotically, rather than analyze how the players update with the signals one at a time, it is easier to consider updating based on the mean of the full set of signals: $\overline{x}_{L,t} = \frac{1}{t}(\hat{x}_{L,1} + \hat{x}_{L,2} + ... + \hat{x}_{L,t})$. This sample mean naturally accounts for the way in which the observations are correlated (their dependence on fixed s_L and τ_L), with $E_{R,0}(\overline{x}_{L,t}|s_L) = s_L + \tau_L^{R,0}$, and $Var_{R,0}(\overline{x}_{L,t}|s_0) = Var_0(\overline{x}_{L,t}|s_0) = \sigma_{\tau_L}^2 + \sigma_\epsilon^2/t$. The observations are i.i.d. given $r = 0$, so no information is lost in combining them this way. R then updates his expectation (from the period 0 prior, after observing the first t signals) to:

$$\begin{aligned} s_L^{R,t} &= \lambda_t s_L^{R,0} + (1-\lambda_t)(\overline{x}_{L,t} - \tau_L^{R,0}) \\ &= \lambda_t \mu_s + (1-\lambda_t)(s_L + \tau_L - \tau_L^{R,0} + \frac{1}{t}\sum_{i=1}^{t} \epsilon_i^L), \end{aligned} \tag{10}$$

with $\lambda_t = \frac{Var_0(\overline{x}_{L,t}|s_L)}{\sigma_s^2 + Var_0(\overline{x}_{L,t}|s_L)}$. It is then clear that $E_0(s_L^{R,t}|s_L, \tau_L) = \lambda_t \mu_s + (1-\lambda_t)(s_L + \tau_L - \tau_L^{R,0})$, and $E_0(s_L^{R,t}|s_L) = \lambda_t \mu_s + (1-\lambda_t)(s_L + b)$, with λ_t decreasing in t, implying the next result.

Proposition 4.3. *With uncertainty in tastes, false consensus bias b, and $r = 0$:*

1. Expected affective polarization occurs for all $b > 0$, for all t. A sufficient condition for strong affective polarization to never occur is $b < \mu_s$.

2. plim $s_L^{R,t} = \frac{\sigma_\tau^2}{\sigma_s^2+\sigma_\tau^2}\mu_s + \frac{\sigma_s^2}{\sigma_s^2+\sigma_\tau^2}(s_L + \tau_L - \mu_{\tau_L} + b)$, and plim $s_R^{L,t}$ is analogous. The marginal effect of b on plim $s_{-i}^{i,t}$ is increasing in σ_s^2 and decreasing in $\sigma_{\tau_L}^2$. The variance of i's belief about s_{-i} does not converge to zero.

If rightists believe that leftists are on average more conservative than they really are (in

tastes), and acting "liberally" is consistent with leftists acting selfishly, then rightists will on average update beliefs about the selfishness of leftists upward over time. The logic is analogous for leftists' beliefs about rightists, and this is why expected affective polarization occurs. The strong version of affective polarization does not occur if $b < \mu_s$ because if $s_L + \tau_L - \mu_{\tau_L} + b < \mu_s$ then $E_0(s_L^{R,t})$ must be decreasing in t, and $b < \mu_s$ guarantees $s_L + \tau_L - \mu_{\tau_L} + b < \mu_s$ for sufficiently small s_L and τ_L. Neither player ever pins down the other player's s parameter with certainty because i only obtains repeated observations of $s_{-i} + \tau_{-i}$ and cannot separately identify the components (and the players are aware of this identification problem). This is why $E_0(s_i^{-i,t}|s_i)$ is a weighted average of the prior mean and truth even if $b = 0$. The long run bias in this weighted average (the term involving b) is greater when σ_s^2 is larger (i is more uncertain about s_{-i} ex ante, and thus is more influenced by new information) and when $\sigma_{\tau_L}^2$ is smaller (i is more certain about τ_{-i} ex ante, and is thus less likely to attribute apparently self-serving actions to differences in tastes).

While this mean affective polarization result is fairly simple to derive and is almost surely something that has been identified in other contexts it is not necessarily intuitive—we might assume that a Bayesian updater with unbiased signals would learn the truth over time. In fact, biased dislike increases over time, on average. Bias in beliefs about taste influence long-run beliefs about character *more* than bias in beliefs about character (out-group bias).

However, there is no polarization of actions, and the long run affective polarization is bounded, potentially quite low if initial bias is small, and does not occur for all realizations of the s and τ parameters. I next incorporate reciprocity to see how this affects the results (but maintain the assumption of no out-group bias). To simplify, assume $r = 1$; the intuition from the previous section that this causes the maximal level of belief divergence continues to hold.

In $t = 1$, R now subtracts off $s_R^{R,L,0}$ and τ_L^R from $\hat{x}_{L,1}$ to obtain a signal with expected value s_L, so R's updated expectation for s_L is:

$$\begin{aligned} s_L^{R,1} &= \lambda_1 s_L^{R,0} + (1-\lambda_1)(\hat{x}_{L,1} - s_R^{R,L,0} - \tau_L^{R,0}) \\ &= \lambda_1 \mu_s + (1-\lambda_1)\big((s_L + s_R^{L,0} + \tau_L + \epsilon_1^L) - s_R^{R,L,0} - \tau_L^{R,0}\big) \\ &= \lambda_1 \mu_s + (1-\lambda_1)\big(s_L + \tau_L + \epsilon_1^L - (\mu_{\tau_L} - b)\big). \end{aligned} \tag{11}$$

The simplification in the last line uses the fact that $s_R^{L,0} = s_R^{R,L,0}$. Meanwhile,

$$s_L^{L,R,1} = \lambda_1\mu_s + (1-\lambda_1)(s_L + \tau_L + \epsilon_1^L - \mu_{\tau_L}), \tag{12}$$

since L is unaware of R's biased prior about τ_L. It is clear that, again, second-order beliefs converge to the normatively ideal value, given available information Thus,

$$s_L^{R,1} - s_L^{L,R,1} = (1-\lambda_1)b, \tag{13}$$

which is also the value of $s_R^{L,1} - s_R^{R,L,1}$ by symmetry. This value is the same as that of the case of $r = 0$.

However, things are different from the $r = 0$ case for $t = 2$ and beyond. Again, it is useful now to consider R updating conditional on the mean of observations, $\overline{x}_t^L$, but this now takes a more complicated form:

$$\overline{x}_{L,t} = \frac{1}{t}\sum_{i=1}^{t}\hat{x}_i^L = s_L + \tau_L + \frac{1}{t}\sum_{i=1}^{t}(s_R^{L,i-1} + \epsilon_i^L). \tag{14}$$

Since R knows that each $\hat{x}_t^L$ is driven in part by L's beliefs about s_R given information available prior to t, R will adjust $\overline{x}_{L,t}$ accordingly, as he sees fit, when updating beliefs about s_L. That is, R will also subtract off her beliefs about $\frac{1}{t}\sum_{i=1}^{t} s_R^{L,i-1}$, in addition to τ_L, so R's updated expected value for s_L in period t will be a weighted average of the prior, μ, and $\overline{x}^L - (\frac{1}{t}\sum_{i=1}^{t} s_R^{R,L,i-1} + \tau_L^{R,0})$.

So, in period 2, R updates by:

$$\begin{aligned}
s_L^{R,2} &= \lambda_2 s_L^{R,0} + (1-\lambda_2)\left(\overline{x}_{L,2} - \left(\frac{1}{2}(s_R^{R,L,0} + s_R^{R,L,1}) + \tau_L^{R,0}\right)\right) \\
&= \lambda_2\mu_s + (1-\lambda_2)\left(s_L + (1/2)\big((s_R^{L,0} - s_R^{R,L,0}) + (s_R^{L,1} - s_R^{R,L,1}) + \epsilon_1^L + \epsilon_2^L\big) + \tau_L - \tau_L^{R,0}\right) \\
&= \lambda_2\mu_s + (1-\lambda_2)\Big(s_L + (1/2)\big((1-\lambda_1)b + \epsilon_1^L + \epsilon_2^L\big) + \tau_L - (\mu_{\tau_L} - b)\Big),
\end{aligned} \tag{15}$$

The belief is biased upward now due to both the initial false consensus bias and the bias this causes in beliefs about s_R at the end of period 1 (in this case, $s_R^{L,1} - s_R^{R,L,1} = (1-\lambda_1)b$). That is, the false consensus bias causes L to over-estimate s_R after period 1, which causes L to take a more extreme action in period 2 due to reciprocity, which in turn causes R to

over-estimate s_L even more after period 2 as compared to after period 1.

Let $b_t := s_{-i}^{i,t} - s_{-i}^{-i,i,t}$ denote the difference between first and second-order beliefs about s_{-i} in period t. As in the analysis of Section 3.1, this difference is deterministic because both players condition on the ϵ's in the same way, symmetric because of the assumed symmetry in priors, and second-order beliefs are objectively correct because players hold unbiased beliefs about their own types' distributions and are unaware of the other types' bias. R's beliefs about s_L in any period t can be written as a function of the earlier b_t's:

$$s_L^{R,t} = \lambda_t \mu_s + (1-\lambda_t)\left(s_L + \frac{1}{t}\sum_{i=1}^{t}\epsilon_i^L\right) + (1-\lambda_t)\left(\frac{1}{t}\sum_{i=1}^{t-1} b_i + \tau_L - \tau_L^{R,0}\right), \qquad (16)$$

which implies

$$b_t = (1-\lambda_t)\left(b + \frac{1}{t}\sum_{i=1}^{t-1} b_i\right) \text{ for } t > 1, \text{ with } b_1 = (1-\lambda_1)b. \qquad (17)$$

In the appendix I show that if $(1-\lambda_t) = 1$, then b_t is equal to the harmonic series $(1 + 1/2 + 1/3 ...)$. This would imply that b_t not only always increases, but diverges! The next proposition follows from this.

Proposition 4.4. *With uncertainty in tastes, false consensus bias b and $r = 1$:*

1. Expected polarization occurs for all values of the parameters, for all t.

2. plim $s_L^{R,t} = \frac{\tau_s^2}{\sigma_s^2+\sigma_{\tau_L}^2}\mu_s + \frac{\sigma_s^2}{\sigma_s^2+\sigma_{\tau_L}^2}(s_L + \tau_L - \mu_{\tau_L} + b) + \lim_{t\to\infty} b_t$, and plim $s_L^{R,t}$ is analogous. plim $s_{-i}^{i,t}$, for each i, can be arbitrarily large for sufficiently large σ_s^2 and small $\sigma_{\tau_L}^2$ and σ_ϵ^2, for any b, s_L, s_R, τ_L and τ_R.

3. There exists t' such that strong polarization occurs for $t \geq t'$, for all $b > 0$, for sufficiently large σ_s^2 and small $\sigma_{\tau_L}^2$ and σ_ϵ^2.

The divergence in $s_{-i}^{i,t}$ occurs when the variance of the prior about tastes goes to zero. Strong polarization, which is indeed quite a strong result—again, this means increasing $s_{-i}^{i,t}$ for each i, for all realizations of the s's and τ's—can occur because the term $\frac{1}{t}\sum_{i=1}^{t-1} b_i$ increases for all realizations of these parameters. Again, this is because of the difference in first and second-order beliefs that occurs for all parameter values. Even when player i is very "good," she over-estimates how "bad" $-i$ is, causing i to next act (excessively) "badly," and this misunderstanding compounds over time. It may not dominate in early periods if s_i is sufficiently

below the mean, but since s_i is constant, its effect on changes in s_i^{-i} across periods shrinks to zero, while b_t always grows.

Figure 3 shows that strong polarization occurs for reasonable parameter values and typically starts quite early—as early as $t = 2$ for all or nearly all of the simulations shown. In the left graphs in the figure, I present beliefs (s^i_{-i} for just the first 10 periods so these can be seen more clearly. While beliefs sometimes decline from $t = 1$ to $t = 2$, when s_{-i} is indeed low (and perhaps also due to small ϵ's), beliefs quickly and steadily begin to rise again across *all* simulations (all parameter values drawn). The right graphs show how beliefs continue to grow over time; however, the growth rate seems to decline to zero quickly when σ^2_τ is larger, and even for the smaller value of this parameter, first-order beliefs do not grow very large. However, these beliefs still do approximately quadruple on average, which still indicates quite substantial polarization and bias.

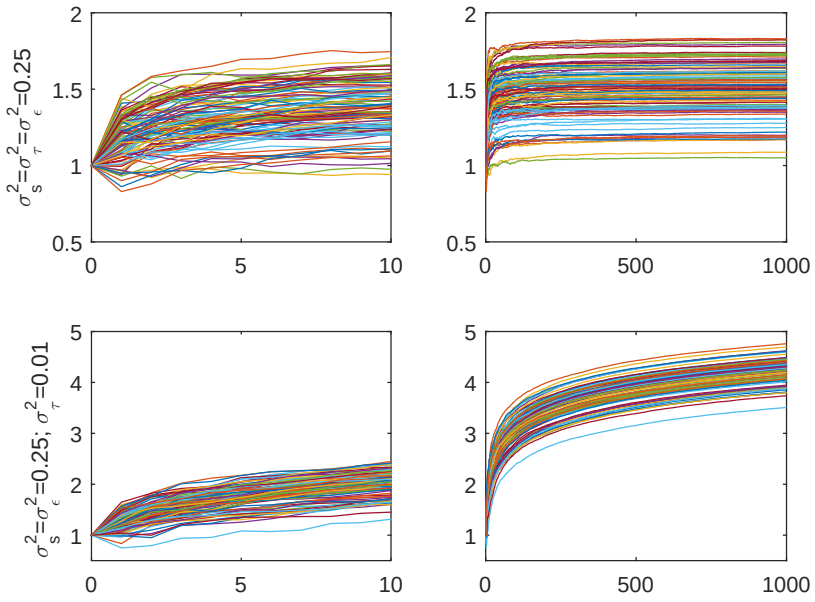

Figure 3: Simulated first-order beliefs ($s^{i,t}_{-i}$) for $t = 0-10$ (left graphs) and $t = 0-1000$ (right graphs). One hundred simulations per set of variance parameter values; $\mu_s = 1$, $\mu_{\tau_L} = 1 = -\mu_{\tau_R}$, $b = 0.5$ in all graphs.

The strength of these results is driven by the strength of the unawareness assumption. This assumption seems reasonable for the false consensus bias in particular, since this is a subtle bias (in fact, it is rarely referred to in economics work), and there is no ego-driven, or otherwise self-serving, motive to believe others are more susceptible to this bias than ourselves. As a

result of unawareness, the assumption that the agents lack reputation concerns is much less of an issue, since the agents do not realize their own reputations are being distorted. This does imply that the agents are unaware of how much they are disliked by the opposition, which is questionable, and in fact worth discussing further, which I do in the concluding remarks.

4.3 (Limited) strategic thinking

Two key questions that remain at this point are: 1) to what extent do the results above depend on the assumption that actions are chosen unilaterally? How might strategic choice affect the results? And, 2) to what extent are the strong results for the false consensus bias unique to that bias? To address both of these questions, I now assume that actions are chosen jointly in the simplest possible way, $x = (x_L + x_R)/2$, and consider a different bias, limited strategic thinking, described just below. Return to the baseline assumptions of correct priors on s_{-i} and $\tau_i = 0$ for each i, and to simplify assume $r = 0$.

To simplify algebra, drop the (1/2) from α_i, implying the stage-game best response functions are now:

$$\begin{aligned} x^*_{L,t}(x_{R,t}) &= s_L - x_{R,t}, \\ x^*_{R,t}(x_{L,t}) &= -s_R - x_{L,t}. \end{aligned} \tag{18}$$

There is no Nash equilibrium in the stage game. For any given action by one player, the other has an incentive to take a more extreme action (in the other direction) to push things slightly in the preferred direction.[25] If there were bounds to the x_L, x_R that could be chosen, there would be a stage game NE in which L would choose x_L at the upper bound and x_R would equal the lower bound, and neither player would update beliefs about the other's s parameter. That is to say, although extreme actions would be taken, they would occur in the first period, and there would be no updating regarding the s's.

Thus, in this model, equilibrium strategic thinking cannot explain growth in extremism of actions or affective polarization. Next, consider behavioral strategic thinking. The rest of this subsection uses the level-k model of strategic thinking. This has become the benchmark behavioral alternative to equilibrium (Crawford, Costa-Gomes, and Iriberri, 2013). A level-k

[25]Plug player i's best response function into $-i$'s to eliminate x_i and obtain a function of just x_{-i}, and it is immediate that this equation has no solution.

strategic thinker best responds to a level $k-1$ opponent. A level-0 player's action is determined by assumption, as this type of player is non-strategic. This action is typically assumed either to be based on a salient benchmark, or a uniform randomization.

The level-k model is usually thought to best apply to games in which players have limited experience, especially one shot games. Thus, it seems perhaps not ideal for the repeated interaction that I study. I use it because it is a tractable model for capturing the key issue—under-estimation of strategic motives. Moreover, while in reality, political interactions are repeated, each one is unique and may involve different actors. It is not at all implausible that strategic motives could be consistently under-appreciated.

Again, to both simplify the analysis and to focus on the specific effects of the new element of the model (strategic thinking), I assume away the reciprocity preference (there is an endogenous strategic reciprocity motive) and both biases in prior beliefs, and also assume there is common knowledge in tastes. W.l.o.g. assume $\tau_L = \tau_R = 0$. Let $x_{i,t}^{\mathcal{L}k}$ denote the best response of player i when she is a level k thinker in period t. For $k > 0$, these are:

$$\begin{aligned} x_{L,t}^{\mathcal{L}k} &= s_L - E_L(x_{R,t}^{\mathcal{L}k-1}), \\ x_{R,t}^{\mathcal{L}k} &= -s_R - E_R(x_{L,t}^{\mathcal{L}k-1}). \end{aligned}$$

There are then two key questions. First, what level of "thinking" should be used for L and R. Level-0 thinkers are rare or do not exist at all in most empirical contexts, and since their actions are non-strategic, the analysis would be degenerate. Level-1 is a better option, but level-2 is preferable for two reasons: first, level-2 is typically more common empirically, and second, since level-2 thinkers believe their opponents engage in some strategic thinking, level-2 thinkers should update beliefs about their opponents' types after observing their actions, while this would typically not be the case for level-1 thinkers. Results would be more similar to the Nash Equilibrium benchmark if players were higher level strategic thinkers.

The second question to address before proceeding is how the non-strategic level-0 players behave. I consider a natural benchmark in which $\mathcal{L}0$ players choose x in each period equal to their taste parameter (zero). This is what players would choose if they were completely non-strategic in the sense of not considering their private interests, and also happens to be both a salient reference point, and the expected action from uniform randomization over the

action space (a common assumption made for level-0 play).

Given these assumptions, $x_{L,t}^{\mathcal{L}1} = s_L$, $x_{R,t}^{\mathcal{L}1} = -s_R$, and $x_{L,t}^{\mathcal{L}2} = s_L + s_R^{L,t-1}$, $x_{R,t}^{\mathcal{L}2} = -s_R - s_L^{R,t-1}$. The belief terms will be determined given the assumption that the other player is $\mathcal{L}1$, so for R this is

$$s_L^{R,t} = \lambda_t \mu_s + (1-\lambda_t)(\frac{1}{t})(\hat{x}_1^L + \hat{x}_2^L + ...\hat{x}_t^L). \tag{19}$$

Since L is actually a level-2 thinker,

$$s_L^{R,t} = \lambda_t \mu_s + (1-\lambda_t)(\frac{1}{t})\left((s_L + s_R^{L,0}) + (s_L + s_R^{L,1}) + ... + (s_L + s_R^{L,t-1}) + \sum_{i=1}^{t} \epsilon_i^L\right). \tag{20}$$

$s_R^{L,1}, s_R^{L,2}$, etc, will be determined analogously as (increasing) functions of s_R and $s_L^{R,t}$'s for prior periods, which will in turn be functions of s_L and earlier $s_R^{L,t}$'s. Thus, to proceed in the characterization of $s_{-i}^{i,t}$ in general, it is simplest to look for a lower bound, based on $s := \min\{s_L, s_R\}$. This can be used to then show that (20) implies

$$\begin{aligned} E_0(s_L^{R,t}|s_L, s_R) &\geq \lambda_t \mu_s + s_t, \text{ in which} \\ s_t &= (1-\lambda_t)(s + \frac{1}{t}\sum_{i=1}^{t-1} s_i) \text{ for } t > 1 \text{ with } s_1 = (1-\lambda_1)s. \end{aligned} \tag{21}$$

This expression is equivalent to (16), and thus implies the following.

Proposition 4.5. *With common knowledge of tastes, $r = 0$, and strategic payoffs, if L and R are $\mathcal{L}2$ thinkers who assume $\mathcal{L}0$ thinkers play $x_i = \tau_i = 0$, then:*

1. Expected affective polarization occurs for all t.

2. The variance of i's beliefs about s_{-i} converges to zero. Still, if $\sigma_s^2 \geq \sigma_\epsilon^2$: $\lim_{t\to\infty} E_0(s_{-i}^{i,t}|s_L, s_R) = \infty$ for all s_L and s_R, and strong affective polarization occurs for sufficiently large t.

These results are comparable to those of Proposition 4.4 but even stronger: $E_0(s_{-i}^{i,t}|s_L, s_R)$ diverges (for each i), and strong polarization occurs, for a large, plausible, well-defined range of parameters. The left graphs in Figure 4 show how $s_{-i}^{i,t}$ almost always increases immediately, and the right graphs show how this belief rises to much higher levels than it does for the false consensus bias case.

The intuition is fairly straightforward: in period 1, R expects L to play s_L (the best

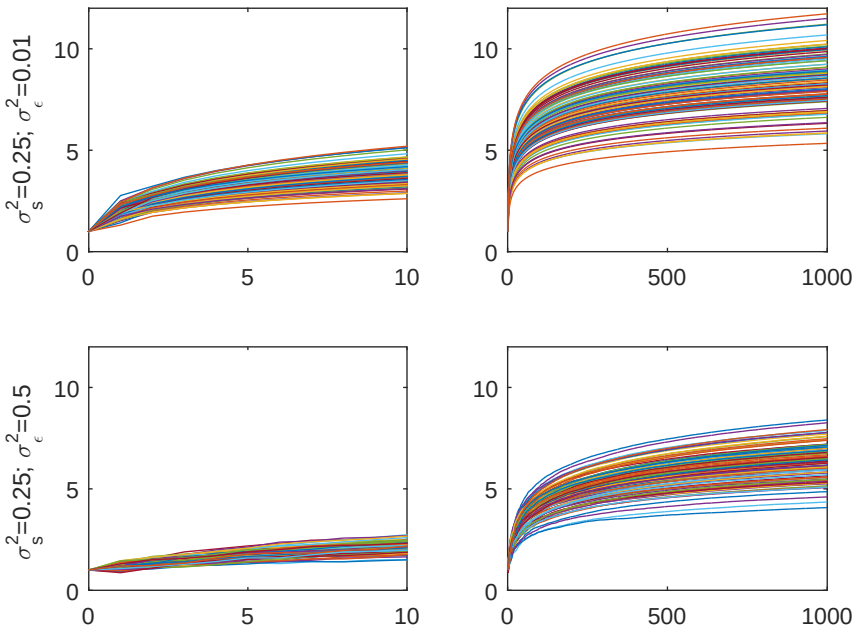

Figure 4: Simulated first-order beliefs ($s^{i,t}_{-i}$) for $t = 0-10$ (left graphs) and $t = 0-1000$ (right graphs). One hundred simulations per set of variance parameter values; $\mu_s = 1$ in all graphs.

response to $x_R = 0$), but L plays $s^{L,0}_R + s_L = \mu_s + s_L$ (the best response to $E(x_R) = E(s_R)$). Thus, $x_{L,1}$ is objectively expected to exceed R's expectation for all $s_L > 0$. This causes each player i to update beliefs about s_{-i} upward, causing future actions to become more extreme, causing beliefs to continue to grow, etc. Again there is a race between precision of beliefs and extremism of actions, and in this case a minimal parameter condition for extremism to "win" is easy to characterize ($\sigma^2_s \geq \sigma^2_\epsilon$). Note that the unawareness assumption again applies here in that agents are fully unaware that their belief about the other agent's level of thinking is inaccurate.

5 Concluding Remarks

Why does disagreement lead to dislike? Why do repeated interactions often cause small conflicts to escalate and lead to extreme actions and hostility? This paper shows how cognitive biases can cause these phenomena and, in particular, seemingly unrelated bias, in conjunction with unawareness of bias, can cause affective polarization. Such biases, even if small, can cause large degrees of dislike even between quite good people. By contrast, bias that appears

more directly relevant to dislike is less pernicious to the exacerbation of dislike. The paper also highlights and supports the connection between cognitive and affective psychological processes, perhaps opening up new topics for economic analysis.

Again, I do not claim that the biases I study are the only explanations for partisan affective polarization. The growth in availability of negative information (due to the increase in campaign advertising, social and political media, etc) in conjunction with (dis)confirmation bias and motivated reasoning is still likely a huge factor. Institutional changes (e.g., gerrymandering) and changing norms in political behavior (regarding, e.g., filibusters) are also likely relevant. Gridlock caused by strategic obstructionism can be a signal of poor character (Stone, 2013). However, it is also possible that cognitive biases such as those that I study exacerbate reactions to these other factors. For example, an institutional change could increase the incentives for certain types of strategic behavior and behavior favoring certain political tastes, and media changes could make partisan citizens more likely to be informed about this new behavior. Limited strategic thinking, the false consensus bias, and likely other biases could then result in this information being misinterpreted, resulting in greater out-party dislike.

Given the importance of the unawareness assumption, it is worth discussing its validity further. Is it plausible that individuals would maintain such unawareness over time? Yes, at least to some extent; in many contexts people remain unaware of their biases for long periods of time (e.g., think of those who continue to arrive 15 minutes late to appointments throughout their lives). Persistent unawareness is especially plausible for political beliefs due to low incentives to increase awareness and potential psychological incentives to maintain it due to motivated reasoning—partisans may enjoy feeling superior to the opposition and the idea of "fighting the good fight."

Another key aspect of the model that is questionable is second-order belief truth convergence—that player i thinks $-i$ eventually knows s_i. I conjecture that in reality partisans indeed do have overly-optimistic beliefs about how their social-mindedness is perceived by the out-party. But I also would not deny that in reality partisans are likely aware of out-party claims of dislike. This could be reconciled with second-order truth convergence, if the out-party's claimed dislike was perceived as exaggerated for strategic or psychological reasons. But partisans likely do often know (or believe) they are truly disliked by the out-party. Even this dislike could be reconciled with second-order belief convergence if the dislike was due to a mechanism other

than the one in the model. That is, L could feel R knows that s_L is low (L is "good"), but also that s_R is higher, making R resent and dislike L. Causes of dislike like this surely exist and do not invalidate the model—I do not claim it explains all partisan dislike; the model may apply more to one party than the other, more to some members of a party than others, and more to partisan relations at some points in time than at others.

An important alternative bias worth discussing is confirmation bias. It is possible that anti-out-group prior bias plus confirmation bias could lead to polarization. However, confirmation bias seems more likely to maintain bias against the out-group, rather than exacerbate it. Furthermore, confirmation bias is a bias that people are relatively likely to believe afflicts others and not ourselves (Shermer, 2011). This would cause us to discount the inferences we make about others' character based on their actions, attributing these actions to mere cognitive bias and not character flaws. Finally it is worth noting that the model prediction that disagreement is bound to spiral out of control is obviously not always realistic. Hostile disagreements are resolved, or perhaps do not even grow in the first place, because of many elements the model excludes, such as reputation concerns, outside information, and possible awareness of bias. These elements are excluded to focus on the forces driving the most severe polarization.

Topics for future work include empirical analysis of second (and perhaps higher) order beliefs about character and motives across the parties, of the relationship between such beliefs and out-party dislike, and of the relationships between various biases and partyism in the real world. Another certainly important open topic is the practical issue of increasing awareness of, or simply reducing, the biases, or reducing partisan dislike in other ways. It is possible that simply spreading the word about research showing the connection between bias and dislike could help shame people into reducing hostility toward the opposition and strengthen social norms supporting cooperation and against partyism (Iyengar and Westwood, 2015).

References

Abramowitz, A., and K. Saunders (2008): "Is polarization a myth?," *Journal of Politics*, 70(2), 542–55.

Acemoglu, D., and A. Wolitzky (2014): "Cycles of conflict: An economic model," *The American Economic Review*, 104(4), 1350–1367.

Andreoni, J., and T. Mylovanov (2012): "Diverging opinions," *American Economic Journal: Microeconomics*, 4(1), 209–232.

Andris, C., D. Lee, M. J. Hamilton, M. Martino, C. E. Gunning, and J. A. Selden (2015): "The Rise of Partisanship and Super-cooperators in the US House of Representatives," *PloS one*, 10(4), e0123507.

Aumann, R. J. (1976): "Agreeing to disagree," *The annals of statistics*, pp. 1236–1239.

Baliga, S., E. Hanany, and P. Klibanoff (2013): "Polarization and ambiguity," *The American Economic Review*, 103(7), 3071–3083.

Barber, M. J., and N. McCarty (2015): "Causes and Consequences of Polarization," *Solutions to Political Polarization in America*, p. 15.

Batson, C. D., and A. A. Powell (2003): "Altruism and prosocial behavior," *Handbook of psychology*.

Benoît, J.-P., and J. Dubra (2014): "A Theory of Rational Attitude Polarization," *Available at SSRN 2529494*.

Bernheim, B. D., and N. Kartik (2014): "Candidates, Character, and Corruption," *American Economic Journal: Microeconomics*, 6(2), 205–46.

Blomberg, B., and J. E. Harrington (2000): "A theory of rigid extremists and flexible moderates with an application to the US Congress," *The American Economic Review*, 90(3), 605–620.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (forthcoming): "Stereotypes," *Quarterly Journal of Economics*.

Brennan, G., and A. Hamlin (1998): "Expressive voting and electoral equilibrium," *Public choice*, 95(1-2), 149–175.

Bullock, J. G. (2009): "Partisan bias and the Bayesian ideal in the study of public opinion," *The Journal of Politics*, 71(03), 1109–1124.

Butler, J. V., P. Giuliano, and L. Guiso (2015): "Trust, values, and false consensus," *International Economic Review*, 56(3), 889–915.

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri (2013): "Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications," *Journal of Economic Literature*, 51(1), 5–62.

Duffy, J., and M. Tavits (2008): "Beliefs and voting decisions: A test of the pivotal voter model," *American Journal of Political Science*, 52(3), 603–618.

Epley, N. (2014): *Mindwise: Why we misunderstand what others think, believe, feel, and want*. Vintage.

Esponda, I., and D. Pouzo (2016): "Berk–Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models," *Econometrica*, 84(3), 1093–1130.

Fiorina, M., S. Abrams, and J. Pope (2008): "Polarization in the American public: Misconceptions and misreadings," *Journal of Politics*, 70(2), 556–560.

Glaeser, E. L. (2005): "The political economy of hatred," *The Quarterly Journal of Economics*, pp. 45–86.

Graham, J., B. A. Nosek, and J. Haidt (2012): "The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum," *PloS one*, 7(12), e50092.

Haidt, J. (2012): *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Hetherington, M. J., and T. J. Rudolph (2015): *Why Washington Won't Work: Polarization, Political Trust, and the Governing Crisis*, vol. 104. University of Chicago Press.

Hill, S. J., and C. Tausanovitch (forthcoming): "A Disconnect in Representation? Comparison of Trends in Congressional and Public Polarization," Discussion paper, Working Paper.

Iyengar, S., G. Sood, and Y. Lelkes (2012): "Affect, not ideology a social identity perspective on polarization," *Public Opinion Quarterly*, 76(3), 405–431.

Iyengar, S., and S. J. Westwood (2015): "Fear and loathing across party lines: New evidence on group polarization," *American Journal of Political Science*, 59(3), 690–707.

Klumpp, T., and H. M. Mialon (2013): "On hatred," *American law and economics review*, p. aht004.

Lelkes, Y., G. Sood, and S. Iyengar (2015): "The hostile audience: The effect of access to broadband Internet on partisan affect," *American Journal of Political Science*.

Levine, D. K. (1998): "Modeling altruism and spitefulness in experiments," *Review of economic dynamics*, 1(3), 593–622.

Mann, T. E., and N. J. Ornstein (2013): *It's even worse than it looks: How the American constitutional system collided with the new politics of extremism*. Basic Books.

Mason, L. (2015): "I Disrespectfully Agree: The Differential Effects of Partisan Sorting on Social and Issue Polarization," *American Journal of Political Science*, 59(1), 128–145.

Modica, S., and A. Rustichini (1999): "Unawareness and partitional information structures," *Games and economic Behavior*, 27(2), 265–298.

Nussbaum, M. C. (2003): *Upheavals of thought: The intelligence of emotions*. Cambridge University Press.

Ortoleva, P., and E. Snowberg (2015): "Overconfidence in Political Behavior," *American Economic Review*, 105(2), 504–535.

Pessoa, L. (2008): "On the relationship between emotion and cognition," *Nature reviews neuroscience*, 9(2), 148–158.

Rempel, J. K., and S. Sutherland (2016): "Hate: Theory and Implications for Intimate Relationships," in *The Psychology of Love and Hate in Intimate Relationships*, pp. 105–129. Springer.

Rogowski, J. C., and J. L. Sutherland (2015): "How Ideology Fuels Affective Polarization," *Political Behavior*, pp. 1–24.

Ross, L., D. Greene, and P. House (1977): "The false consensus effect: An egocentric bias in social perception and attribution processes," *Journal of experimental social psychology*, 13(3), 279–301.

Roux, N., and J. Sobel (2015): "Group polarization in a model of information aggregation," *American Economic Journal: Microeconomics*, 7(4), 202–232.

Sethi, R., and M. Yildiz (2012): "Public disagreement," *American Economic Journal: Microeconomics*, 4(3), 57–95.

Shermer, M. (2011): "The believing brain," *Scientific American*, 305(1), 85–85.

Stone, D. F. (2013): "Media and gridlock," *Journal of Public Economics*, 101, 94–104.

A Proofs

A.1 Proof of Proposition 4.2

Proof. To prove the first part, note (4) implies that second-order beliefs, $s_i^{i,-i,t}$, do not change in expectation (over s_i) over time ($E_0(s_i^{i,-i,t}) = \mu_s$ for all t). (8) implies that first-order beliefs, $s_i^{-i,t}$, in expectation either approach second-order beliefs (if $r < 1$) or remain equal to second-order beliefs plus b (if $r = 1$). Thus, first-order beliefs do not increase in expectation, so expected affective polarization cannot occur. Since expected polarization is a necessary condition for strong polarization, this also cannot occur. The second part of the claim follows directly from combining (8) and Lemma 4.1, given the probability limit of a sum is equal to the sum of the probability limits. The third part follows directly from the second part. □

A.2 Proof of Proposition 4.4

Proof. To prove the first part, note expected polarization occurs because

$$E_0\left(\lambda_t\mu_s + (1-\lambda_t)(s_L + \sum_{i=1}^{t}\epsilon_i^L) + (1-\lambda_t)(\frac{1}{t}\sum_{i=1}^{t-1} b_i + \tau_L - \tau_L^{R,0})\right) = \\ \mu_s + E_0\left((1-\lambda_t)(\frac{1}{t}\sum_{i=1}^{t-1} b_i + b)\right),$$

which is increasing in t since both $(1-\lambda_t)$ and $\frac{1}{t}\sum_{i=1}^{t-1} b_i$ increase in t.

To prove the second part, I first prove the following lemma

Lemma A.1. *Let* $x_t = 1 + (\frac{1}{t})\sum_{i=1}^{t-1} x_i$, *with* $x_1 = 1$. *Then* $x_t = H_t = 1 + 1/2 + 1/3 + ... + 1/t$ *and therefore* $\lim_{t\to\infty} x_t = \infty$.

Proof. The proof is by induction. It is easily confirmed that $x_1 = H_1$ and $x_2 = H_2$. Assume $x_t = H_t$. We then want to show $x_{t+1} = H_{t+1}$.

$$\begin{aligned} x_{t+1} = 1 + (1/(t+1))\sum_{i=1}^{t} x_i) \leftrightarrow \ &= 1 + (1/(t+1))(H_1 + H_2 + ...H_t) \leftrightarrow \\ &= 1 + (1/(t+1))\Big(1 + (1+1/2) + ... + (1+1/2+...+1/t)\Big) \\ &= 1 + (1/(t+1))\Big(t + (t-1)(1/2) + ... + 1/t\Big) \\ &= 1 + t/(t+1) + (t-1)/(2(t+1)) + ... + 1/(t(t+1)) \\ &= 1 + (1 - 1/(t+1)) + (1/2 - 1/(t+1)) + ... + (1/t - 1/(t+1)) \\ &= 1 + 1 + 1/2 + 1/3 + ... + 1/t - t/(t+1) = H_{t+1}. \end{aligned} \tag{22}$$

□

The lemma implies that b_t approaches bH_t as $1-\lambda_t$ approaches 1 for all t. Since H_t diverges, bH_t diverges for all $b > 0$. $1-\lambda_t$ always increases in t, and can be arbitrarily close to 1 for any t for sufficiently large σ_s^2 and small σ_τ^2 and σ_ϵ^2. Thus, for sufficiently large t and σ_s^2 and small σ_τ^2 and σ_ϵ^2, b_t can be arbitrarily large for any $b > 0$.

To prove the third part, note

$$E_0(s_L^{R,t} - s_L^{R,t-1}|s_L, \tau_L) =$$
$$\Delta\lambda_t(\mu_s - s_L - (\tau_L - \mu_\tau + b)) + \left((1-\lambda_t)(\frac{1}{t}\sum_{i=1}^{t-1} b_i) - (1-\lambda_{t-1})(\frac{1}{t-1}\sum_{i=1}^{t-2} b_i)\right) =$$
$$\Delta\lambda_t(\mu_s - s_L - (\tau_L - \mu_\tau + b)) + (1-\lambda_t)\Big(\frac{1}{t}\sum_{i=1}^{t-1} b_i - \frac{1}{t-1}\sum_{i=1}^{t-2} b_i\Big)$$
$$+(1-\lambda_t)(\frac{1}{t-1}\sum_{i=1}^{t-2} b_i) - (1-\lambda_{t-1})(\frac{1}{t-1}\sum_{i=1}^{t-2} b_i) =$$
$$\Delta\lambda_t(\mu_s - s_L - (\tau_L - \mu_\tau + b) - \frac{1}{t-1}\sum_{i=1}^{t-2} b_i) + (1-\lambda_t)\Big(\frac{1}{t}\sum_{i=1}^{t-1} b_i - \frac{1}{t-1}\sum_{i=1}^{t-2} b_i\Big). \quad (23)$$

$\Delta\lambda_t$ is always negative, and so by part two of the proposition $(\mu_s - s_L - (\tau_L - \mu_\tau + b) - \frac{1}{t-1}\sum_{i=1}^{t-2} b_i)$ can be guaranteed to be negative for sufficiently high t, making its product with $\Delta\lambda_t$ positive. And the second term, $(1-\lambda_t)\Big(\frac{1}{t}\sum_{i=1}^{t-1} b_i - \frac{1}{t-1}\sum_{i=1}^{t-2} b_i\Big)$, is always positive since b_i is increasing. □

A.3 Proof of Proposition 4.5

Proof. Expected polarization occurs because

$$\begin{aligned} E_0(s_{-i}^{i,t}) &= \mu_s + m_t, \text{ in which} \\ m_t &= (1-\lambda_t)(\mu_s + \frac{1}{t}\sum_{i=1}^{t-1} m_i) \text{ for } t > 1 \text{ with } m_1 = (1-\lambda_1)\mu_s, \end{aligned} \quad (24)$$

and m_t is increasing.

To prove part 2, first note $1 - \lambda_t = \sigma_s^2/(\sigma_s^2 + \sigma_\epsilon^2/t)$. This is increasing in σ_s^2, so showing the claim holds for the case $\sigma_s^2 = \sigma_\epsilon^2$ is sufficient. In this case $1 - \lambda_t = t/(t+1)$, and thus it is sufficient to prove $s_t = (t/(t+1))(1 + (\frac{1}{t})\sum_{i=1}^{t-1} s_i)$ for $t > 1$ (and $s_1 = 1/2$) diverges. This is implied by the following result, which is related to A.1 and so I again state and prove as a separate lemma.

Lemma A.2. *Let* $x_i = (t/(t+1))\left(1 + (\frac{1}{t})\sum_{i=1}^{t-1} x_i\right)$ *for* $t > 1$ *with* $x_1 = 1/2$. *Then* $x_t + 1 = H_{t+1} = 1 + 1/2 + 1/3 + ...1/(t+1)$, *and thus* x_t *diverges.*

Proof. Again, the proof is by induction. The claim is true for x_1. Assume it is true for x_i for

all $i < t$. Then

$$\begin{aligned}
x_i &= (t/(t+1))\left(1+\frac{1}{t}(H_2-1+H_3-1+...+H_t-1)\right)\\
&= t/(t+1)+(1/(t+1))\Big((1/2)+(1/2+1/3)+...(1/2+1/3+...+1/t)\Big)\\
&= t/(t+1)+(t-1)/(2(t+1))+(t-2)/(3(t+1))+...+1/(t(t+1))\\
&= t/(t+1)+(t+1-2)/(2(t+1))+(t+1-3)/(3(t+1))+...+(t+1-t)/(t(t+1))\\
&= t/(t+1)+1/2-1/(t+1)+1/3-1/(t+1)+...1/t-1/(t+1)\\
&= t/(t+1)+(H_t-1)-(t-1)/(t+1)=(H_t-1)+1/(t+1)=H_{t+1}-1. \qquad (25)
\end{aligned}$$

□

This implies that $E_0(s_{-i}^{i,t}|s_L,s_R)$ diverges for $\sigma_s^2 \geq \sigma_\epsilon^2$. The proof of strong polarization is analogous to the corresponding proof for Proposition 4.4. □